UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS

POSTGRADO EN CIENCIAS DE LA COMPUTACIÓN

"ESTUDIO DE DOS PARADIGMAS DE MODELADO DE TÓPICOS EN UN CORPUS DE DOCUMENTOS TOMADOS DE UNA RED SOCIAL"

Trabajo de Grado de Maestría presentado ante la ilustre Universidad Central de Venezuela por la Dra. Luz Marina Barreto para optar al título de Magister Scientiarum mención Ciencias de la Computación.

Tutora: Dra. Haydemar Núñez

Caracas-Venezuela

Marzo de 2018

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
COMISIÓN DE ESTUDIOS DE POSTGRADO

Comisión de Estudios de Postgrado

VEREDICTO

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias de la Universidad Central de Venezuela, para examinar el **Trabajo de Grado** presentado por: **Dra. Luz Marina Barreto, Cédula de identidad Nº 5.310.785**, bajo el título "**Estudio de dos paradigmas de modelado de tópicos en un corpus de documentos tomados de una red social**", a fin de cumplir con el requisito legal para optar al grado académico de **MAGÍSTER SCIENTIARUM, MENCIÓN CIENCIAS DE LA COMPUTACIÓN**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 17 de Abril de 2018 a las 11:00 AM.,para que la autora lo defendiera en forma pública, lo que ésta hizo en el Auditotio Manuel Bemporad, mediante un resumen oral de su contenido, luego de lo cual respondió satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **aprobarlo**, por considerar, sin hacerse solidario con la ideas expuestas por la autora, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado

Para dar este veredicto, el jurado estimó que el trabajo examinado resulta un aporte para la comprensión de las técnicas frecuentitas y probabilísticas para el modelado de tópicos en corpus de documentos, y cómo éstas pueden ser útiles en la reconstrucción de la semántica o determinación de los conceptos latentes de conjuntos de textos recuperados a partir de redes sociales. En particular, se logró determinar las connotaciones semánticas alrededor del concepto "Derechos", que subyace a las emisiones de los usuarios de Twitter que mencionan este término en el ámbito Iberoamericano.

En fe de lo cual se levanta la presente ACTA, a los 17 días del mes de Abril del año 2018, conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinadora del jurado la Dra. Haydemar Núñez.

El presente trabajo fue realizado bajo la dirección de Dra. Haydemar Núñez.

Dr. Eugenio Scalise C.I.10.184.983
U.C.V

Dr. Rhadamés Carmona C.I.10.804.242
U.C.V.

Dra. Haydemar Núñez C.I. 5.538.772
U.C.V
Tutor(a)

POSTGRADO EN CIENCIAS DE LA COMPUTACION
Universidad Central de Venezuela

Resumen

El presente trabajo estudia tres técnicas de modelado de tópicos, o reducción de la dimensionalidad, en conjuntos de datos de texto para la recuperación de la semántica en matrices dispersas de bolsas de palabras. Son técnicas destinadas al modelado de tópicos o conceptos latentes en conjuntos desestructurados de datos de texto. Estas tres técnicas se estudian en el marco de dos paradigmas generales del análisis estadístico de datos: el paradigma frecuentista y el paradigma de la inferencia Bayesiana. Las tres técnicas son: el análisis semántico latente o LSA, el análisis semántico latente de índole probabilística o PLSA y la atribución latente de Dirichlet o LDA. El trabajo estudia los fundamentos teóricos que subyacen al desarrollo de sus algoritmos y aplica sus implementaciones, en el lenguaje de programación Python, a un corpus de documentos tomados de la red social Twitter. El corpus consta de tres conjuntos de datos de texto en los cuales se busca reconstruir la semántica del concepto "derechos", tal y como se expresa en emisiones de usuarios de Twitter provenientes del entorno Iberoamericano. Al analizar los resultados obtenidos en las aplicaciones se pudo comprobar que el algoritmo de la LDA ofrece una semántica más general y profunda del concepto de "derechos", al atravesar transversalmente los documentos, que el algoritmo del PLSA, cuyos resultados dan mejor cuenta de la semántica *ad intra* de los documentos. Al mismo tiempo, fue posible constatar que los algoritmos que implementan modelos de inferencia Bayesiana son más eficientes para la tarea de modelado de tópicos que los algoritmos que calculan valores singulares en matrices factorizadas. También se pudo comprobar un manejo competente de la semántica de la noción de "derechos" por parte de los usuarios de esa red social, el cual evidencia familiaridad con el significado teórico e institucional de dicho concepto. No obstante, ese manejo se mantiene siempre en un nivel elevado de convencionalidad.

Palabras claves: modelado de tópicos, reducción de la dimensionalidad, semántica latente, factorización de matrices, modelos de inferencia Bayesiana, algoritmos LSA, PLSA y LDA.

A la memoria de mi padre

Esteban Barreto Ravell

Ingeniero Electricista egresado de la UCV

Tabla de contenidos:

Índice de tablas

Índice de figuras

"Si tuviera una hora para resolver un problema,
ocuparía 55 minutos pensando en el problema
y 5 minutos pensando en su solución."

Albert Einstein

"Los algoritmos para el modelado de tópicos son,
con frecuencia, adaptaciones de métodos más generales
para aproximarse a la distribución posterior."

David Blei

"No preguntes si un enunciado es verdadero
hasta que sepas bien qué significa"

Errett Albert Bishop

Introducción.

El propósito general del siguiente trabajo de grado es estudiar tres técnicas para la reconstrucción automatizada o el reconocimiento automatizado de la semántica en conjuntos de datos de texto, también conocidos como *modelado de tópicos o conceptos*, y aplicar sus algoritmos sobre tres conjuntos de datos de texto tomados de la red social Twitter.

Las tres técnicas objeto de estudio son:

1. El análisis semántico latente o LSA (por sus siglas en inglés: *latent semantic analysis*)

2. El análisis semántico latente de índole probabilística o PLSA (por sus siglas en inglés *probabilistic latent semantic analysis*) y

3. La atribución o "colocación" latente de Dirichlet o LDA (*latent Dirichlet allocation*).

Estas tres técnicas pueden enmarcarse en dos paradigmas más amplios que son objeto de mucha discusión en las ciencias de la computación contemporáneas, a saber:

1. Los modelos de estadística "frecuentista" y

2. Los modelos de inferencia probabilística o modelos Bayesianos.

La diferencia entre los paradigmas de análisis frecuentista y de inferencia probabilística ha recibido últimamente una particular atención por parte de los investigadores en ciencias de la computación porque, desde hace relativas pocas décadas y con el aumento de la capacidad computacional, se ha abierto por primera vez para el investigador en ciencias naturales y sociales la posibilidad de escalar los cálculos derivados de conjuntos de datos sobre los que se han estimado frecuencias de variables (el ámbito tradicional de los análisis frecuentistas y los de la estadística descriptiva, incluyendo el cálculo de probabilidad sobre modelos paramétricos), en la dirección de un genuino cálculo de inferencia estadística, el ámbito del cálculo de probabilidades de cantidades y variables *desconocidas* en modelos no paramétricos.

La distinción entre la estadística frecuentista y la inferencia probabilística es también una distinción relativamente reciente. Los filósofos y teóricos en ciencias políticas y jurídicas que estudian la teoría de la elección racional, en particular los que estudian ética y filosofía política, están familiarizados con ella gracias al trabajo clásico de Duncan Luce y Howard Raiffa, *Games and Decisions* (1957), quienes extendieron al campo de la ciencia social aplicada y al de los modelos conductistas de actores sociales, los desarrollos pioneros de los matemáticos John Von Neumann y Oskar Morgenstern, en *Theory of Games and*

Economic Behavior (1947), para comprender de manera matemáticamente precisa los juegos de n-personas en donde se producen conflictos de interés y en los cuales hay que tomar decisiones sobre estrategias cuyas probabilidades de éxito se estiman como distribuciones de probabilidad.

En Luce y Raiffa (1957), esta teoría de las decisiones racionales formalizadas, teoría que, entre tanto, se constituyó en uno de los pilares de la inteligencia artificial, se divide conforme a los siguientes criterios:

1. Si la decisión es hecha por un individuo o un grupo (dos o más individuos) y

2. Si la decisión se realiza bajo condiciones de: *certidumbre*, *riesgo* e *incertidumbre*.

En el momento en el que Luce y Raiffa escriben su trabajo, la inferencia estadística combina en un mismo ámbito de indagación decisiones en condiciones de *riesgo* (cuando se conocen los parámetros de un modelo estadístico) y decisiones en condiciones de *incertidumbre* (cuando no se conocen los parámetros, por ejemplo, cuando no se sabe qué tipo de estrategia elegirá un contendor en un juego de suma no nula). No obstante, con el paso de los años, y con el crecimiento de nuestra capacidad de computación, ambos tipos de decisión se han ido distanciando una de la otra.

Se define como decisiones en condiciones de *certidumbre* aquella en la cual es posible calcular con precisión qué acciones conducen a resultados específicos (Luce y Raiffa, 2012, 13). Este es el ámbito de la programación lineal y la investigación de operaciones: con un número acotado de datos se trata de poder ajustar un conjunto de datos, del modo más preciso posible, a una función matemática.

Luce y Raiffa definen seguidamente las decisiones en condiciones de *riesgo* como aquellas en la que pueden calcularse las distribuciones de probabilidad de que las acciones conduzcan a resultados específicos a partir de las variables más frecuentes en matrices de conjuntos de datos.

Finalmente, las decisiones tomadas en condiciones de *incertidumbre* se definen para un conjunto de resultados que se especifican conforme a una distribución de probabilidad *desconocida* en modelos no paramétricos.

La distinción entre los tres tipos de decisiones es muy importante en las ciencias de la computación actuales porque muchos investigadores están preocupados por diseñar mejores y más eficientes algoritmos destinados a modelar conjuntos de datos que crecen exponencialmente. De este modo, la creciente popularidad que comienzan a gozar los modelos computacionales Bayesianos, de los que, por ejemplo Geoffrey Hinton, Radford Neal o David Blei son

pioneros y entusiastas defensores, descansa en su capacidad (o presunta capacidad) para permitir a las máquinas de computación la toma de decisiones en condiciones de cada vez mayor *incertidumbre*, es decir, en donde el cálculo de los parámetros probabilísticos y de la probabilidad marginal de una variable se vuelven intratables computacionalmente o se van perdiendo en cada nueva iteración de un algoritmo. Por esta razón, en contraste con el escenario que Luce y Raiffa tienen ante sí, hoy en día el cálculo de la incertidumbre está en manos de modelos computaciones que se apoyan en la capacidad de la regla de Bayes para encarar el cálculo de la probabilidad posterior sobre familias exponenciales profundas. [1]

Como se verá a lo largo de este trabajo, el LSA y el PLSA son modelos típicamente frecuentistas, en donde el PLSA, como su nombre lo indica, calcularía las probabilidades que se desprenden de las frecuencias detectadas o parámetros disponibles en un conjunto de datos.[2] Por contraste, la LDA aspira a dar un salto cualitativo para el cálculo de la probabilidad posterior, -una cantidad desconocida, es decir, no detectada como frecuencia en un conjunto de datos-, en el modelado de tópicos de índole probabilístico y por eso los Bayesianos lo consideran un enfoque orientado ya al cálculo de la incertidumbre (por su enfoque orientado a optimizar el cálculo de la máxima probabilidad posterior o MAP, *maximum a posteriori*, y su capacidad de conjugar la probabilidad previa para sucesivas iteraciones de los algoritmos).

La LDA es, por ello, el punto de partida del diseño de redes neuronales Bayesianas de aprendizaje profundo para modelado de tópicos (Cfr. D. Blei, 2016). También es interesante examinar la tesis doctoral de Radford Neal, un discípulo de Geoffrey Hinton, para encontrar allí una justificación del enfoque Bayesiano para el cálculo de la incertidumbre que plantea el diseño de redes neuronales, (Cfr. Neal, 1995). Se retomará en el capítulo 4 de este trabajo la pregunta por la eficacia de la LDA para el cálculo de la incertidumbre sobre distribuciones de probabilidad desconocidas.

Las tres técnicas de modelado de tópicos que son objeto de estudio en el trabajo tienen interés para el científico social y para el filósofo porque atañen a un problema central del análisis del lenguaje ordinario: la reconstrucción de la semántica en conjuntos de datos de texto.

---

[1]Véase también Bernardo, José (2003) para entender la diferencia entre los modelos frecuentistas y bayesianos, de los que hablaremos más adelante en el presente trabajo de grado.

[2]El LSA también puede decirse que calcula una probabilidad a partir de una frecuencia detectada. En efecto, el coeficiente de *x* en una función de regresión puede definirse como la *probabilidad* de que un punto en el conjunto de datos se acerque a la *media* de la función de regresión y, por lo tanto, pueda ser modelado por ella. Son los modelos Bayesianos los únicos que calculan una probabilidad sobre cantidades desconocidas. Por lo tanto, a diferencia de los otros dos, lidian con la incertidumbre. De este modo, si bien tanto el PLSA como el LDA son modelos que invocan probabilidades de modo explícito, el único que lidia con la incertidumbre o con una probabilidad desconocida es el LDA. Veremos esto con detalle más adelante.

En efecto, la capacidad de *entender* el sentido que emerge de un conjunto de palabras, que es también la capacidad de *clasificar* algo como algo, es, sin duda, el problema más antiguo del conocimiento y, por lo tanto, uno de los problemas más importantes de la filosofía. Se trata de comprender cómo se pueden formar enunciados y juicios que son considerados verdaderos o válidos por hablantes competentes. Por esta razón, el problema de la comprensión del sentido o de la comprensión de la semántica de una oración está ligado al establecimiento de la verdad, de lo que *sea el caso*. Se trata de un problema que define la metafísica antigua, especialmente la aristotélica, pero es en el siglo XX que los filósofos intentan encarar este problema con los desarrollos más recientes de una estrategia lógico-matemática.

En efecto, Aristóteles definió, en su tratado *Los analíticos posteriores,* lo que en los siglos subsiguientes seguía siendo considerado el método de la fundamentación del conocimiento: un conjunto de proposiciones ligadas por relaciones de fundamentación es verdadero si es posible hacer un recuento exhaustivo de todos los axiomas que se encuentran en su base. De este modo, para Aristóteles, el modelo por excelencia de una ciencia verdadera era la geometría euclídea, con sus cinco axiomas, de los cuales podían derivarse todos sus teoremas. La tarea de todas las ciencias era, de acuerdo con el Estagirita, encontrar sus axiomas básicos, de los cuales dependerá la validez de sus teoremas, reglas de inferencia y verdad de sus enunciados. La comprensión racional o semántica no sería sino un derivado natural de este proceso.

De acuerdo con el filósofo español Jesús Mosterín, a finales del siglo XIX el trabajo de fundamentación ontológica del conocimiento seguía siendo el mismo que era 25 siglos antes para Aristóteles: el de encontrar los fundamentos lógicos básicos de una ciencia, o sus axiomas y teoremas. El gran matemático David Hilbert define, en 1900, la principal tarea de la filosofía en el siglo que apenas empieza como un recuento completo de los fundamentos lógico-matemáticos de toda la ciencia, tal que sea posible encontrar el conjunto completo de los axiomas y teoremas que nos permitan modelar de manera matemática las relaciones entre entidades y eventos en la realidad física y computar todos los problemas planteados en ese ámbito de modo exhaustivo o completo. Dicha reconstrucción era, para Hilbert en1900, como para Aristóteles, el fin último del conocimiento (Cfr. Mosterín, 2000, 171 y ss.)

En 1931, el matemático de origen checo Kurt Gödel demostrará que esta empresa es una tarea imposible: una ilusión. Lo descubre en su intento de axiomatización de la aritmética, en el que demuestra que si un sistema de axiomas es completo, será inconsistente, y si es consistente, será incompleto. Se trata de su famoso *teorema de la incompletitud*, que da al traste con el proyecto formulado por Hilbert un par de décadas antes, en la reunión de 1900 del Congreso de Matemática de París.

No obstante, antes de que la ciencia del siglo XX asimile verdaderamente el descubrimiento de Gödel, buena parte de los esfuerzos de la lógica matemática de principios de siglo estaba orientada a llevar a cabo el plan de Hilbert para la fundamentación del conocimiento. En Inglaterra y en Alemania destacan los esfuerzos de Bertrand Russell y Gottlob Frege en esta dirección. Ahora bien, a medida que el proyecto de encontrar los axiomas últimos, (o verdades autoevidentes que no necesitan de una demostración porque pueden ser comprendidas de modo inmediato), de la ciencia va desintegrándose, el problema de explicar cómo es que podemos decir que algo es verdadero, o lo entendemos, comienza a exigir otras explicaciones.

El gran filósofo austríaco del siglo XX, Ludwig Wittgenstein, que en 1921 ya atisba la inutilidad del trabajo de su maestro Bertrand Russell, sugiere que nos entendemos sobre algo si es posible señalar con el dedo, por decirlo así, aquel "estado de cosas" que hace que un enunciado sea comprensible a un interlocutor y, por lo tanto, verdadero en algún tipo de contexto. Por ejemplo, si alguien dice "la pelota es roja" (para dar un ejemplo clásico), el enunciado es verdadero si y sólo si…la pelota es roja. Con ello, Wittgenstein deja de lado la necesidad de encontrar axiomas últimos que sean susceptibles de un tratamiento matemático. Sólo aquel que "tiene un mundo", bien sea porque vive en él con otros hablantes o porque comparte un lenguaje común con otros, puede decir que sabe lo que alguien quiere decir cuando dice que la pelota es roja.

Pero con ello se abre este otro problema: ¿por qué entendemos un enunciado cuando no podemos reducirlo a un estado de cosas en el mundo? Gran parte de la filosofía del siglo XX es un intento de responder a esta cuestión, que atañe a la semántica de un enunciado.

Ahora bien, cuando se habla de adquisición de conocimiento en las ciencias de la computación se alude a algo distinto al conocimiento en este sentido analítico-conceptual. En computación, de acuerdo con el profesor de ciencias de la computación de la Universidad de Washington, Pedro Domingos, el conocimiento es el resultado de aplicar distintas herramientas de análisis estadístico a conjuntos de datos (Cfr. Domingos, 2015).[3]

El objetivo del presente trabajo de grado es, precisamente, estudiar aquel conjunto de herramientas destinadas a reconstruir de manera automatizada la

[3]El problema de cómo interpretar la adquisición de conocimiento sigue siendo objeto debate filosófico. Los filósofos de la mente que piensan que el cerebro funciona como una súper computadora estarían de acuerdo con Domingos, pero no así filósofos contemporáneos como Raymond Tallis (Tallis, 2014), quien, al igual que Kant o Quine, piensa que la capacidad de orientarse en el mundo y de comprenderlo no puede interpretarse como el resultado de haber hecho muchas experiencias que resultarían estadísticamente relevantes para la formación de creencias sobre el mundo. Es al revés: se posee a priori una serie de conceptos básicos que ayudan al observador a clasificar experiencias.

semántica de una oración, o de un documento, que en la literatura de las ciencias de la computación se conoce como *métodos de modelado de tópicos*.

El modelado de tópicos constituye una de las herramientas más interesantes y utilizadas para la tarea del descubrimiento de la semántica que subyace a conjuntos de datos de texto. Busca el descubrimiento, con herramientas automatizadas, de los *conceptos* o *tópicos* a la base de conjuntos de datos de texto. Este trabajo, que, hasta hace relativamente poco tiempo en la historia, era realizado por un filósofo solitario llevado por su intuición o capacidad de análisis, ¿cómo puede hacerse de manera automatizada?

La pregunta es importante porque, por primera vez en la historia, los humanistas y científicos sociales tienen acceso a enormes cantidades de datos con los cuales es posible trabajar para llegar a conclusiones generales de carácter predictivo. Como sugiere Pedro Domingos, en el libro ya citado, la ciencia física se desarrolló espectacularmente al inicio de la era moderna porque pensadores como Kepler o Newton contaban ya con suficientes datos estadísticos que les permitieron corroborar si el movimiento de los planetas y estrellas se conformaban o no a las leyes de la física que intentaban formular.

De un modo análogo, los científicos sociales tienen ahora suficientes datos, en la forma de millones de archivos de texto en la Web, como para saber si el modelado conceptual de un tópico, hecho de manera automatizada, confirma o no lo que es posible formular desde el análisis conceptual, la herramienta de trabajo de un filósofo (Cfr. Strawson, 1992).

Por esta razón, el propósito del siguiente trabajo de grado es estudiar tres técnicas de reconstrucción de la semántica en tres conjuntos desestructurados de datos de texto, y hacerlo en el marco de dos paradigmas claramente discernibles: el frecuentista (típico de la estadística descriptiva) y el probabilístico (que se divide, a su vez, en dos: el enfoque probabilístico de tipo frecuentista y el enfoque probabilístico que se apoya en la inferencia Bayesiana en condiciones de incertidumbre).

Estas tres técnicas serán proyectadas sobre tres conjuntos de datos de texto relacionados con el área académica y profesional de trabajo de quien esto escribe: la fundamentación de la ética y la filosofía del derecho. Los conjuntos de datos de texto provendrán de la red social Twitter. En las aplicaciones se examinarán las connotaciones semánticas o los tópicos más frecuentes alrededor del uso de la palabra o concepto "*Derechos*", que funcionará como operador simbólico para explorar la semántica latente que subyace a las emisiones de los usuarios de Twitter que mencionan este término en sus tweets. Todos los documentos analizados provienen del ámbito Iberoamericano.

Este trabajo se divide 7 capítulos, incluyendo el capítulo que recoge las conclusiones.

El capítulo 1 es una introducción general a distintos métodos para el análisis automatizado y tratamiento de conjuntos de datos de texto. En este capítulo, se ofrecerá un contexto para las tres técnicas de reconstrucción de la semántica que son el objeto de este trabajo de grado, así como se ofrecerá un breve recuento de algunas de las aplicaciones tratadas por la literatura. Se inicia con el estudio de la naturaleza de las matrices de conjuntos de datos de texto y el análisis de su preparación, como matrices de términos-documentos, para su tratamiento por parte de los algoritmos de modelado de tópicos y reducción de la dimensionalidad. Dado que el LSA es un método de análisis de conjunto de datos de texto que aplica una técnica de factorización y descomposición de matrices en valores singulares, en este capítulo se definirán también los dos procedimientos de factorización de matrices más conocidos para la reducción de la dimensionalidad de matrices dispersas: el análisis de componentes principales o PCA, por sus siglas en inglés *principal component analysis*, y la descomposición en valores singulares o SVD, *singular value decomposition*.

Los siguientes tres capítulos ofrecerán un estudio más detallado de los fundamentos teóricos de cada una de las tres técnicas usadas en el presente trabajo de grado y se ofrecerán pequeños ejemplos para ilustrar sus algoritmos.

El capítulo 2 entra de lleno en el análisis de la primera técnica para el descubrimiento de la estructura semántica latente en conjuntos de datos de texto: el LSA o *latent semantic analysis*, de la cual se ofrecerá sus fundamentos teóricos y matemáticos, así como un ejemplo para ayudar a la intuición y posterior comprensión de la aplicación automatizada que se realizará en este trabajo.

El capítulo 3 se ocupa de la versión probabilística del LSA. El PLSA o análisis semántico latente de índole probabilística (*probabilistic latent semantic analysis*). En un sentido, es una extensión del análisis semántico latente, y una técnica frecuentista, pero en donde la factorización de matrices y la consiguiente reducción de la dimensionalidad se realiza calculando probabilidades (o los parámetros) de que las palabras apunten a un tópico en vez de a otro en un corpus de documentos dado. Ahora bien, dado que con este capítulo se inicia el análisis de matrices de términos-documentos usando el cálculo de probabilidades de manera explícita, este capítulo se inicia con una reflexión más precisa que la ofrecida en esta introducción sobre las diferencias entre los enfoques frecuentistas y probabilísticos y sus implicaciones para las técnicas de modelado de tópicos. El capítulo continúa con el estudio de los fundamentos teóricos propiamente dichos del análisis semántico latente de índole probabilística o PLSA y ofrece ilustraciones de su aplicación.

El capítulo 4 estudia la atribución latente de Dirichlet o LDA y sus implicaciones para el desarrollo de un tratamiento automatizado de la incertidumbre en la inferencia Bayesiana aplicada al modelado de tópicos. Ahora bien, dado que las distribuciones de probabilidad de Dirichlet definen un enfoque matemático destinado a calcular distribuciones de probabilidad en conjuntos de datos que crecen exponencialmente, se dedica una primera parte de ese capítulo al estudio de las propiedades formales de las distribuciones de Dirichlet. Seguidamente, se estudia la aplicación de la LDA en modelado de tópicos de la mano de uno de sus mayores exponentes y pioneros, el profesor de la Universidad de Columbia, David Blei, y se ofrece un pequeño ejemplo de aplicación para ilustrar el uso de sus algoritmos. Este capítulo termina con el estudio de la interpretación y la evaluación de los distintos métodos de modelado de tópicos.

El capítulo 5 expone las aplicaciones de las técnicas, métodos y algoritmos estudiados sobre tres conjuntos de documentos o tweets tomados de la red social Twitter. Se inicia con las implementaciones y algoritmos que se han usado en este trabajo, en el cual el lenguaje de programación ha sido *Python* y distintas bibliotecas de métodos implementadas en ese lenguaje.

El capítulo 6 interpreta y evalúa los resultados obtenidos por las aplicaciones.

Por último, el capítulo 7 ofrece las conclusiones a este trabajo.

Finalmente, no quisiera concluir esta introducción sin expresar mi profundo agradecimiento a mi tutora, la Dra. Haydemar Núñez, Profesora Titular de la Facultad de Ciencias de la UCV, por su generosa disposición para orientarme cuando me acerqué al Postgrado de Ciencias de Computación por primera vez para indagar sobre la posibilidad de iniciar un estudio riguroso sobre el análisis de conceptos filosóficos con herramientas automatizadas, el punto de partida de mi interés por la minería de conjuntos de datos de texto y el modelado de tópicos latentes. La atenta dedicación de mi tutora, y colega de la UCV, por ayudarme a comprender la perspectiva de un científico de la computación, su rigor como investigadora y docente, y su paciencia, no carente de interés, con mis hábitos de pensamiento filosófico, ponen de relieve lo mejor del espíritu académico y científico de nuestra Universidad Central de Venezuela, la cual, gracias a personas como ella, sigue y seguirá siendo un centro de investigación de primer nivel en Venezuela y en Iberoamérica.

También quisiera agradecer a todos los que han sido mis profesores del Postgrado de Ciencias de la Computación a lo largo de mis estudios de Maestría: Esmeralda Ramos, Nora Montaño, Iván Flores, Francisca Losavio, Jesús Lares y

Eugenio Scalise. Haber recibido clases u orientaciones con ellos me ha dado una lección de verdadera humildad y ha enriquecido enormemente mis propios estilos y métodos de enseñanza y estudio personal.

Quisiera recordar aquí también a mis compañeros de estudio del Postgrado: Victoria Noguera, Richard Serrano, Paulo Pérez, Livia Borjas, Juan Manuel Acosta, quienes, con su compañerismo, me hicieron revivir la alegría y el solaz que ofrece el aprendizaje que se hace entre amigos. Como alguien que venía de una Facultad distinta, mis compañeros fueron enormemente generosos y atentos al explicarme conceptos y el uso de herramientas con los que yo no estaba familiarizada. En este sentido, mis compañeros José Ramón García y Wilmer González merecen una mención especial por haberme ayudado, en momentos cruciales, con algunas líneas de código especialmente intrincadas y difíciles. Ambos expresaron también su disposición a ayudarme con cualquier dificultad y la seguridad de su ayuda, en último término, impulsó y sostuvo mi confianza mientras aprendía yo misma a utilizar el lenguaje de programación Python y a desenvolverme con mayor seguridad entre sus distintas bibliotecas de métodos. Ellos me hicieron recordar que el aprendizaje realizado dentro de una comunidad de individuos que se esfuerzan todos por aprender es uno de los grandes logros y valores del ambiente universitario.

Capítulo 1. Aspectos generales del estudio del modelado de tópicos en conjuntos de datos de texto y su lugar en la literatura de las ciencias de la computación y el análisis de datos.

De acuerdo con Barbara Tabachnick y Linda Fidell (2013), la elección de una técnica de análisis estadístico viene determinada por el tipo de pregunta de investigación que se hace el científico o el grupo de científicos que examina un conjunto de datos. En este sentido, Tabachnick y Fidell identifican cinco tipos de preguntas básicas de investigación para el análisis estadístico. Estas son (2013, 29-31):

1. Las preguntas que examinan el grado de relación entre variables.

2. Las preguntas que examinan el significado de las diferencias de grupo.

3. Las preguntas que buscan predecir la membresía de grupo.

4. Las preguntas por la estructura latente en un conjunto de datos.

5. Las preguntas que interrogan sobre el decurso temporal de un evento.

El presente trabajo se centra en el estudio de tres técnicas para lo que ambas investigadoras han llamado "el análisis de la estructura latente de un conjunto de datos (4)": el análisis semántico latente o LSA, por sus siglas en inglés (*latent semantic analysis*), el análisis semántico latente de índole probabilística o PLSA (*probabilistic semantic analysis*) y la atribución o colocación latente de Dirichlet o LDA (*latent Dirichlet allocation*).

Este capítulo se encuentra estructurado del siguiente modo. El parágrafo 1.1 está orientado a explicar la naturaleza y características de los conjuntos de datos de texto y ofrece un contexto teórico para sus posibles campos de análisis, en contraste con otros tipos de datos y técnicas de análisis estadístico. Se explicará lo que define a los conjuntos de datos de texto en tanto que tales, así como se pondrán de relieve los problemas que conciernen a la representación y preparación de los conjuntos de datos de texto. El parágrafo 1.1 está dividido en dos segmentos: 1.1.1, el cual se ocupa de la preparación de matrices multidimensionales de texto para su análisis estadístico y 1.1.2, que atañe al estudio de la técnica de ponderación de frecuencias de palabras en documentos *tf·idf*, una técnica habitual de conteo de frecuencias de palabras que será importante a lo largo de este trabajo, dado que será usada en las aplicaciones de los algoritmos sobre su corpus de *tweets*.

Seguidamente, en el parágrafo 1.2 se estudiarán algunos fundamentos teóricos generales que subyacen al estudio de la estructura latente en conjuntos de datos de texto en la literatura sobre análisis automatizado de textos. Se mencionarán las dos técnicas pioneras de reducción de la dimensionalidad: la

descomposición en valores singulares o SVD, *singular value decomposition,* y el PCA o análisis de componentes principales, *principal component analysis*. Se explicarán los aspectos matemáticos que fundamentan estos enfoques y en qué sentido estas dos técnicas son el punto de partida del LSA, PLSA y LDA.

Finalmente, en 1.3. se ofrecerá una visión general de algunas de las aplicaciones tratadas por la literatura que se ocupa de las técnicas de modelado de tópicos, a fin de ofrecer un contexto a la investigación del presente trabajo de grado en el marco actual disciplinar que define a las ciencias de la computación.

1.1. La representación y preparación de conjuntos de datos de texto.

1.1.1. Preparación de matrices multidimensionales.

El minado de conjuntos de datos textos es una disciplina especial del análisis de datos que aprovecha la enorme presencia de documentos de texto en la Web, las redes sociales, los servicios de difusión de noticias y bibliotecas online. La paulatina digitalización de la palabra escrita y hablada, así como la proliferación de bibliotecas digitales, aplicaciones en la Web, y servicios de difusión de noticias, han crecido y seguirán creciendo exponencialmente a medida que pasan los años. Con este incremento del lenguaje humano digitalizado se han sustituido, por ejemplo, libros impresos por archivos digitales, facilitado la búsqueda e investigación sobre aspectos puntuales de un dominio científico, acumulado información escrita a la que se puede acceder como hipertexto y analizado los productos de las agencias de información y la creación de noticias (Aggarwal, 2015, 411). Todo ello abre para el científico social la exploración y comprensión científica del ámbito de la expresión humana comunicada, sin duda un acceso privilegiado a los procesos de creación de la subjetividad humana.

Las herramientas de análisis de texto aprovechan la cada vez más creciente disponibilidad de datos de texto digitalizados que pueblan la *world wide web*. Esta disponibilidad ha venido también acompañada por avances de tecnologías de software y hardware que permiten mejores capacidades de computación y mejoras en la gestión de la complejidad. Todo ello ha impulsado la innovación en el diseño de algoritmos orientados a descubrir patrones de distintos tipos en los datos de texto, muchos de los cuales han sido posibles gracias a los logros alcanzados con datos de otro tipo.

Ahora bien, siempre de acuerdo con Aggarwal (2015), a diferencia de la gestión habitual de conjuntos de datos, que se expresan en bases de datos etiquetadas previamente por los investigadores o expertos para tareas de clasificación, los datos de texto, al inicio de la tarea de análisis, son típicamente conjuntos de datos desestructurados. Por esta razón, se suelen tratar preferentemente con algoritmos de agrupación o *clustering*. Adicionalmente, la "*search engine*" o un buscador Web es desde hace tiempo la herramienta preferida

en la gestión de datos de texto orientados a la recuperación o captación de información destinada al *clustering* o agrupación de textos, categorización de textos, sumarios o resúmenes de textos y sistemas de recomendación.

Pero la recuperación de información por sí sola no es el objetivo de la minería de textos. En efecto, el reconocimiento o descubrimiento de patrones ocultos o latentes en vistas de su análisis debe ser considerada la meta primaria de la minería de textos, de acuerdo con Aggarwal y ChengXiang (2012). Precisamente por esta razón, la minería de texto que busca patrones en conjuntos de datos de texto desestructurados puede ser usada para *reconstruir la semántica* de una oración o documento. La recuperación de información pudiera decirse que se limita a facilitar el acceso a ésta, conectando la información adecuada con los usuarios que quieren acceder a ella. Pero el procesamiento o transformación de la información recuperada a la que aspira la minería de texto busca algo más: digerir la información para que los usuarios puedan analizarla o entenderla mejor y tomar mejores decisiones racionales con base en ellas, entre otras cosas.

Los datos de textos tienen una serie de características que les son específicas, tales como su dispersión y elevada dimensionalidad, lo que los vuelven susceptibles de técnicas destinadas a disminuirlas. La representación de los datos de texto y su normalización son también aspectos críticos en la gestión de este tipo de datos.

Así pues, un aspecto muy importante de la adecuada representación de los conjuntos de datos de texto para su manipulación es el esfuerzo orientado a la *conservación de la semántica*. Un nivel de representación adecuado para los conjuntos de datos de texto es, pues, aquel que mejor da cuenta de la semántica del término. Esto permite, desde luego, una representación mejor que la de "bolsas de palabras" o matrices de conjuntos de palabras, el tipo de representación que, como veremos en seguida, está al inicio del análisis automatizado de este tipo de datos.

Como ya se ha señalado, el presente trabajo de grado se concentrará en el estudio de tres técnicas de recuperación de la semántica en conjuntos de datos de texto. Estas técnicas tienen en común la tarea de reducir la dimensionalidad de matrices de datos de texto multidimensionales. La intuición teórica que subyace a la idea de identificar la semántica con una tarea de reducción de la dimensionalidad consiste en suponer que la semántica de una oración descansa, básicamente, en conceptos amplios que engloban y dan sentido al texto completo. Si una herramienta de inteligencia artificial es capaz de identificar esos conceptos más amplios u ontológicamente más englobadores podrá reconocer su semántica, es decir, "de qué trata" el texto en cuestión. Ahora bien, las tareas de reducción de la dimensionalidad ocupan un lugar particular entre otras posibles del análisis de conjuntos de datos de texto. Examinemos brevemente algunas de estas posibilidades antes de entrar de lleno en el presente estudio.

Las tareas y técnicas posibles y más importantes para la minería de datos de texto son, de acuerdo con Aggarwal y ChengXiang (2012):

1. Extracción de información. En particular, la extracción de entidades y sus relaciones. Examinarlas como vienen representadas en el texto puede conducir a revelar información de una semántica subyacente y, por lo tanto, a descubrir conocimiento "oculto" en el texto, para hacer inferencias racionales o significantes a partir de él.
2. Sumarios o resúmenes de textos. Se trata de extraer de los textos lo esencial, en forma de un breve sumario que resume un texto más amplio. Estos resúmenes pueden ser de dos tipos: los sumarios "extractivos", que toman algunas unidades singulares de los textos para hacer el resumen y, en segundo lugar, sumarios "abstractivos", que ofrecen una síntesis original, por decirlo así, que no necesariamente está presente en el texto como tal. Estos últimos ofrecen una posibilidad interesante, dado que permiten la construcción de texto original con herramientas de inteligencia artificial.
3. Métodos de aprendizaje no supervisado para datos de texto: el ámbito teórico en el cual se mueve este trabajo. Los métodos de aprendizaje no supervisado más importantes, en cuanto se aplican a datos de texto, son dos: el *clustering* o agrupamiento y el modelado de tópicos. La tarea de agrupamiento consiste en segmentar un corpus de texto de modo que cada uno arroje u ofrezca al invetigador un grupo, o *cluster*, alrededor de un tópico determinado. El agrupamiento y el modelado de tópicos se relacionan entre sí. Los autores sugieren, en efecto, que el modelado de tópicos posibilita un agrupamiento *suave* (*soft*), en donde cada documento comporta una probabilidad estimada de formar parte de un *cluster*. Por lo tanto, se opone a una segmentación rígida de los documentos.
4. La tarea que caracteriza el modelado de tópicos se relaciona de modo estrecho con un tipo de técnica muy importante para el preprocesamiento de conjuntos de datos de textos: aquella que encara los problemas que atañen a la reducción de la dimensionalidad. Al estudiar el modelado de tópicos desde la perspectiva de la reducción de la dimensionalidad, la representación de baja dimensión puede permitir la identificación de palabras "afines", las cuales dan mejor cuenta de la semántica al apuntar a conceptos con mayor capacidad englobadora o aglutinadora que la de esas palabras. El presente trabajo examina con detalle esta posibilidad del análisis automatizado de conjuntos de texto: la recuperación de la semántica a través de una reducción de la dimensionalidad en conjuntos de datos de texto y examina tres técnicas asociadas a este tipo de tareas.
5. Una técnica de este último tipo es el LSA o análisis semántico latente, un método de reducción de dimensionalidad que será examinado en el capítulo segundo del presente trabajo. La reducción de la dimensionalidad persigue representar los datos de manera comprimida, a fin de facilitar su indexación

y recuperación. Uno de sus aspectos importantes es el de resaltar los aspectos claves de la semántica de los datos de texto reduciendo el ruido que introduce en los mismos la polisemia y la sinonimia.

6. Otros métodos de aprendizaje para datos de texto que están disponibles para el investigador en la literatura actual serían métodos de *machine learning* que "entrenan" datos (es decir, pares de puntos de entrada en relación con una función de salida deseada) para "aprender" un clasificador o una función de regresión lineal. Se trata aquí de métodos supervisados. Se mencionan brevemente aquí para facilitar la comprensión del contexto del presente trabajo de grado. No obstante, no se entrará en estos métodos. Los métodos de clasificación usados en la minería de datos también se utilizan con éxito en la minería de conjunto de datos de textos, tales como: los clasificadores basados en reglas, árboles de decisión, clasificadores basados en algoritmos de vecinos cercanos y clasificadores basados en cálculos de probabilidades.
7. Al lado del las técnicas de análisis semántico latente o LSA, tenemos también técnicas probabilísticas para datos de texto, como el PLSA, el análisis semántico latente de índole probabilístico, y la LDA, la atribución latente de Dirichlet. El PLSA, o análisis semántico latente de índole probabilística, es una de las técnicas objeto del presente trabajo y se analizará en el capítulo 3. La LDA será estudiada en el capítulo cuatro, para completar así el escrutinio de las tres técnicas de análisis semántico latente o de reducción de la dimensionalidad que son objeto del presente estudio.
8. Entre las otras tareas posibles en el análisis de conjuntos de datos de texto se encuentra también el minado de flujos o transmisiones de datos de texto. Las redes sociales, pero también las agencias de noticias online, crean un gran flujo de noticias continuas que pueden ser minadas como datos de texto. Se trata de tareas de minado de texto difíciles, dado que mucho de este flujo de datos no se puede guardar en repositorios para preprocesamiento. En este caso, el diseño de algoritmos es mucho más difícil.
9. Minado de conjuntos de datos de texto en diferentes idiomas, lo que ofrece un contexto mucho más rico para la comprensión de la semántica de un documento. Las técnicas relevantes aquí son: traducción automática, recuperación de información multilingüe, análisis de corpora paralela, entre otras.
10. Minado de texto en redes multimedia. Se trata de ver si es posible enriquecer nuestra comprensión de los textos tomando datos de distintos tipos.
11. Minería de texto en las redes sociales. Aunque las redes sociales son una fuente muy importante de datos de texto, ellas implican la dificultad adicional de ser pobremente redactadas o poseer un vocabulario no estándar. Aquí vale la pena explorar también métodos que minan

simultáneamente texto y links o redes de usuarios, a fin de enriquecer la comprensión de éstos. El conjunto de datos que se usará en este trabajo será, precisamente, tomado de una red social, Twitter, lo que ofrecerá la oportunidad de observar cuáles son las dificultades que comporta este tipo de conjuntos.

12. Minería de opiniones y análisis de sentimiento en conjuntos de datos de texto constituyen también un campo de aplicación muy importante, al analizar *reviews* o reseñas de productos para apoyar la toma de decisiones en distintas áreas.
13. Minería de datos biomédicos. Lo que permite a investigadores aprovechar el conocimiento implícito en toneladas de literatura médica que se encuentran disponibles actualmente en distintas bases de datos. Ello permite el reconocimiento de entidades biomédicas desconocidas y su relación con ya existentes, a fin de posibilitar el desarrollo de nuevas ontologías que apunten a relaciones interesantes entre patologías.

En general, el crecimiento exponencial de datos de texto obliga a la creación de técnicas de minado cada vez más poderosas. El minado de texto es cada día más relevante para distintas comunidades de investigadores en distintas áreas, y exige de ellos el dominio de técnicas como las mencionadas de minado de texto y procesamiento de lenguaje natural. Desde el punto de vista de las técnicas que se requieren para mejorar nuestra capacidad de analizar estos datos se ha de tener en cuenta que ellas implican necesariamente:

1. Métodos robustos y escalables para el procesamiento de lenguaje natural.
2. Adaptación al dominio y transferencia de aprendizaje.
3. Análisis contextual de datos de texto, para mejorar su profundidad.
4. Desarrollo de herramientas robustas para la minería de texto.

Entre los conceptos generales que definen este dominio de análisis de datos se encuentran:

1. El concepto de *léxico*: mientras que en lingüística el léxico es el conjunto de términos de los que consta un lenguaje, en minería de texto, *léxico* es el conjunto de atributos, características o dimensiones (si se ven como integrando una matriz) de un texto. Por esta razón, los términos de un léxico se incorporan al análisis de conjuntos de datos de texto como *dimensiones*.
2. El concepto de *corpus*: como en lingüística, un corpus es una colección de documentos, en donde un documento puede verse como una secuencia o un registro multidimensional o como una secuencia discreta de palabras o una *cadena* (*string*). De este modo, una secuencia de 140 o 240 caracteres tomada de la red de Twitter es, en cierto sentido, un documento tan merecededor de ese nombre como un artículo o un ensayo largo publicado

en la web y así será tratada en este trabajo. Los documentos susceptibles de análisis de textos suelen ser secuencias de cientos de miles de palabras, involucrando así, pues, un léxico enorme.

En todas las tareas de análisis de conjuntos de datos de texto, los datos se representan como matrices de datos multidimensionales que registran la frecuencia con las que ocurren en un documento dado con una técnica que se conoce como "*bag of words*", bolsas de palabras o BOW por su siglas en inglés. Las palabras también se conocen como "términos".

En estas matrices, se registra simplemente la frecuencia de un término o una dimensión en un documento o, en representaciones categoriales, solo si está presente o no. Obsérvese el siguiente ejemplo (Tabla 1), que registra en una matriz de bolsas de palabras la frecuencia de las palabras en un corpus de 7 documentos:

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	1	1	1	0	0
B	3	3	3	0	0
C	4	4	4	0	0
D	5	5	5	0	0
E	0	2	0	4	4
F	0	0	0	5	5
G	0	1	0	2	2

Tabla 1: Matriz de documentos/términos.

Como es evidente por este pequeño ejemplo, el problema con este tipo de representación es que con ella se pierde el orden de las palabras en el lenguaje natural, con lo que desaparece precisamente aquel tipo de información que contribuye a la semántica de una palabra.

Por esta razón, la representación matricial, que rinde buenos resultados en otras tareas de estadística descriptiva que miden la frecuencia en un conjunto de datos, es una condición necesaria pero no suficiente para la reconstrucción de la semántica de una oración o un documento. La frecuencia relativa de un dato respecto de una variable no es una buena definición de la semántica, dado que la comprensión semántica exige el dominio del contexto en el que se emite la oración.

Por otro lado, precisamente porque una matriz de bolsas de palabras requiere el conteo de las frecuencias de todas las palabras presentes en un documento, con independencia de si ellas son relevantes para su semántica o no, este tipo de matrices suele ser enorme y dispersa. Esta propiedad, muy citada en la literatura especializada, es específica de conjuntos de datos de texto y se conoce como la dispersión (*sparsity*) de una matriz de datos de texto. Dado que no es posible saber a priori qué término es decisivo para la semántica, en la preparación inicial de una matriz como bolsa de palabras todas ellas deberían, en principio,

poder estar. Esto decide habitualmente el tamaño enorme de este tipo de matrices, su dimensionalidad, con muchos valores iguales a 0.

De esta manera, el primer paso en la construcción de modelos de análisis de conjuntos de datos de texto, entonces, es preprocesar el documento o el corpus para abstraer de él las palabras que tienen menos peso semántico (palabras comunes o frecuentes de un documento respecto de otro, también llamadas "*stopwords*", palabras tales como preposiciones, artículos definidos e indefinidos, signos de puntuación, etc.), así como consolidar las distintas variaciones de una palabra a través de su lematización (que consiste en recuperar sólo la raíz lingüística de la palabra: 'mensaje' en vez de 'mensajeros', 'mensajeras', 'mensajería', etc.).

El orden semántico es sustituido así por un "orden" que descansa únicamente en la frecuencia con la cual una palabra es usada o, en otros casos, en la presencia o ausencia del término, que se denota en la matriz con una variable categórica igual a 1, por ejemplo. Con ello, la dimensionalidad total de este conjunto de datos es igual al número de palabras que conforman el espacio vectorial, menos las que se le han restado. Si, además, decimos que las palabras ausentes en este espacio obtendrán un valor 0, es posible tratar un conjunto de documentos con técnicas de análisis de datos que no son distintas a los procesamientos estadísticos generales de variables, en donde de lo que se trata es encontrar la función que permite una adecuada clasificación o agrupación de un conjunto de datos (Cfr. Aggarwal, 2015, 412).

De esta manera, los textos, en tanto que datos, comportan las siguientes características únicas:

1. La "dispersión" (o *sparsity*) de la matriz de términos-documentos puede ser muy elevada, dado que un mismo documento puede tener sólo un número muy pequeño de palabras relevantes, siendo las demás celdas 0.
2. El otro problema, que se desprende del punto anterior, es su así llamada "no-negatividad": las palabras que aparecen frecuentemente siempre tomarán valores positivos o no negativos (es decir ≥ 0).
3. En tercer lugar, también se hace necesario poder discernir, cuando se hace minería de datos de texto online, la presencia de información asociada, en la forma de metadata o hipertextos, que pudieran oscurecer el análisis que se quiere emprender.

Un primer paso en el análisis de textos consiste, pues, en su modificación de datos desestructurados a una representación multidimensional en matrices de frecuencia de términos como la que vimos en la Tabla 1.

Siempre de acuerdo con Aggarwal (2015), las dos técnicas más importantes para la normalización de la frecuencia de los términos en un documento son dos: *la*

frecuencia inversa del documento (IDF) y la *atenuación de la frecuencia* (FD o *frequency damping*). De ambas, nos interesa en particular la técnica de la frecuencia inversa de un documento, dado que ella permitirá la representación *tf·idf* (frecuencia del término -*term frequency*- por frecuencia inversa del documento, -*inverse document frequency*), el modo de representación más utilizado para la construcción de matrices de conjuntos de datos de texto. Esta última es una técnica estándar que se usa en el procesamiento de los conjuntos de datos de textos presente en los modelos de reducción de la dimensionalidad que se estudiarán a lo largo de este trabajo. De este modo, suponemos que la frecuencia de un término en un corpus se ha establecido con la técnica *tf·idf*.

1.1.2. Técnica de la frecuencia inversa de un documento y representación *tf·idf*.

La representación *tf·idf* pondera la frecuencia con la que aparece un término en *un* documento con la frecuencia del mismo término en la *colección* de los documentos que forman parte del corpus, para dar cuenta de la intuición de que algunas palabras, para ser consideradas características de un documento, deberían, en principio, no ser características de otros o ser menos frecuentes en otros, o bien estar ausentes de ellos.

El procedimiento de normalización *tf·idf* se realiza en dos pasos. En primer lugar, el tipo de normalización *idf* se obtiene dividiendo el logaritmo del número total de documentos en el corpus entre el número de documentos que contienen la palabra cuya frecuencia (*tf*) se examina.

La frecuencia inversa del término *i*-ésimo del documento, que se denota idf_i, es una función decreciente del número total de documentos tal que:

$$idf(t, C) = \log\left(\frac{C}{(d \in C : t \in d)}\right) \qquad (1:1)$$

En donde C es el número de documentos que forman parte de un corpus y $\{(d \in C : t \in d)\}$ designa el número de documentos en donde aparece el término t.

La ponderación tf·idf se calcula entonces, para cualquier término t como:

$$tf \cdot idf(x_i) = tf(x_i) \cdot idf_{x_i} \qquad (1:2)$$

Muchos algoritmos de minería de textos usan, por lo general, este tipo de representación *tf·idf*.

La normalización *idf* supone que palabras o términos con una frecuencia muy alta agregan ruido a las operaciones de minería de conjuntos de datos de texto. Normaliza en el sentido de que da menos peso a las palabras que aparecen con mucha frecuencia en un corpus de documentos, restándole especificidad al documento que se normaliza.

1.2. Los algoritmos de factorización de matrices y los métodos de modelado de tópicos revelan atributos latentes y reducen la dimensionalidad en conjuntos de datos de texto desestructurados.

Las matrices de bolsas de palabras, como ya se ha señalado, se caracterizan por poseer una enorme dimensionalidad, dado que en ellas cada dimensión o atributo, en el caso de conjuntos de datos de texto, equivale a un solo término. El reto de computación consiste, entonces, *en reducir esa dimensionalidad de modo que cada dimensión pueda atribuirse a un concepto o a un tópico común o a otros términos o palabras observadas* y, por lo tanto, distintas palabras puedan agruparse bajo un número menor para iluminar la semántica de dicho tópico o concepto. Dicho de otro modo, puedan ser subsumidas en un número menor de tópicos y supertópicos (Cfr. Crain et al, 2012, 131).

El uso de la factorización de matrices para el modelado de tópicos son una suerte de agrupamiento *soft* o suave, que puede asociar un mismo documento a distintos *clusters*, con la consiguiente reducción de la dimensionalidad. Los documentos pueden ser asociados a distintos tópicos latentes, los cuales pueden ser, tanto grupos de documentos, como representaciones compactas de éstos. Distintos "pesos" o parámetros asignan cada documento a un tópico, tópico que ofrece tanto el criterio de pertenencia de un documento a sí mismo como sus coordenadas espaciales en un corpus. Como resultado, tanto la reducción de la dimensionalidad, como el modelado de tópicos, al conservar en lo posible la representación original de los documentos, reduciéndola, ofrece una representación más significativa de los temas o tópicos de los que se ocupan los documentos que pertenecen al corpus(Cfr. Crain et al., 2012, 132).

Las primeras técnicas de reducción de la dimensionalidad que se encuentran en la literatura utilizan técnicas estándar de álgebra lineal para la factorización de matrices y su descomposición en valores singulares (SVD, *singular value decomposition*, por sus siglas en inglés) o en componente principales (PCA, *principal component analysis*).

Los presupuestos teóricos que subyacen a estas dos técnicas son los siguientes:

Téngase presente, en primer lugar, que todo lo que es un producto se puede factorizar. En el álgebra lineal, el proceso de factorización, que consiste en la *descomposición* de una matriz en otras matrices que, al ser multiplicadas de nuevo entre sí, ofrecerían como producto la matriz original, busca ayudar a descubrir atributos latentes que revelan las relaciones implícitas entre las entidades que conforman la matriz que se descompone. La primera técnica de modelado o descubrimiento de tópicos que se estudia en este trabajo es el análisis semántico latente o LSA, que utiliza un algoritmo de descomposición en valores singulares o SVD y, por lo tanto, es una forma de SVD aplicada a los conjuntos de datos de texto.

La descomposición en valores singulares SVD, y también el análisis de componentes principales o PCA, descubren relaciones implícitas entre términos al interior de una matriz de términos-documentos grande y dispersa. Estas relaciones implícitas con frecuencia son agrupamientos o *clusters*. Cuando la matriz es pequeña esas relaciones implícitas pueden observarse a simple vista, con un conjunto de dimensiones que parecen aludir a un conjunto específico de documentos agrupados arriba a la izquierda, mientras otras son más frecuentes alrededor de un conjunto de documentos agrupados al otro lado de la matriz, por ejemplo. La descomposición en valores singulares intenta así capturar matemáticamente lo que sería fácilmente visible si la matriz en cuestión pudiera ser abarcada completamente con una mirada, lo que no es el caso en conjuntos de datos de texto con muchas dimensiones o términos.

La factorización de matrices es una técnica de reducción de la dimensionalidad porque permite agrupar las entidades de una matriz bajo atributos o tópicos (o conceptos) de mayor envergadura o capacidad englobadora, de forma que entidades aisladas que pudieran haber sido consideradas como su propia clase o categoría, por decirlo así, pueden ahora ser agrupadas bajo categorías analíticas más abarcadoras o de mayor profundidad ontológica.

La descomposición en valores singulares, SVD, y el análisis de componentes principales, PCA, de matrices de datos multidimensionales supone, de acuerdo con Aggarwal (2012, 39), un procedimiento análogo la operación de centrado en la media que ajusta los datos a una función de regresión en modelos lineales. Esta operación de centrado consiste, como en todas las medidas de error en modelos lineales, en sustraer la media del conjunto de datos con respecto a cada punto del mismo; su resultado es un conjunto de datos "centrado" ya en el origen.

Sin embargo, en conjuntos de datos dispersos, y, por lo tanto, no susceptibles de ser representados directamente en modelos lineales, un centrado en la media es especialmente difícil. Por esta razón, las técnicas de SVD y PCA proponen un análogo de este tipo de centrado, como veremos en la Figura 1. El "centrado en la media" que propone el SVD y el PCA no es sino un método ingenioso para encontrar el valor esperado medio que definiría una línea de regresión en datos cuya elevada dimensionalidad impide un cálculo más sencillo, como la suma del cuadrado de los errores por ejemplo, para encontrar correlaciones potenciales entre datos.

Desde el punto de vista de la representación geométrica de su estrategia algorítmica, el objetivo del SVD y PCA es poder rotar los datos a un sistema de ejes en donde la mayor cantidad de varianza de los datos sea capturada con un número mínimo de dimensiones. Desde el punto de vista intuitivo, puede verse en la siguiente Figura 1 que el nuevo sistema de ejes correlaciona de manera "cuasi-lineal" los atributos o dimensiones (componentes) principales:

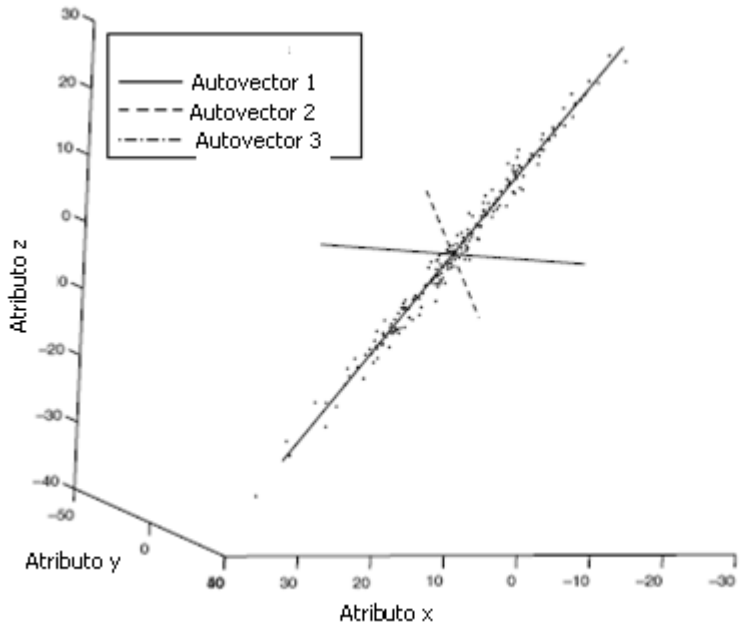


Figura 1: Simulación de una línea de regresión por rotación de ejes en el plano (Adaptado de Aggarwal, 2012, 39). Como puede observarse, en un plano de tres dimensiones o tres atributos, una técnica algebraica de reducción de la dimensionalidad crea una correlación lineal al simularla con un cálculo de "autovectores" o vectores propios.

De este modo, tanto la SVD como el PCA intentan capturar la varianza a lo largo de un eje o dirección particular expresándola con una matriz de covarianza. Lo importante en un estimado de covarianza no es el valor numérico en sí, como sucede en los modelos lineales más sencillos, sino el *signo* de la misma, que relaciona los vectores que componen la matriz con los cuadrantes positivos o negativos en el plano. Un valor positivo nos indica que la relación es directa o creciente. Un valor negativo indica que la relación lineal decrece (los puntos en el plano se dispersan o se alejan unos de otros).

La propiedad algebraica de ortonormalidad de los vectores es muy importante para estas técnicas dado que permiten discernir la covarianza entre los valores vectorizados. Esas direcciones definidas por la propiedad de ser ortonormales se pueden determinar en la medida en que es posible diagonalizar la matriz de covarianza tal que:

$$C = P\Lambda P^T \qquad (1:3)$$

En donde las columnas de la matriz P contienen los vectores propios o autovectores ortonormales de C y Λ en una matriz diagonal que contiene valores propios no negativos. La entrada Λ_{ii} es el valor propio correspondiente al i-ésimo autovector o columna de la matriz P.

Como resultado de esta diagonalización, sugiere Aggarwal (2012, 39-40), tanto los vectores propios como los valores propios tienen una interpretación geométrica en términos de la distribución de los datos subyacente. Si el sistema de

ejes que respresenta los datos es rotado conforme al conjunto de autovectores ortonormales en las columnas de *P*, se puede mostrar que todas las covarianzas de todos los valores de los atributos transformados son ahora iguales a 0. Esto significa que las direcciones que preservan la mayor varianza suprimen las direcciones que expresan correlaciones, que no interesan como componentes principales, y la matriz de covarianza resultante será la matriz diagonal Λ, en donde los autovalores representan las varianzas de los datos a lo largo de los autovectores correspondientes.

De este modo, vectores propios con valores propios grandes preservan una mayor varianza y por eso se denominan *componentes principales*. Dada la naturaleza de la fórmula de optimización usada para derivar esta transformación, un nuevo sistema de ejes que contiene sólo los autovectores con los autovalores más grandes se optimiza de modo que ese nuevo sistema de ejes retenga la máxima varianza en un número fijo de dimensiones. Como veremos en el capítulo 2 este tipo de técnica ofrece los fundamentos matemáticos para los modelos de LSA.

1.3. Algunos ejemplos en la literatura de aplicaciones de modelado de tópicos como los estudiados en este trabajo.

El LSA, de acuerdo con Crain y Ke Zhou (2012), se ha utilizado desde finales de los años 80 para indexar sistemas de recuperación de información, por ejemplo, para asignar artículos a árbitros o en traducciones de documentos. La mayoría de los sistemas de LSA utiliza una SVD parcial o truncada que apela a un algoritmo para computar los valores propios y vectores propios, es decir, la forma espectral, de matrices grandes y dispersas a través de la multiplicación de matrices-vectores. El trabajo clásico que expone la metodología para la recuperación de tópicos es debido a Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer y Richard Harshman "Indexing by Latent Semantic Analysis"(1990), pero también son importantes Martin y Berry (2006) y Golub y Van Loan (1996).

Ahora bien, métodos más poderosos que el LSA se han venido estudiando en la literatura para la reducción de la dimensionalidad en matrices de conjuntos de datos de texto grandes. A medida que crece la disponibilidad de los datos de texto, el LSA ha sido paulatinamente desplazado a favor de otros métodos. Uno de estos métodos es la atribución o colocación latente de Dirichlet, de la que nos ocuparemos en detalle en el capítulo 4.

En un ejemplo de aplicación del método LDA, uno de estos métodos más idóneos para tratar conjuntos de datos de texto de grandes dimensiones, Blei y sus colaboradores infirieron la estructura tópica oculta de una colección de 17.000 artículos científicos de la Revista *Science*, para una probabilidad previa de 100 tópicos. Con esta información, se pudo computar la distribución en tópicos de las palabras de un nuevo artículo, aquella que reflejaba mejor su conjunto de palabras.

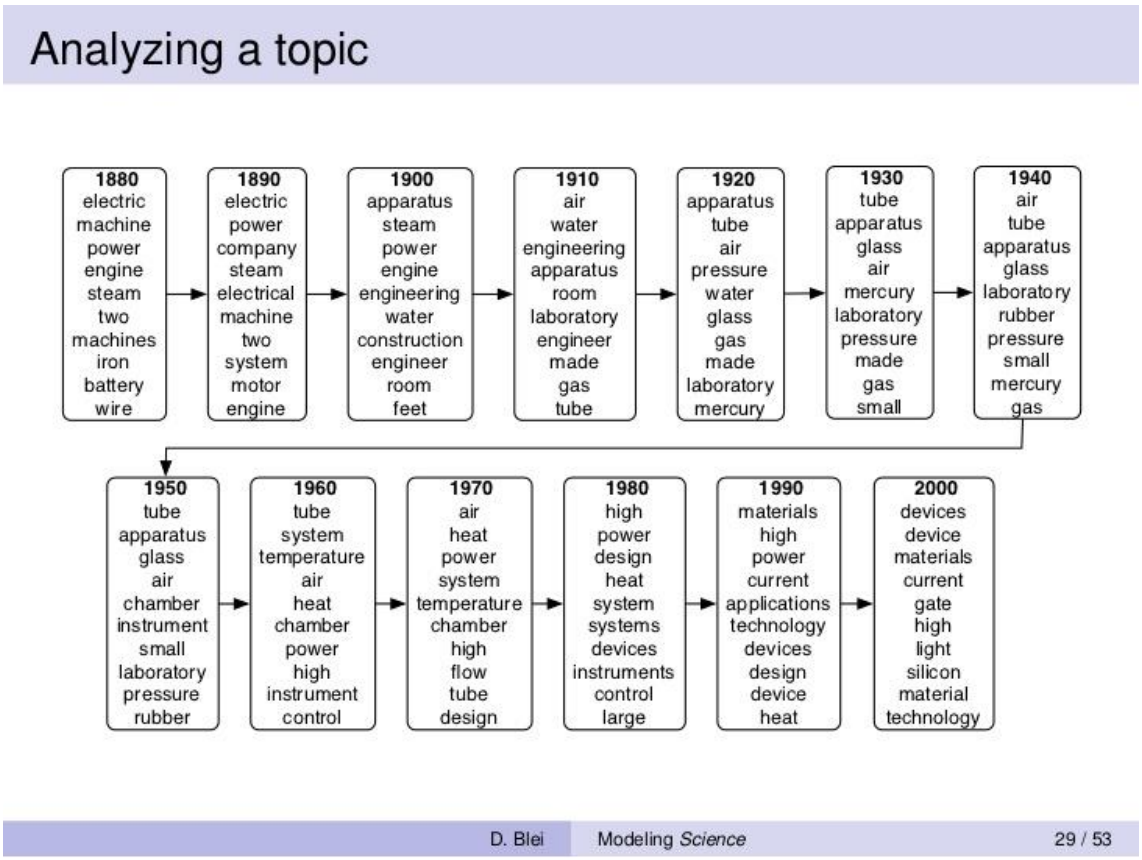


Figura 2: 100 años de tópicos de la Revista *Science*.

En la lámina de la Figura 2, se observa que el modelado de tópicos de más de cien años de artículos de la revista *Science* permite analizar un tópico (en este caso, "aparatos en artículos científicos"), a lo largo de más de un siglo, y los conceptos u ontología asociado a él.

De este modo, mientras que en 1880 los conceptos asociados al tópico "aparatos científicos" son: eléctrico, máquina y máquinas, poder, motor, vapor, dos, hierro, batería y cable, cien años después, en 1980, los conceptos asociados son: alto, poder, diseño, calor, sistema y sistemas, dispositivos, instrumentos, grande. En 1900 emerge una nueva palabra "ingeniería" e "ingeniero"(*engineering*), derivada, en inglés, de motor (*engine*), mientras que cien años después, en el 2000, las palabras importantes asociadas a los aparatos científicos son: sistemas, tecnología, dispositivo, material o silicón.

Esta lámina da una idea de cómo una tarea no supervisada de modelado de tópicos puede rendir una suerte de ontología simplemente a través de la exploración automática de tópicos asociados a aparatos vinculados a la investigación científica a lo largo de un siglo.[4]

[4]La lámina se encuentra disponible en https://www.slideshare.net/ajayohri/modeling-science. La traducción al castellano de todos los términos es la siguiente: 1880: electrizado, máquina, poder, motor, vapor, dos, máquinas, hierro, batería, cable. 1890: electrizado, poder, compañía, vapor,

En otro de los ejemplos que ofrece Blei y su equipo, también es posible que un historiador examine miles de telegramas transmitidos entre embajadas para detectar qué eventos importantes se han producido a lo largo de décadas. Esto se llama "modelado de tópicos dinámico". También es posible tratar de indexar o anotar automáticamente los elementos de una imagen.

La capacidad de indexar documentos permite también predecir los comportamientos de los usuarios dependiendo de lo que la gente lee (o dice). Por otro lado, el comportamiento de la gente, lo que la gente *hace*, constituye una señal adicional del significado de los documentos para ellos y del modo como ha sido organizada la colección. Saber lo que hay en la biblioteca de Charles Darwin puede ayudar a comprender cómo se formó su teoría de la evolución. Lo que está en los estantes de Darwin también puede interpretarse como una *señal* (Blei, 2017 a).

Este tipo de modelado se llama modelado de tópicos colaborativo. Los modelos colaborativos pueden ayudar a los lectores descubrir documentos, viejos y nuevos. También pueden ayudar a describir lectores en términos de sus preferencias de tópicos e identificar documentos que pudieran ejercer algún tipo de impacto en la semántica del tópico.

También es importante entender que los lectores nos dicen algo respecto de los artículos que leen o prefieren. De este modo, el modelado de tópicos nos puede dar información valiosa respecto de artículos interdisciplinarios y los artículos influyentes en un campo e influencias exteriores en un campo.

En Hong y Davidson (2010), otro ejemplo, los autores buscan clasificar los usuarios de la red social Twitter y sus mensajes por el tipo de tópicos que los caracterizan. En su metodología, apelan a la representación de los tweets con la técnica *tf.idf* y el método para el descubrimiento de tópicos usado es el LDA.

En Waal, Alta, Jacobus Venter y Etienne Barnard (2008), los autores exploran igualmente el potencial del método para LDA para recuperar, en emails y otros documentos, información que pudiera ser relevante a los investigadores de policías científicas, quienes muchas veces se encuentran abrumados por la cantidad de documentos que deben revisar en una investigación de carácter forense.

---

eléctrico, máquina, dos, sistema, motor. 1900: aparato, vapor, poder, motor, ingeniería, agua, construcción, ingeniero, cuarto, pie. 1910: aire, agua, ingeniería, aparato, cuarto, laboratorio, ingeniero, hecho, gas, tubo. 1920: aparato, tubo, aire, presión, agua, vidrio, gas, hecho, laboratorio, mercurio. 1930: tubo, aparato, vidrio, aire, mercurio, laboratorio, presión, hecho, gas, pequeño. 1940: aire, tubo, aparato, vidrio, laboratorio, goma (rubber), presión, pequeño, mercurio, gas.1950: tubo, aparato, vidrio, aire, cámara, instrumento, pequeño, laboratorio, presión, goma. 1960: tubo, sistema, temperatura, aire, calor, cámara, poder, alto, instrumento, control. 1970: aire, calor, poder, sistema, temperature, cámara, alto, flujo, tubo, diseño. 1980: alto, poder, diseño, calor, sistema, sistemas, dispositivos, instrumentos, control, grande. 1990: materiales, alto, poder, corriente, aplicaciones, tecnología, dispositivos, diseño, dispositivo, calor. 2000: dispositivo, dispositivos, materiales, corriente, compuerta (gate), alto, ligero, silicón, material, tecnología.

Estas y otras aplicaciones que han sido tratadas por la literatura citada evidencian la flexibilidad y complejidad de los algoritmos de modelado de tópicos, o reducción de la dimensionalidad, aplicados a conjuntos de datos de texto desestructurados.

Capítulo 2. El análisis semántico latente (LSA).

El primer método de modelado de tópicos que se estudiará en este trabajo es el análisis semántico latente o LSA, una forma especial de la técnica matemática de reducción de la dimensionalidad denominada *descomposición en valores singulares* (SVD, *singular value decomposition*). Recordemos que, en el ámbito de los conjuntos de datos de texto, los procedimientos de factorización de matrices son aplicados en la búsqueda de atributos latentes en matrices de términos, los cuales pueden ser descritos efectivamente como verdaderos tópicos o conceptos ocultos que dan sentido al corpus de documentos cuyos términos han perdido su semántica en las matrices dispersas.

El LSA o LSI es un tipo de *análisis* o *indexado* automático (este último sentido de *indexado* se usa en diccionarios o en búsquedas en la web y se lee, por sus siglas en inglés, *latent semantic indexing*) que proyecta tanto términos como documentos en un espacio de dimensiones reducidas, el cual, precisamente por reducir el número de atributos o dimensiones del documento, produce un resultado que puede interpretarse como una recuperación de la semántica del documento original. La intuición que subyace al LSA es que el proceso de reducción de la dimensionalidad captura los términos o tópicos más importantes o básicos desde el punto de vista ontológico o conceptual, de manera que el resultado, al capturar los términos más generales, puede resultar más inteligible, tener más *sentido*. De allí que estos métodos de reducción de la dimensionalidad se consideren también métodos para la recuperación de la semántica de un conjunto de datos de texto disperso y multidimensional.

Por ejemplo, si el LSA se lleva a cabo con la información que arroja el historial de búsqueda de un conjunto de usuarios, una misma inquietud que motiva la búsqueda será naturalmente descrita con términos diferentes por los distintos usuarios. ¿Cómo saber, pues, que están buscando lo *mismo* aunque usando palabras *diferentes*? Como sugieren Crain et al. (2012, 133): "…proyectar los documentos en el espacio semántico le permite a la máquina de búsqueda encontrar documentos con los mismos conceptos aunque usen términos diferentes." Como se trata de un método que reduce la dimensionalidad, al subsumir cada término a un concepto o tópico más general, términos que se asocian a conceptos diferentes también pueden ser correctamente "ubicados" o indexados. Por esta razón, una de sus tareas es resolver los casos de polisemia y sinonimia que complican las matrices de datos de texto.

2.1. La descomposición en valores singulares o SVD aplicada al descubrimiento de la semántica latente en una matriz de conjuntos de datos de texto.

Como ya hemos señalado, el LSA es un caso especial del método más general de factorización de matrices SVD.

La descomposición en valores singulares se encuentra estrechamente relacionada con el análisis de componentes principales dado que apela al mismo tipo de estrategia, a saber, el análisis de la forma espectral de una matriz. Básicamente, el indexado semántico latente se apoya siempre en una descomposición de valores singulares (SVD) parcial de una matriz de términos-documentos que usa un algoritmo para encontrar los autovalores de grandes matrices simétricas. Se trata de un algoritmo iterativo que computa autovalores y autovectores de matrices grandes y dispersas, matrices como las que consideramos aquí, usando la multiplicación de matrices vectores.

La SVD es más general que el PCA (el análisis de componentes principales, *principal components analysis*) porque provee dos conjuntos de vectores de base en vez de uno solo. En efecto, de acuerdo con Aggarwal (2015), la SVD ofrece vectores de base tanto para las filas como para las columnas de la matriz de datos, mientras que PCA lo hace sólo para las filas. Al mismo tiempo, la SVD puede ofrecer los mismos vectores que PCA para las filas de la matriz de datos en los conjuntos de datos en donde la media de cada atributo es 0. Los vectores de base del PCA son invariantes a la traslación de la media, mientras que los de la SVD no. Cuando el conjunto de datos, en ambos tipos de técnicas, no está centrado en la media, los vectores de base de la SVD y del PCA serán cualitativamente diferentes. Una SVD no centrada en la media se aplica usualmente a datos dispersos no negativos como las matrices documento-término o palabra (Aggarwal, 2012, 41).

Una manera formal de definir la SVD consiste en verla como el producto de la descomposición de tres matrices (o de la factorización de tres matrices), tal que, dada la matriz *D* de dimensiones *n*x*d*, se tiene:

$$D = Q\Sigma P^T \qquad (2:1)$$

Aquí Q es una matriz $n \; x \; n$ con columnas ortonormales que son los vectores singulares de la izquierda. Σ es una matriz diagonal $n \; x \; d$ que contiene los valores singulares que son siempre no negativos y ordenados y que, como en el PCA, están arreglados en un orden decreciente. P es una matriz $d \; x \; d$ con columnas ortonormales que son los vectores singulares de la derecha. La matriz Σ es rectangular y no cuadrada pero, a todos los efectos, es una matriz diagonal porque sólo las entradas Σ_{ii} son distintas de 0. El número de entradas diagonales distintas de cero de Σ es igual al rango de la matriz *D*, el cual a lo sumo es $min\{n, d\}$. Gracias a la ortonormalidad de los vectores singulares tanto $P^T P$ como $Q^T Q$ son matrices de identidad (Aggarwal, 2012, Ibídem).

Dado que la matriz de covarianza de los datos centrados en la media es $\frac{D^T D}{n}$ y los vectores singulares de la derecha de la SVD son autovectores de $D^T D$, los autovectores del PCA son los mismos que los vectores singulares de la derecha de la SVD, para el caso de los datos centrados en la media. Los valores singulares

elevados al cuadrado en la SVD son *n* veces los autovalores del PCA. Este tipo de equivalencia muestra por qué la SVD y el PCA pueden proveer la misma transformación para datos centrados en la media.

En general, también de acuerdo con Aggarwal (2012), el PCA proyecta los datos en un hiperplano de baja dimensionalidad pasándolos a través de la media, mientras que la SVD proyecta los datos sobre un hiperplano de baja dimensionalidad pasándolos a través del origen. PCA captura toda la varianza posible de los datos (o la distancia euclidiana cuadrada a lo largo de la media), mientras que la SVD captura la sumatoria de la distancia euclidiana cuadrada alrededor del origen (Aggarwal, 2012, 43).

La SVD maximiza las distancias euclideas cuadradas agregadas (también denominadas *energía*) de los puntos de datos transformados alrededor del origen. Ahora bien, *maximizar la energía preservada es lo mismo que minimizar el cuadrado de los errores* o la pérdida de energía para la aproximación al rango *k*, los autovectores o vectores propios que indicarán los tópicos más relevantes. Esto es así porque la suma de energía en el subespacio preservado, y la pérdida de energía en el subespacio complementario que ha sido descartado, siempre es una constante, que es igual a la energía del conjunto de datos original *D menos las matrices factorizadas* $Q_k \Sigma_k P_k^T$. En palabras de Aggarwal:

> "Dado que los vectores singulares de la derecha son autovectores de $D^T D$, se sigue de ello que los autovectores (de los vectores singulares de la derecha) con los autovalores *k* más grandes (los valores singulares cuadrados) proveen una base que maximiza la energía preservada en la matriz de datos transformada y reducida $D'_k = DP_k = Q_k \Sigma_k$. Dado que la "energía", que es la suma de distancias euclidianas cuadradas desde el origen, es invariante con respecto a la rotación del eje, la energía en D'_k es la misma que en $D'_k P_k^T = Q_k \Sigma_k P_k^T$. Por esta razón, *SVD de rango k* es una factorización que preserva al máximo la energía. Esto se conoce como el teorema de Eckart-Young". (Aggarwal, Ibíd.).

SVD es más general que PCA y puede ser usado para determinar simultáneamente un subconjunto *k* de vectores de base para la matriz de datos y su transposición con la máxima conservación de "energía". Un ejemplo en donde estas propiedades se aplican pudiera ser aquel en que, en una matriz para establecer un rating que vincule usuarios e ítems, pudiera ser importante determinar una representación reducida de los usuarios y de los ítems. La SVD *truncada*, la técnica que será usada en este trabajo en la aplicación del algoritmo para el análisis semántico latente, expresa los datos en términos de componentes latentes *k*-dominantes (Aggarwal, 2012, 44).

Veamos ahora un pequeño ejemplo de aplicación de las propiedades de la descomposición en valores singulares, que será luego importante para comprender el análisis o indexado semántico latente. El ejemplo de aplicación que ofreceremos ha sido tomado de Leskovec, Rajaraman y Ullman (2016), pero modificado por nosotros para adaptarlo al problema que es objeto del presente trabajo: el modelado de tópicos.

2.2. Ejemplo de aplicación sobre una matriz de datos de texto

De acuerdo con Leskovec, Rajaraman y Ullman (2016), en la SVD una matriz de documentos y términos *m x n*, es factorizada para su representación como el producto de tres matrices diferentes o "latentes".

En efecto, sea la matriz de entrada:

$$A_{mxn} = U_{mxr}\Sigma_{rxr}V_{nxr}^{T} \qquad (2:2)$$

En donde *m* representan las filas de documentos y *n* las columnas de términos, mientras que *r* representarían los conceptos o tópicos (también: factores o dimensiones "latentes"). A_{mxn} es una matriz de valores discretos muy dispersa, en donde los valores no negativos (ó ≥ 0) representan que el término se encuentra presente en el documento.

Decimos entonces que A_{mxn} es el producto de tres matrices $U_{mxr}\Sigma_{rxr}V_{nxr}^{T}$, en donde *U* es de tamaño *m x r* (m documentos y r conceptos o tópicos, en donde podemos pensar que *r* es un número más pequeño que los otros componentes de la matriz de entrada) y es una matriz que comporta los vectores singulares de la izquierda.

Luego tendríamos una matriz cuadrada en donde la diagonal comporta los valores singulares. Es, pues, una matriz de valores singulares *r x r,* Σ. La diagonal está ordenada de manera decreciente y ello expresa la fuerza de cada tópico o concepto. Los valores singulares más grandes van primero y así sucesivamente. *R* designa también el rango de la matriz de entrada A. Finalmente *V* designa los vectores singulares de la derecha y es una matriz *n x r***,** en donde *r* designa los tópicos o conceptos.

"*r*" viene de "rank" (rango de una matriz), que es el número de columnas o filas que son linealmente independientes en una matriz, siendo las otras sumas o "dependientes" de las primeras. La búsqueda del rango de una matriz, pues, equivale a encontrar las columnas o filas "realmente" importantes, las que son, en principio, productos o sumas de otras filas o columnas de una matriz. Ello equivale a conceptos o tópicos subyacentes en una matriz de términos.

Desde un punto de vista intuitivo, esta estrategia puede representarse como sigue.

Supóngase un conjunto de puntos en un plano, que en la Figura 3 se representa en dos dimensiones, cuya agrupación pareciera sugerir una correlación lineal entre las variables dependientes e independientes de una hipotética función $f(x) = y$:

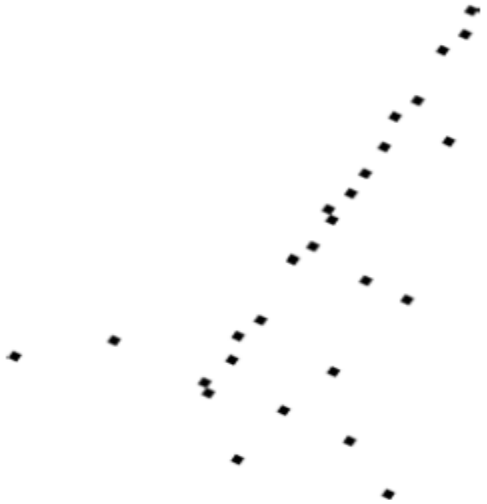


Figura 3: Diagrama de dispersión de conjuntos de datos cuya linealidad es sugerida (Domingos, 2015).

Como se puede observar, la correlación no es del todo lineal: algunos puntos en el plano se encuentran dispersos o alejados de la línea del gráfico que traza la función. La descomposición en valores singulares busca reducir los "errores" o residuales de la función con un algoritmo que, al calcular los vectores y valores propios de la matriz, "rota" o mueve los ejes del plano y captura con ello los datos que expresan una mayor varianza, de modo que pueda representarse el conjunto de datos como una distribución normal. Esta transformación puede representarse como en la Figura 4:

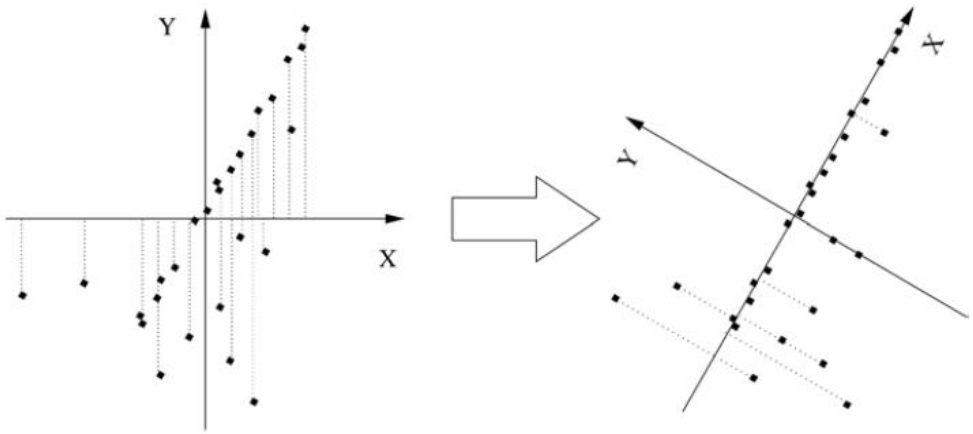


Figura 4: Ejemplo de rotación de los ejes en técnicas de reducción de la dimensionalidad (Domingos, 2015)

Y comporta las siguientes propiedades:

1. Una de las propiedades de esta descomposición es que es única. Para cada matriz A existe una única descomposición $U\Sigma V^T$.
2. U y V son matrices de columnas ortonormales, lo que significa que las columnas de ambas matrices tienen una distancia euclidea de extensión 1, de modo que la suma de los cuadrados de los valores de cada columna es igual a 1. Dos vectores ortonormales tienen, a su vez, la propiedad de formar entre ellos un ápex con un ángulo recto, lo que posibilita la distinción en el plano de los vectores que definen los valores singulares o las propiedades o atributos principales de un conjunto de datos.
3. En tercer lugar, las columnas de U y V son ortogonales, lo que significa que el resultado del producto punto de dos columnas de U o V es igual a 0.
4. $\sum$, por su parte, es una matriz diagonal, en donde las entradas o valores de la diagonal son positivos y están ordenados en orden decreciente: $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_0)$

De este modo, la relación entre estas tres matrices se pueden interpretar como:

1. U: Matriz de similitud de documento en su relación con los tópicos o conceptos.
2. V: Matriz de similitud de términos en su relación con los tópicos o conceptos.
3. Σ: "Fuerza" de cada concepto.

La operación de factorización puede expresarse como sigue:

$$A = U\Sigma V^T = \Sigma_i \sigma_i u_i \text{ ó } \Sigma_i \sigma_i v_i^T \qquad (2:3)$$

En donde σ_i es un escalar, y u_i ó v_i^T son vectores.

Los valores singulares de una matriz A son las raíces cuadradas de los valores propios σ_n de la matriz A^TA.

Encontramos un valor propio de la matriz A si encontramos un valor λ tal que:

$$\det(A^TA - \lambda I) = 0 \qquad (2:4)$$

Se ofrece ahora un ejemplo, modificado de Leskovec, Rajaraman y Ullman (2016) para que aluda al problema de este trabajo, el modelado de tópicos. Sea una matriz A de documentos-términos tal que las columnas representan términos y las filas 7 documentos distintos (Tabla 2):

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	1	1	1	0	0
B	3	3	3	0	0
C	4	4	4	0	0
D	5	5	5	0	0
E	0	2	0	4	4
F	0	0	0	5	5
G	0	1	0	2	2

Tabla 2: Matriz de documentos y términos.

Los documentos sugieren la presencia de dos tópicos o conceptos latentes en el corpus de documentos: "teoría del derecho" y "política", por ejemplo, o tal vez "topografía". Cada uno de los índices representa las veces que el término es mencionado en cada uno de los documentos. Podemos observar que el documento C presenta una preponderancia de ciertos términos ligados a la teoría del derecho sobre los que parecen sugerir el tema de la política o la topografía. Mientras que los documentos E, F y G se inclinan a usar términos ligados a este último tópico.

El cálculo de la matriz *U* tiene como resultado la siguiente tabla, en donde arriba a la izquierda se distinguen los valores propios de los documentos vinculados con temas de teoría del derecho, mientras que abajo a la derecha se resaltan en negrita los valores propios de los documentos vinculados con el otro tópico.

Matriz *U*:

0,13	0,02	-0,01
0,41	0,07	-0,03
0,55	0,09	-0,04
0,68	0,11	-0,05
0,15	**-0,59**	**0,65**
0,07	**-0,73**	**-0,67**
0,07	**-0,29**	**0,32**

Tabla 3: Distinción de autovectores correspondientes a dos tipos de términos.

Esta matriz *U* indica en qué medida *un documento contribuye a un tópico o concepto*, de modo que los valores propios pueden interpretarse de esta manera.

Sigma ∑: En esta matriz se calcula la "fuerza" del concepto "teoría del derecho" (el primero de la diagonal), en contraste con el segundo ("política" o tal vez "topografía") y hay un tercer valor 1,3 que es tan bajo que no es necesario tomarlo en cuenta o resolver la ambigüedad.

Matriz ∑:

12,4	0	0
0	**9,5**	0
0	0	**1,3**

Tabla 4: Matriz diagonal que indica la fuerza de los tópicos.

Y V^T, la tercera matriz, expresa la fuerza de la relación entre términos y tópicos, es decir, la contribución de los términos 'Derecho', 'Humanos' y 'Ley' al tópico "teoría del derecho", mientras que los otros dos "Derecha" e "Izquierda" al tópico o concepto "política" o tal vez "topografía".

Matriz V^T:

0,56	**0,59**	**0,56**	0,09	0,09
0,12	-0,02	0,12	**-0,69**	**-0,69**
0,4	**-0,8**	0,4	0,09	0,09

Tabla 5: Matriz de relación entre términos y tópicos.

¿Qué es lo que "hace" aquí el análisis semántico latente como aplicación de un algoritmo de SVD a un conjunto de datos de texto? Supóngase ahora que se proyecta sobre un espacio semántico latente de dos dimensiones un documento $\boldsymbol{d}$ que tenga 'derecho' entre sus términos, lo que equivale a decir que se ha de encontrar otros documentos que comporten ese término. En el ejemplo, es un hecho que existen solo dos dimensiones relevantes: la del tópico "teoría del derecho" y la del tópico "política" (o tal vez, como se ha dicho, "topografía"). La tercera dimensión dada por la SVD, como ya se ha señalado, arroja un número tan pequeño que es irrelevante.

Se imagina, pues, ese documento como un punto de los datos, o un vector, y se multiplica por cada vector de términos $\boldsymbol{v}_i$, dado que esa multiplicación *transforma* o *mueve* el vector en el espacio de dos dimensiones en la dirección de una u otra dimensión o eje (esto es lo que "hace" básicamente la multiplicación entre vectores y matrices). Supóngase que $\boldsymbol{v}_1$ representa el vector de la matriz V que contiene el término 'derecho'. Al multiplicar dicho vector $\boldsymbol{d}$ por $\boldsymbol{v}_1$, el vector $\boldsymbol{d}$ se moverá en dirección al eje más cercano a los puntos de documentos que exhiben el término 'derecho'. Como $\boldsymbol{v}_2$ es ortogonal con el vector anterior (y forma en el gráfico un ángulo recto con éste), multiplicar el vector del nuevo documento proyecta también $\boldsymbol{d}$ en el espacio latente del término, digamos, 'izquierda', y lo aleja de $\boldsymbol{v}_1$. Ahora bien, recuérdese que el documento $\boldsymbol{d}$ tiene un valor de 1 en $\boldsymbol{v}_1$ y 0 en $\boldsymbol{v}_2$.

Obsérvese gráficamente el movimiento de $\boldsymbol{d}$ al eje en donde domina $\boldsymbol{v}_1$ en virtud de la multiplicación de ambos vectores $\boldsymbol{d}^* \boldsymbol{v}_1$:

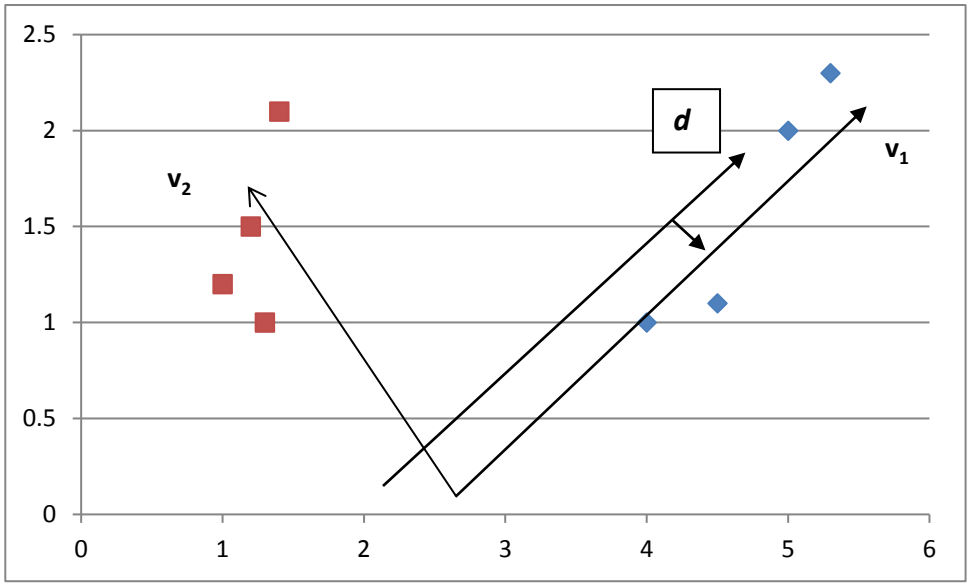


Figura 5: Representación gráfica de la vectorización de dos dimensiones latentes.

En donde la serie romboide (a la derecha del gráfico) representaría, en este ejemplo sencillo, el espacio en donde los documentos comportan el término 'derecho'.

De este modo, sea un vector de documentos:

$$\boldsymbol{d} = [5,0,0,0,0]$$

En donde el índice con el valor 5,$\boldsymbol{v_1}$, representa la presencia cinco veces del término 'derecho' en el documento, el cálculo de SVD obtendrá:

$$\boldsymbol{d} = [5,0,0,0,0] \times \begin{bmatrix} \mathbf{0.56} & 0.12 \\ \mathbf{0.59} & -0.12 \\ \mathbf{0.56} & 0.12 \\ 0.09 & -0.09 \\ 0.09 & -0.09 \end{bmatrix}$$

$$= [\mathbf{2,8}\ \ 0,6] \qquad (2{:}5)$$

Lo que da las coordenadas de ese documento en el espacio de tópicos "teoría del derecho" (resaltado en negrillas) vs "política".

Ahora supóngase que tenemos un segundo documento

$$\boldsymbol{d}' = [0,4,5,0,0]$$

En el que *no* aparece las frecuencias del término 'derecho' pero sí las frecuencias de 'humanos' y 'ley'. Si se hace la misma multiplicación anterior:

$$\boldsymbol{d'} = [0{,}4{,}5{,}0{,}0] \times \begin{bmatrix} \mathbf{0.56} & 0.12 \\ \mathbf{0.59} & -0.12 \\ \mathbf{0.56} & 0.12 \\ 0.09 & -0.09 \\ 0.09 & -0.09 \end{bmatrix}$$

$$= [\mathbf{5}, \mathbf{2}\ \ 0{,}4] \qquad (2{:}6)$$

En virtud de esta similitud de los coeficientes de la operación de SVD, aunque los dos documentos son distintos porque no comparten en común el término 'derecho', sin embargo, *sí pertenecen al mismo grupo que comparte el mismo tópico "teoría del derecho"*. Esta es la fuerza de la descomposición en valores singulares o SVD, que posibilita vincular dos documentos diferentes, **d** y **d'**.

De este modo, básicamente lo que "hace" la SVD es encontrar el mejor vector para proyectar los datos, haciendo lo que hacen todas las técnicas que quieren encontrar ese mejor ajuste en modelos lineales: minimizar la suma del cuadrado de los errores. En el caso del gráfico anterior, el vector $\boldsymbol{v_1}$ representa simplemente la primera fila de la matriz *V*, que correlaciona términos a tópicos o conceptos y que evidencia la mejor línea de ajuste para aquellos términos relacionados con el tópico "teoría del derecho".

Del mismo modo σ_1: (12,4), en la matriz diagonal Σ, recuérdese:

12,4	0	0
0	**9,5**	0
0	0	**1,3**

El eje o vector $\boldsymbol{v_1}$, que es una nueva base o un nuevo eje para la proyección de datos descubierta por la factorización de la matriz original, representa la varianza alrededor de $\boldsymbol{v_1}$, es decir, la dispersión de los datos que exhiben el tópico "teoría del derecho", cuya varianza ha sido, precisamente, capturada por el cálculo de valores y vectores propios que realiza la SVD.

De acuerdo con Aggarwal (2015, 45), en el *análisis o indexado semántico latente* o LSA/LSI la dispersión del conjunto de datos sugiere una dimensionalidad baja implícita. De este modo, LSA puede ejecutar una reducción bastante drástica de la dimensionalidad en los conjuntos de datos de texto. Por ejemplo, un léxico de 100.000 dimensiones pudiera verse reducido a 300 dimensiones.

El LSA ejemplifica el hecho de que perder alguna información a través del descarte de dimensiones resulta en una mejora en la representación de los datos.

El mayor problema de los conjuntos de datos de texto son la polisemia y la sinonimia. La sinonimia implica que dos palabras pudieran compartir un mismo significado, mientras que la polisemia indica que una misma palabra pudiera tener

dos significados diferentes. Por esta razón, dar cuenta de la semántica de una palabra equivale a comprender el contexto en el cual se usa, o, lo que es lo mismo, el sentido de las otras palabras que la acompañan.

De esta manera, una reducción de la dimensionalidad como la que ofrece el LSA puede ser de mucha ayuda, al reducir el ruido que incorporan la polisemia y la sinonimia. Esto se produce porque los vectores singulares en la representación reducida, como vimos en el ejemplo anterior, mantienen toda la "energía" del conjunto de datos original al representar las direcciones de la correlación de los datos de los vectores en el plano y los contextos implícitos que se representan a lo largo de esas direcciones. Las direcciones de "baja energía", o cuya varianza respecto de la media no es significativa, que se truncan en la representación reducida, codifican las diferencias individuales en el uso de una palabra que son sólo ruido (Crain et al., 2012, 134). El SVD es, pues, muy efectivo para la reducción del ruido en bases de datos de texto multidimensionales.

De este modo puede decirse que tanto PCA como SVD son métodos muy efectivos para la reducción del ruido, lo que es posible, a su vez, por la eliminación de vectores propios o vectores singulares de valores pequeños, como se vio en el ejemplo anterior, los cuales, al ser ignorados, implican una mejora en la recuperación de la información relevante de la matriz de datos de texto. Esto es así porque la variación a lo largo de vectores propios de pequeño valor se debe mayormente al ruido. En LSA, la remoción de los componentes pequeños conduce a la mejora de la representación de la semántica de un texto. El beneficio es, de acuerdo con Aggarwal, verdaderamente cualitativo. El LSA o LSI proyecta tanto los términos como los documentos en un espacio semántico latente *K* dimensional.

Ahora bien, uno de los problemas que presenta el análisis de la estadística multivariante "frecuentista" aplicadas sobre este tipo de matrices grandes y dispersas, típicas de los modelos que registran, precisamente, la frecuencia de términos en documentos, es el de que, a medida que pasa el tiempo, se vuelve cada vez más difícil e impráctico actualizarlas cuando se producen cambios en el corpus, bien sea porque se agrega, cambia o se desagrega un documento. En el capítulo siguiente veremos de manera más sistemática esta dificultad, que afecta, en general, a todas las estrategias de carácter "frecuentista" o, dicho de otro modo, que afecta a los métodos de estadística descriptiva que analizan básicamente una sola forma de incertidumbre, a saber, la que concierne a los resultados de la computación.

Los modelos que persiguen el reconocimiento de tópicos latentes modelan o establecen la probabilidad de que un autor utilice un término con base en los tipos de tópicos de los que el documento se ocupa. Mientras que el análisis semántico latente o LSA reduce la dimensionalidad de los documentos proyectando los vectores de una matriz de bolsas de palabras dispersa en un espacio semántico

construido a partir de la descomposición en valores singulares de la matriz términos-documentos original, el modelado probabilístico de tópicos ofrece un mecanismo diferente orientado al reconocimiento explícito de tópicos latentes. A ese tipo de modelado estará dedicado el próximo capítulo.

Capítulo 3. El análisis semántico latente de índole probabilística o PLSA.

3.1. El PLSA también calcula probabilidades a partir de frecuencias de palabras en conjuntos de datos de texto.

Las técnicas de factorización de matrices que extraen valores singulares de las matrices de bolsas de palabras dejan de funcionar bien con conjuntos masivos de datos de textos, en donde los "*outliers*", que en presencia de una matriz de frecuencias acotada o finita se pueden reconocer y dejar de lado, pudieran pasar a tener un peso inesperado en la computación. En este caso, cuando todos los datos en un plano, por el crecimiento exponencial de los datos, se vuelven cada vez más difíciles de describir estadísticamente con una función de regresión o con un análogo de un centrado en la media, -como con los métodos de factorización de matrices-, solo nos queda calcular probabilidades, es decir, tomar decisiones en condiciones de *riesgo*.

La decisión automatizada de una aplicación de modelado de tópicos basada en una técnica de PLSA es, pues: ¿Cuál es la probabilidad de que una palabra observada en un conjunto de datos de textos apunte a un tópico en vez de otro? En el ejemplo anterior de aplicación de LSA, que descubría valores singulares latentes en una matriz de datos, un conjunto de datos perfectamente acotado nos permitía visualizar dos o más vectores ortonormales que distribuían nitídamente las palabras en dos tópicos diferentes.

En una aplicación de PLSA, en donde tenemos muchos datos y tópicos diferentes y un cierto desconocimiento de su atribución a un tópico en vez de otro dado el tamaño de la matriz y su eventual crecimiento exponencial, la apuesta por el cálculo de probabilidades es el camino a seguir.

Por esta razón, en la literatura sobre el modelado de tópicos que emerge a partir de aplicaciones de LSA, comienzan a tomar cada vez mayor relevancia dos nuevos tipos de técnica: el análisis o indexado semántico latente de índole probabilística (PLSA o PLSI, *probabilistic semantic analysis or indexing*, por sus siglas en inglés) y la atribución o colocación latente de Dirichlet (LDA, *latent Dirichlet allocation*). Lo que distingue, en particular, este último enfoque del anterior, el PLSA, que sigue formando parte un paradigma frecuentista de la estadística descriptiva, es la asunción, en la LDA, de un enfoque Bayesiano, enfoque que es el más adecuado en presencia de grandes conjuntos de datos de texto, en donde la tarea de computación pudiera volverse intratable y tablas o matrices de frecuencia pudieran no ser muy confiables para predecir tendencias o resultados futuros.

El PLSA es, ante todo, una continuación del LSA, en el sentido de que atribuye palabras a tópicos o conceptos latentes con base en la frecuencia ponderada de unos términos en algunos documentos en vez de en otros, aunque

interpreta esas frecuencias en términos de *probabilidad*. El PLSA sigue siendo una técnica de estadística descriptiva de tipo frecuentista en la medida en que la probabilidad de que un término forme parte de *clusters* de términos pertenecientes a un tópico o concepto dado depende de parámetros arrojados por un conteo de frecuencias dadas en una matriz de bolsas de palabras basadas en un cálculo de probabilidad multinomial. Este aspecto del PLSA ya apunta a su mayor problema: el del crecimiento exponencial de los parámetros, a medida que aumentan los términos y tópicos posibles en un conjunto de datos de texto, y el riesgo implícito de sobreajuste del conjunto de datos por la función.

Los modelos frecuentistas, en efecto, ofrecen solo parámetros que son válidos para el conjunto de datos que ya se conoce. No comportan, pues, ninguna técnica para encarar o inferir distribuciones de probabilidad desconocidas. En la literatura especializada sobre teoría de las decisiones, como ya se ha señalado en la introducción, los modelos frecuentistas caracterizan más bien la toma de decisiones en situaciones de *riesgo*: dada una tabla de frecuencias y las probabilidades o parámetros asociados a los eventos que ella describe, puede formularse un valor esperado, o probabilidad de un resultado, como guía o parámetro para la toma de decisión (Luce y Raiffa, 1957, 19).

El cálculo de probabilidades en condiciones de *riesgo*, sin embargo, contrasta con el cálculo de probabilidades en condiciones de *incertidumbre*, en donde las probabilidades, o son totalmente desconocidas, o sólo pueden *presumirse* con base en información previa (Luce y Raiffa, 2012, 13). Aquí entrará en juego el paradigma Bayesiano, muy importante en el modelado de tópicos probabilístico que define a la LDA, objeto del capítulo siguiente, que expresa, en el fondo, un cierto grado de creencia subjetiva con respecto al resultado de un ensayo o experimento.

En los cálculos de probabilidades en situaciones de riesgo, el valor esperado o la probabilidad esperada viene dado por una distribución de probabilidad tal que (en donde p designa una probabilidad):

$$E(v) = a_1 p_1 + a_2 p_2 + \cdots a_n p_n \qquad (3{:}1)$$

Que asocia, mutiplicándolos, eventos, jugadas o ensayos con sus distintas probabilidades. La suma de todos los eventos y sus probabilidades ofrece un resultado estimado para la probabilidad de un resultado sobre el cual se apuesta: es decir, una distribución de probabilidad para situaciones de riesgo. Ello abre para el decisor humano o automatizado la posibilidad de calcular la probabilidad más favorable para elegir su estrategia, acción o, como en el caso que nos interesa, su tópico.

3.2. Ejemplo de aplicación sobre una matriz de datos de texto.

De acuerdo con Aggarwal y ChengXiang (2012, 2016), el PLSA concibe un documento como una muestra o mezcla de tópicos diferentes $\{\theta_1, \theta_2, \ldots \theta_k\}$, en donde cada tópico θ_k supone una cierta proporción de palabras que tienen una mayor probabilidad de aparecer en ese tópico en vez de otro.

Recuérdese nuevamente la Tabla 2, el ejemplo usado en el estudio del LSA:

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	1	1	1	0	0
B	3	3	3	0	0
C	4	4	4	0	0
D	5	5	5	0	0
E	0	2	0	4	4
F	0	0	0	5	5
G	0	1	0	2	2

Se puede reinterpretar esta u otra matriz de términos-documentos cualquiera de forma probabilística y suponer que cada documento comporta una proporción de palabras, calculadas con un conteo de frecuencias ponderado (se supone igualmente que en el corpus existen documentos con otras palabras de trasfondo que no se definen como palabras de tópicos y que son comunes a ambos documentos –como las palabras con poco peso semántico o palabras "*stop*"):

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	0,30	0,15	0,10	0	0
B	0,20	0,10	0,30	0	0
C	0,20	0,20	0,40	0	0
D	0,10	0,20	0,10	0,15	0,15
E	0	0,10	0	0,20	0,50
F	0	0	0	0,15	0,25
G	0	0,10	0	0,20	0,30

Tabla 6: Simulación de una interpretación en términos probabilísticos de una tabla de frecuencias para los mismos términos de la Tabla 2. Distribución de probabilidad de palabras en documentos.

Sea la Tabla 6, en la que se simula una distribución de probabilidad cualquiera de los mismos términos en documentos que aparecían en la Tabla 2. Para desarrollar el argumento a continuación, supóngase igualmente que las frecuencias de la Tabla 6 son ligeramente diferentes a las frecuencias de la Tabla 2 y que, por lo tanto, la distribución de la Tabla 6 ofrece parámetros diferentes a los de la tabla anterior: en la Tabla 6 los términos de ambos tópicos presentan una distribución de probabilidad más dispersa y, por ende, más ambigua, entre los dos tópicos. En este sentido, obsérvese, por ejemplo, la distribución posible de los

términos en el documento D de la Tabla 6, en donde se evidencia una distribución de probabilidad previa tanto para términos del tópico "teoría del derecho", como para términos del tópico "política" y, por tanto, es más difícil discernir la atribución del documento D a uno de los dos tópicos del corpus.

El objetivo del PLSA es estimar la distribución de probabilidad multinomial de algunas palabras en un tópico en contraste con el otro, de modo que, por ejemplo, sea posible encontrar una proporción o distribución de probabilidad de tópicos en un documento cualquiera, dados un conjunto de palabras observadas en el mismo, que permita luego clasificarlos como documentos que atañen a uno de los dos tópicos presentes en el ejemplo.

De este modo, si en la Tabla 6, examinamos un documento cualquiera, por ejemplo el D, podremos observar una cierta distribución de probabilidad de que las palabras que conforman el documento sean palabras del tópico "teoría del derecho" (θ_1) o del tópico "política" (θ_2), en donde θ_k es la probabilidad de que una distribución de palabras aluda a un tópico determinado.

Así, si se examina, en la Tabla 6, el documento D, se observará que la distribución de probabilidad de las palabras pareciera sugerir que, en ese documento, es prevalente el tópico "política".

$$\theta_1 = \{'derecho'\ 0{,}1, 'humanos'\ 0{,}2, 'ley'\ 0{,}1\}$$

$$\theta_2 = \{'derecha'\ 0{,}15, 'izquierda' 0{,}15\}$$

Los modelos de PLSA también computan una distribución de probabilidad restante (0,30 para ese documento D), que puede definirse como un "ruido" de fondo o *background*:

$$Trasfondo\ o\ background = \ \theta_B = \{'la'\ 0{,}02, 'de' 0{,}03, \}\ Etc.$$

Este ruido de fondo puede definirse como probabilidades de que palabras "stop" o palabras frecuentes formen parte del documento como, por ejemplo, pronombres, artículos, etc. En general, palabras con poco peso semántico.

En los modelos de PLSA se trata, pues, de "decodificar" o calcular las probabilidades de los distintos tópicos que pudieran estar latentes en un solo documento o en un conjunto de ellos en un corpus de documentos.

En una típica aplicación de PLSA, la entrada o *input* está conformada por los documentos del corpus (C), una distribución inicial o previa de tópicos (k) y el vocabulario presente en los documentos (V), tal que:

$$Entrada = C, k, V$$

Para una salida u output que será la distribución conjunta de la distribución previa de probabilidad de ciertas palabras en los tópicos y la probabilidad o "cobertura" que cada documento ofrezca de las palabras típicas de tópico, tal que:

$$Salida = \{\theta_1, \theta_2, \dots \theta_k\}, \{\pi_{i1}, \dots, \pi_{ik}\}$$

En donde la variable aleatoria π_{ik} representa esa cobertura de palabras (o probabilidad marginal) típicas de tópicos que comporta un documento cualquiera del corpus.

En cálculo de probabilidades, una distribución conjunta es el *producto* de la probabilidad de θ_k, la distribución de probabilidad previa de palabras y tópicos, y de la probabilidad condicional o "*likelihood*" de que una palabra de un documento sea una palabra de tópicos, probabilidad que viene dada calculando la probabilidad que se puede presumir para un documento, estimando sus frecuencias, en relación con probabilidades previas estimadas con frecuencias dadas con anterioridad en ese corpus de documentos.

Esto vuelve posible imaginar que si el algoritmo va iterando a través de distintos productos entre la probabilidad previa y la probabilidad condicional, tal que, por ejemplo:

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
Doc A	$\theta_1$0,02	$\theta_1$0,04	$\theta_1$0,01	$\theta_2$0,0	$\theta_2$0,0
Doc B	$\theta_1$0,0	$\theta_1$0,0	$\theta_1$0,0	$\theta_2$0,02	$\theta_2$0,04
⋮	⋮	⋮	⋮	⋮	⋮
Doc N	$\theta_1$0,0 …	$\theta_1$0,0 …	$\theta_1$0,0 …	$\theta_2$0,0 …	$\theta_2$0,0 …

Tabla 7: Distribución de las probabilidades de que las palabras de los documentos sean palabras típicas de los tópicos *k* que se definen para ese corpus.

(En donde θ_k denota la probabilidad de que la palabra *w* sea una palabra del tópico k (dada por una distribución previa de probabilidades de tópicos en un corpus de documentos) y la probabilidad 0,0… denota la probabilidad de que una palabra *w* del documento sea una palabra que corresponda a esa distribución θ_k).

Entonces el algoritmo de PLSA, como resultado de las distintas distribuciones de probabilidad previas de cada palabra para cada tópico, multiplicadas por las probabilidades de que las palabras típicas estén presentes en el documento, calculadas como distribuciones multinomiales con base en sus frecuencias, iría arrojando, en cada iteración, la distribución conjunta que se define como salida de éste:

$$Salida = \{\theta_1, \theta_2, \dots \theta_k\}, \{\pi_{i1}, \dots, \pi_{ik}\}$$

En donde, como se ha señalado, $\{\pi_{i1}, \dots, \pi_{ik}\}$ indica la "cobertura" probable de un documento de un determinado tópico dentro del corpus, de modo que, por

ejemplo, pudiera darse: $\pi_{i1} = 0{,}4$, $\pi_{i2} = 0{,}3$ y $\pi_{ik} = 0{,}5$ para algún documento d, tal que:

Docs/Topic	θ_1	θ_2	⋯	θ_k
Doc A	$\pi_{A1} = 0{,}4$	$\pi_{A2} = 0{,}1$	⋯	$\pi_{Ak} = 0$ …
Doc B	$\pi_{B1} = 0{,}2$	$\pi_{B2} = 0{,}3$	⋯	$\pi_{Bk} = 0$ …
⋮	⋮	⋮	⋮	⋮
Doc N	$\pi_{ik}0{,}0$ …	$\pi_{ik}0{,}0$ …	⋯	$\pi_{ik} = 0$ …

Tabla 8: Estimado de la cobertura de tópicos en documentos dada la probabilidad ofrecida por frecuencias de las palabras típicas de tópicos.

En donde, para cada documento:

$$\sum_{i=1}^{k} \pi_{d,1} = 1 \qquad (3{:}2)$$

Otra manera de representar la Tabla 7 sería la siguiente: en los sistemas de PLSA, las probabilidades de que una palabra de un documento sea generada por un tópico se actualizan a través de un cálculo de la probabilidad conjunta, en donde las probabilidades de que una palabra probablemente fuera generada por un tópico θ_k (la probabilidad condicional o *likelihood*) se multiplican por la distribución de probabilidad previa de la palabra en el tópico θ_k :

Documentos	'Derecho'	'Humanos'	'Derecha'	'Izquierda'
Doc A	$p(\theta_1)p(w\|\theta_1)$	$p(\theta_1)p(w\|\theta_1)$	⋯	⋯
Doc B	⋯	⋯	$p(\theta_2)p(w\|\theta_2)$	$p(\theta_2)p(w\|\theta_2)$
⋮	⋮	⋮	⋮	⋮
Doc N	$\theta_k 0.0$ …	$\theta_k 0.0$ …	$\theta_k 0.0$ …	$\theta_k 0.0$ …

Tabla 9: Cálculo de la probabilidad conjunta de los tópicos y los documentos.

En donde $p(w|\theta_1)$ es la probabilidad condicional de que una palabra sea una palabra típica del tópico θ_1 en el documento A. La probabilidad condicional requiere una presunción previa sobre la distribución marginal de todos los tópicos que conforman el documento, dado que es una función de la evidencia disponible.

$p(\theta_1)$ es cualquier distribución previa, dada por expertos o por una tabla de frecuencias, de palabras para el tópico θ_1 que se usa para ir actualizando iteraciones de la probabilidad conjunta. El producto de la previa y la probabilidad condicional arrojaría la distribución conjunta $\{\theta_1, \theta_2, \ldots \theta_k\}, \{\pi_{i1}, \ldots, \pi_{ik}\}$ que define, como veíamos arriba, la salida de una aplicación de PLSA. Como resultado, cada documento exhibiría una probabilidad de referirse a cada tópico con base en la cobertura total $\pi_{d,1} = 1$ que sus palabras ofrecen de todos los tópicos cuya distribución previa se presume para cada documento, de manera que:

$$p(\theta_1) = \pi_{d,1} \qquad (3{:}3)$$

Adicionalmente, es necesario definir una variable aleatoria para discernir palabras de tópicos de palabras insignificantes que pertenecen al trasfondo o *background* λ_B.

De este modo, de acuerdo con ChenXiang (2016), la función que define la distribución de probabilidad de una palabra en un documento cualquiera se formula:

$$p_d(w) = \lambda_B\, p(w|\theta_B) + \ (1-\lambda_B)\ \sum_{j=1}^{k} \pi_{d,j}\, p(w\big|\theta_j) \qquad (3{:}4)$$

En donde $\pi_{d,j}$ denota la cobertura de un determinado tópico θ_j en un documento.

La distribución de tópicos por documentos se calcula:

$$\log\ p(d) = \sum_{w\in V} c(w,d) \log\ [\lambda_B p(w|\theta_B) + \ (1-\lambda_B)\ \sum_{j=1}^{k} \pi_{d,j}\, p(w\big|\theta_j)] \qquad (3{:}5)$$

Y la probabilidad conjunta para toda la colección se define como (3:6):

$$\log\ p(C|\Lambda) = \sum_{d\in C}\sum_{w\in V} c(w,d) \log\ [\lambda_B p(w|\theta_B) + \ (1-\lambda_B)\ \sum_{j=1}^{k} \pi_{d,j}\, p(w\big|\theta_j)]$$

El estimado del cálculo de optimización que definiría la máxima probabilidad condicional de todo el modelo o su *MLE* por sus siglas en inglés (*maximum likelihood estimation*) (la distribución probable de las palabras de los documentos dada la distribución de los tópicos) se define como:

$$\Lambda^* = argmax_{\Lambda}\ p(C|\Lambda) \qquad (3{:}7)$$

Bajo las siguientes restricciones:

$$\forall j \in [1,k], \sum_{i=1}^{M} p(w_i\big|\theta_j) = 1 \qquad (3{:}8)$$

Y:

$$\forall d \in C, \sum_{j=1}^{k} \pi_{d,j} = 1 \qquad (3{:}9)$$

Se trata de un problema de optimización porque si el conjunto de datos de textos es muy grande, la probabilidad marginal se procesa como una integral intratable.[5] Por otro lado, si el conjunto de datos no es convexo, como suele

[5] Para entender en qué consiste la distribución marginal de probabilidad o la probabilidad marginal, imaginemos que podemos distribuir en una matriz la probabilidad de *todas* las palabras en todos los tópicos y sumamos cada columna, anotando los resultados de esa suma "al margen" de la matriz, tal que la suma de todas estas probabilidades da 1. Un problema es intratable si los algoritmos que se implentan para su solución tomarían un tiempo excesivo. Los problemas que se pueden resolver en un tiempo razonable son los llamados problemas que se pueden resolver en un tiempo descrito por un polinomio. Por ejemplo, la multiplicación de dos matrices *n x n* se puede resolver utilizando alrededor de n^3 multiplicaciones. Estos problemas pertenecen a la llamada clase

suceder con conjuntos de datos de textos en matrices grandes y dispersas, entonces se hace necesario hacer cálculos para determinar mínimos o máximos locales.

Esta probabilidad condicional se computa con el algoritmo EM (Esperanza-Maximización o *Expectation-Maximization*, por sus siglas en inglés), en donde se trata de definir un conjunto de variables escondidas o latentes *z*, cada una de las cuales sería un índice de tópicos, como puede observarse en la Figura 6.

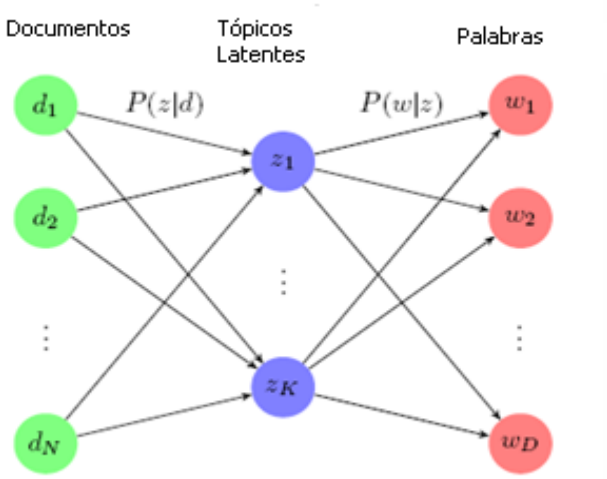


Figura 6: Representación gráfica de un modelo de PLSA (Oneata, 2016). Los documentos están señalados por los nodos en verde, los tópicos latentes por los nodos en azul y las palabras por los nodos en rojo.

Cada *z* denota una palabra típica de tópico tal que:

$$z_{d,w} \in \{B, 1,2, \dots, k\}$$

En el paso E se calcula la probabilidad de que una palabra haya sido generada por el tópico *j*:

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)} p(w|\theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)} p(w|\theta_{j'})} \qquad (3:10)$$

Contra la probabilidad de palabras generadas por tópicos con poco peso semántico del trasfondo:

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)} p(w|\theta_j)} \qquad (3:11)$$

P, por "tiempo polinómico". Por contraste, los problemas que pertenecen a la clase NP, no computables en tiempo polinómico, son aquellos para los que no se conocen un algoritmo polinomial.

En el paso M, se reestima, en primer lugar, la probabilidad de que un documento *d* cubra el tópico θ_j:

$$p\left(\pi_{d,j}^{(n+1)}\right) = \frac{\sum_{w\in V} c(w,d)(1-p\left(z_{d,w}=B\right))p(z_{d,w}=j)}{\sum_{j'}\sum_{w\in V} c(w,d)(1-p\left(z_{d,w}=B\right))p(z_{d,w}=j')} \quad (3:12)$$

Y, en segundo lugar, la probabilidad de que una palabra *w* sea una palabra del tópico θ_j:

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d\in C} c(w,d)(1-p\left(z_{d,w}=B\right))p(z_{d,w}=j)}{\sum_{w'\in V}\sum_{d\in C} c(w',d)(1-p\left(z_{d,w'}=B\right))p(z_{d,w'}=j)} \quad (3:13)$$

El proceso de computación del algoritmo EM procede como sigue:

1. En primer lugar, se inicializa de modo aleatorio todos los parámetros desconocidos y

2. Se itera el algoritmo hasta que las probabilidades condicionales convergen, es decir, cuando la computadora arroja probabilidades condicionales que cada vez son más parecidas a las de las anteriores iteraciones.

De este modo:

En el paso E:

$$p\left(z_{d,w}=j\right) \propto \pi_{d,j}^{(n)} p^{(n)} p(w|\theta_j)$$

$$p\left(z_{d,w}=B\right) \propto \lambda_B p(w|\theta_B) \quad (3:14)$$

En donde:

$$\sum_{j=1}^{k} p\left(z_{d,w}=j\right) = 1$$

Y el paso M:

$$\pi_{d,j}^{(n+1)} \propto \sum_{\mathrm{w}\in\mathrm{V}} \mathrm{c(w,d)(1-p(}\left(z_{d,w}=B\right))p\left(z_{d,w}=j\right)$$

$$\mathrm{p^{(n+1)}} \propto \left(w|\theta_j\right) \sum_{\mathrm{d}\in\mathrm{C}} \mathrm{c(w,d)(1-p(}\left(z_{d,w}=B\right)\right) p\left(z_{d,w}=j\right) \quad (3:15)$$

Con la siguiente restricción:

$$\forall j \in [1,k], \sum_{w\in V}^{M} p(w|\theta_j) = 1$$

Ahora bien, de acuerdo con Crain et al. (2012, 41), *maximizar la probabilidad condicional en sistemas de PLSA es lo mismo que minimizar la divergencia de Kullback-Leibler*, un método de evaluación de la distancia o de la entropía relativa entre dos distribuciones de probabilidad, una aproximada y una "real", que ofrece una medida de ganancia o pérdida de información. La divergencia KL entre dos funciones de densidad de la probabilidad evalúa la distancia entre la distribución medida empíricamente o aproximada, y la distribución "verdadera". Por esta razón, el PLSA utiliza básicamente un método para la minimización de la divergencia KL.

De este modo, el cálculo de optimización de la probabilidad condicional se lleva a cabo con un proceso que estima la medida de distancia denominada la divergencia de Kullback-Leibler.

La Figura 7 ofrece una representación intuitiva de la divergencia de Kullback-Leibler:

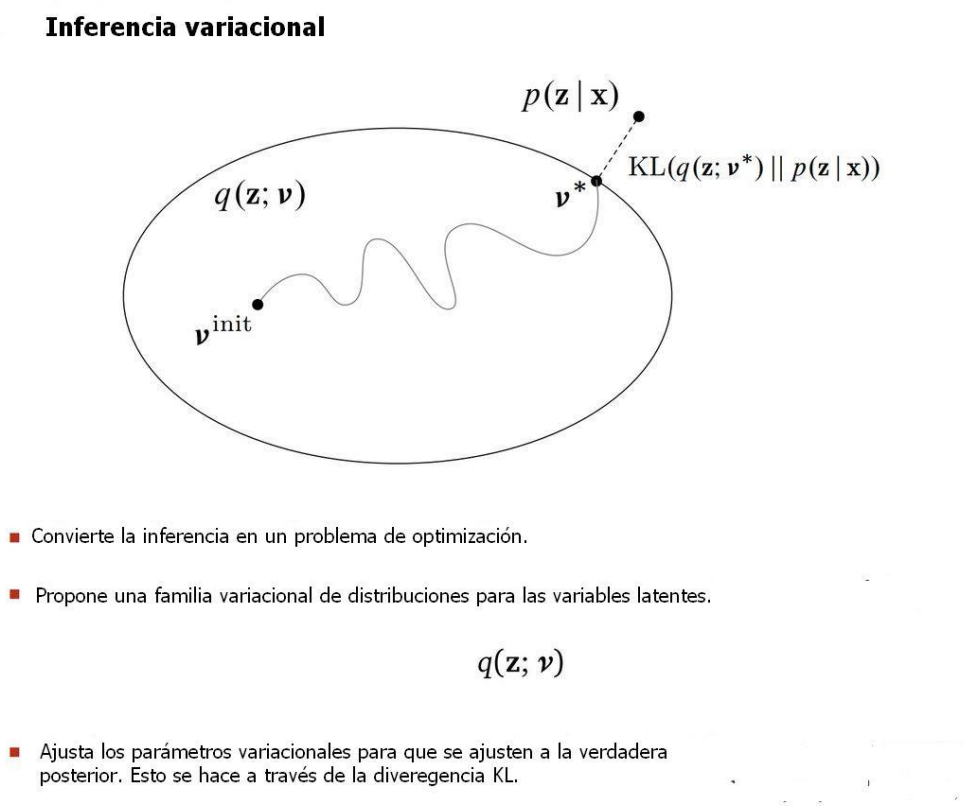


Figura 7: Representación gráfica de la divergencia de Kullback-Leibler (adaptado de Blei, 2017 b)

En donde en cada punto del espacio de puntos representado por la elipse hay una distribución diferente de tópicos sobre palabras. Se empieza con alguna particular realización de la distribución o con una propuesta de distribución (ν^{init}),

que se va ajustando u optimizando hasta encontrar ν^*, una distribución que está más cerca de la probabilidad que se busca.

La divergencia de Kullback-Leiber es un método de evaluación de la distancia de la entropía relativa que ofrece una medida de ganancia o pérdida de información entre dos funciones de densidad de la probabilidad, en donde se trata de acercar una función "proxy" a la "verdadera" o última distribución de probabilidad en la convergencia del modelo.

No obstante, los modelos de PLSA comportan, como se ha señalado, un riesgo importante de sobreajuste y no son muy eficientes como proveedores de cálculos de probabilidad de la distribución de tópicos en corpora de documentos. En el siguiente capítulo, cuando se examinen los modelos que aplican LDA, veremos con más detalle estas limitaciones del PLSA, al contrastarlas con el modelo más eficiente, la LDA, una aplicación más efectiva para el cálculo de probabilidad al modelado de tópicos.

Capítulo 4. La atribución latente de Dirichlet (LDA).

El siguiente capítulo está dedicado a la tercera y última técnica de modelado de tópicos: la atribución o colocación latente de Dirichlet o LDA, *latent Dirichlet allocation*.

Lo que distingue, como ya se ha señalado más arriba, los modelos probabilísticos que se apoyan en probabilidades basadas en un conteo de frecuencias de las aplicaciones para modelado de tópicos que utilizan la atribución latente de Dirichlet es la capacidad que exhibe esta última para inferir, no simplemente *la probabilidad condicional de las palabras dados los tópicos en un corpus de documentos (o su MLE), sino distribuciones de probabilidad basadas en cantidades desconocidas, las cuales resultan más apropiadas para calcular probabilidades en conjuntos de datos de texto que, o bien son muy grandes, o bien crecen exponencialmente*. Por esta razón, los modelos de LDA no solo calculan una máxima distribución condicional de la probabilidad de palabras dados tópicos en una colección de documentos que se conoce, y cuya frecuencias de palabras se ha contado (el *MLE*), sino también el así llamado *maximum a posteriori* (o *MAP*), es decir, el cálculo de las distribuciones de la probabilidad posterior (futura o desconocida), en conjuntos de datos que crecen o pudieran crecer exponencialmente.

La LDA supone, pues, una implementación completa de la regla de Bayes en sistemas de modelado probabilístico de tópicos y, por esta razón, posibilitan modelos de decisión sobre tópicos caracterizados no solo por el *riesgo*, sino también por la *incertidumbre*. Con la LDA, el modelado de tópicos se separa realmente del enfoque frecuentista y asume plenamente su carácter de *inferencia probabilística*. Esto hace que, de modo creciente en la literatura del modelado de tópicos, toda la atención se haya ido volcando sobre los modelos de LDA, en desmedro de los modelos de PLSA, que van siendo poco a poco dejados de lado por la literatura especializada.

De este modo, el modelado de tópicos de índole probabilística que *infiere* probabilidades *en condiciones de incertidumbre*, es decir, sobre cantidades desconocidas o cantidades sobre las que ya no se tiene un conteo de frecuencias confiable, va poco a poco monopolizando el interés de los especialistas en el análisis de conjuntos de datos complejos, dada la creciente proliferación de datos que hace posible la Web. Los conjuntos de datos de texto que derivan en la Web, por ejemplo, no pueden ser objeto de un simple enfoque basado en cálculos de probabilidades multinomiales que se apoyen en conteos de frecuencias, simplemente porque un número enorme de nuevos documentos introducirán sesgos en los parámetros que se alejan de aquellos que definían la matriz de términos-documentos inicial.

Así pues, puede decirse que la LDA calcula probabilidades de tópicos a través de un uso completo de la *inferencia Bayesiana*.

La inferencia bayesiana computa la probabilidad posterior conforme al teorema de Bayes, el cual se denota como:

$$p(h|e) = \frac{p(e|h)p(h)}{p(e)} \qquad (4:1)$$

En donde $\frac{p(e|h)}{p(e)}$ se interpreta como el impacto de la evidencia sobre la probabilidad de la hipótesis o *probabilidad condicional*.

→La raya vertical significa "ocurrencia o probabilidad de un evento condicionado por": la probabilidad de que algo ocurrirá cuando se produce otro evento.

→ En tareas de modelado de tópicos, como ya podía observarse en el capítulo anterior, un modelo probabilístico es la distribución conjunta de variables ocultas $\boldsymbol{z}$ y variables observadas $\boldsymbol{x}$, en donde $\boldsymbol{z}$ designa palabras de tópicos o conceptos latentes u "ocultos": $p(\boldsymbol{x}, \boldsymbol{z})$. ($P(\boldsymbol{x}, \boldsymbol{z})$ designa la probabilidad conjunta entre *x* y *z*).

→ En modelos de LDA, las inferencias sobre distribuciones de probabilidad dadas las variables ocultas se realizan como presunciones *posteriores* de probabilidad. La distribución de probabilidad posterior de las variables ocultas dadas las variables observadas viene dada por la siguiente función, en donde la distribución *conjunta* se divide por la probabilidad marginal o la "evidencia":

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{z},\boldsymbol{x})}{p(\boldsymbol{x})} \qquad (4: 2)$$

Es decir, la probabilidad de que la variable oculta *z* ocurra dada la variable observada *x*. Veíamos en el capítulo anterior que la distribución conjunta es un producto de la probabilidad *condicional* de las palabras dados los tópicos $p(\boldsymbol{x}|\boldsymbol{z})$ multiplicada por la probabilidad previa de los tópicos $p(\boldsymbol{z})$. Pero el cálculo de la posterior implica algo más que conocer una distribución conjunta. Como se desprende de la ecuación (4:2) la probabilidad posterior requiere conocer la distribución de la probabilidad marginal $p(\boldsymbol{x})$, lo cual es imposible o muy difícil cuando estamos ante conjuntos de datos que son enormes o crecen exponencialmente. Esto es lo que hace necesario, en los modelos de LDA, que la probabilidad posterior sea objeto de un cálculo de maximum a posteriori o *MAP*, es decir, que no sea *suficiente*, como en los modelos de PLSA, el estimado de la probabilidad condicional o *MLE*.

→En efecto, en la mayoría de los modelos actuales de análisis de tópicos latentes, el denominador *p(x)* o *p(e)* es, de hecho, "intratable" o insoluble, porque resulta cada vez más costoso computacionalmente el cálculo de la distribución de probabilidad marginal de las palabras y los tópicos. La inferencia de la posterior

sólo puede ser aproximada con distintos métodos.[6] Esto es lo que limita, en último término, los modelos frecuentistas como el LSA, o los modelos que infieren probabilidades a partir de contar frecuencias, como el PLSA: la creciente incertidumbre sobre los datos que conforman la evidencia y, por lo tanto, la dificultad de estimar de modo confiable los parámetros probabilísticos.

A medida que crece el corpus, el número total de sus palabras, tópicos ocultos y documentos, la computación de la inferencia Bayesiana se convierte en un problema de cálculo de optimización de funciones, en donde la probabilidad se calcula con una función de densidad para definir el área del gráfico que abarca los datos. En la Figura 8 se ofrece un ejemplo para ilustrar distintas funciones que definen el cálculo de las probabilidades que caracterizan a la regla de Bayes. Obsérvese que el cálculo de la marginal es una integral:

Previa	Condicional	Conjunta	Posterior
$f_1(x)$	$f_2(y\|x)$	$f_3(x,y)=f_1(x)f_2(y\|x)$	$f_3(x\|y)=f_3(x,y)/f_4(y)$
		$f_4(y) = \int f_3(x,y)dx$	
		Marginal	

Figura 8: Un ejemplo de funciones de densidad de la probabilidad derivadas de la regla de Bayes, en donde *x* representa la previa e *y* la evidencia.

Que el cálculo de la posterior exija, primero, elevar una presunción de probabilidad para futuras iteraciones del algoritmo de modelado probabilístico de datos con base en probabilidades condicionales calculadas para conjuntos de datos anteriores y, segundo, que la posterior exija, a su vez, un cálculo de optimización sobre máximos que solamente pueden ser locales, es lo que hace que los modelos de LDA sean sumamente difíciles e intrincados desde el punto de vista conceptual. Comportan un componente subjetivo que resulta polémico a muchos autores, dado no sólo por la distribución inicial de la previa, que se apoya en una opinión facilitada por expertos sobre una posible distribución de tópicos en documentos, sino también porque el cálculo de la probabilidad posterior, que ofrece presunciones racionales basadas en el teorema de Bayes sobre distribuciones futuras, lo hace iterando la regla de Bayes con base en probabilidades posteriores que pasan a definir la previa de la siguiente iteración. Como es evidente, sucesivas iteraciones de las distribuciones de la probabilidad posterior, en tanto que previas de la siguiente iteración del algoritmo, pudieran sesgar el análisis de los datos, por pérdida de los parámetros, alejándolos de una evidencia que ya no es completa. El LDA ofrece, así, una solución ingeniosa, pero siempre polémica, al problema de calcular la probabilidad de tópicos en corpora de documentos cuya probabilidad marginal se desconoce.

[6]Véase nota 5.

Para entender esto, obsérvese que el paradigma Bayesiano asume siempre, como punto de partida, una distribución de probabilidad previa de los tópicos en los documentos: $p(T)$, en donde T denota los tópicos. Al inicio, la previa puede ser más bien vaga: una presunción de probabilidad de la proporción de tópicos por documento, dada, por lo general, por el curador del conjunto de datos o un experto. En la regla de Bayes, como veremos en la ecuación 4:3, la previa se multiplica por la distribución condicional o distribución de la muestra, la "*likelihood*", a fin de obtener una distribución conjunta. La "*likelihood*" o función de esperanza de la condicional, o simplemente probabilidad condicional, es, por lo general, el resultado de un cálculo de tipo frecuentista, como vimos en el capítulo anterior. Da cuenta del conocimiento actual sobre probabilidades reales en conjuntos de datos, las cuales derivan de lo que se sabe respecto de sus frecuencias en una distribución de probabilidad marginal.

La probabilidad condicional expresa cuán probable es la distribución de probabilidad de los datos observados, dados los parámetros (la media, la desviación estándar y, en general, la distribución de la probabilidad de todos los eventos o casos) del modelo. La probabilidad condicional favorece parámetros que explican realmente los datos observados, los datos de la muestra, o los hacen verosímiles, plausibles.

En el enfoque de la inferencia Bayesiana, que calcula una distribución de probabilidad posterior, una distribución de probabilidad previa se multiplica por la distribución condicional, que se apoya en una muestra de los datos observados, y luego este resultado se normaliza por la distribución de probabilidad marginal, es decir, la sumatoria igual a 1 (recordemos que se llama "marginal" porque antiguamente se sumaba "en el margen" de la matriz de datos) de todas las variables aleatorias de la distribución de probabilidad que abarca la evidencia completa. Para una distribución posterior (predictiva):

$$p(T \mid P) = \frac{p(P|T)p(T)}{p(P)} \qquad (4:3)$$

En donde T denota los tópicos de un corpus y P todas las palabras de ese corpus.

En el modelo anterior, el PLSA, pudo observarse que este cálculo ofrece una distribución conjunta de todos los tópicos y su proporción en cada documento de un corpus, que se obtenía con el algoritmo EM, que estima precisamente probabilidades condicionales a través de un cálculo de optimización de la *MLE*. No obstante, como recordará el lector, este tipo de cálculo requiere que podamos tener un estimado confiable de las frecuencias de todas las palabras de un corpus de documentos, es decir, que tengamos la esperanza de un dominio completo de la probabilidad marginal, lo cual es perfectamente posible en un corpus acotado de documentos cuyas frecuencias es posible, en principio, contar. Por esta razón, el

PLSA era un método, en el fondo, frecuentista: calcula probabilidades con base en frecuencias.

La inferencia Bayesiana predictiva, por el contrario, que ofrece, a la par de una probabilidad *condicional* sobre la evidencia disponible $p(P|T)$, un cálculo de la probabilidad *posterior* (como se expresa en la ecuación 4:3: $p(T|P)$), es el método por excelencia de inferencia estadística no frecuentista o que lidia con incertidumbre. El teorema de Bayes se usa, entonces, para estimar la probabilidad posterior de una hipótesis a medida que más información o más evidencia se vuelve disponible y poder utilizarla como previa $p(T)$ para la siguiente aplicación en el conjunto de datos.

La posterior debe ser entendida como una consecuencia de los tres antecedentes ya mencionados en 4:3: una probabilidad previa y una función de probabilidad condicional o *likelihood* que, como función de la evidencia, puede computarse si y solo si se tiene la esperanza de un dominio completo sobre la probabilidad marginal, es decir, sobre el conjunto total de palabras que conforman el corpus. Hasta aquí, un modelo de LDA no se distingue de un modelo de PLSA. El problema surge cuando la posterior se usa como una previa en sucesivas iteraciones. Si el conjunto de datos crece exponencialmente, como es el caso en conjuntos de datos de texto que nacen en la Web, o si es enorme, como los documentos que conforman 100 años de la revista Science estudiados por Blei y sus colaboradores, disponer de las posteriores como si fueran previas es inútil si no se cumplen dos condiciones: si no podemos calcular de modo fiable la probabilidad marginal, que ofrece un modelo estadístico del conjunto de datos, (como se veía más arriba: un estimado de la probabilidad total en la muestra disponible), y, en segundo lugar, si no se conjuga la previa.

Por estas razones, David Blei define las funciones que calculan la probabilidad marginal en los métodos probabilísticos de modelado de tópicos como "integrales intratables", dado que no es posible observarla o tratarla en conjuntos de datos masivos o conjuntos de datos que crecen exponencialmente.

Ello vuelve necesario apelar a métodos de *inferencia variacional* para resolver el problema del cálculo de la posterior en distribuciones que forman parte de familias exponenciales y que, por ello, no pueden dominarse totalmente con métodos frecuentistas. Los métodos de inferencia variacional generalizan, para un conjunto de datos que no se puede dominar en su totalidad, los métodos de inferencia de la posterior, o *MAP*, un método de optimización que sustituye la necesidad de computar toda la evidencia (una tarea imposible) y conocer todos los parámetros para los puntos del espacio de datos.

Optimizar una función de probabilidad condicional (*MLE*) o de probabilidad posterior (*MAP*), significa simplemente esto: si tenemos una función cualquiera $y = f(x)$, optimizar la función es encontrar el punto de la pendiente que es

"estacionario" o "crítico", es decir, en donde la tangente del gradiente es igual a 0, como puede observarse en la Figura 9:

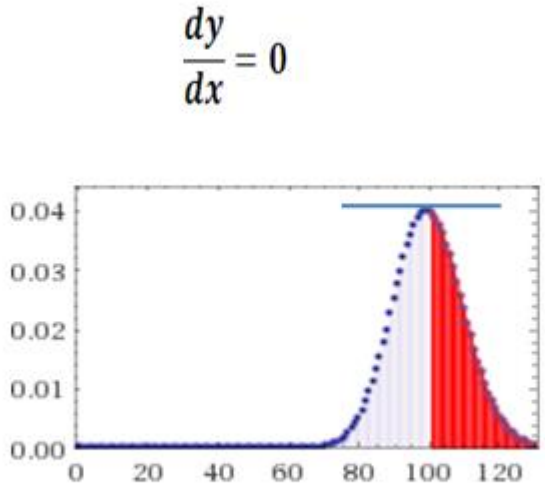


Figura 9: Ejemplo gráfico de optimización con un punto crítico.

La LDA o atribución latente de Dirichlet ofrece una alternativa al problema planteado por distintas iteraciones de la previa que se alejan de la evidencia ofrecida por el conjunto de datos, proponiendo una técnica matemática, la distribución de Dirichlet, que calcula *distribuciones de probabilidad sobre distribuciones de probabilidad posibles*. La LDA aplica este tipo de técnica matemática a la búsqueda de tópicos latentes.

La propiedad matemática que define a la LDA como modelo que computa estimados de probabilidad posterior y los utiliza como previas en la siguiente iteración es su capacidad para *conjugar* esas posteriores como previas. Cuando nos vamos alejando de la evidencia disponible y ya no conocemos cómo están distribuidas las probabilidades de los tópicos en los documentos, la capacidad de la LDA para conjugar las previas o distribuciones de probabilidad en familias exponenciales permite a este tipo de modelo proponer parámetros posibles, lo cual es importante si se quiere evitar que se modele como una distribución gaussiana lo que es una distribución multinomial, por ejemplo. La previa conjugada permite que la posterior de la cual deriva sea del *mismo tipo* de distribución de probabilidad que esta última.

De este modo, la LDA es un modelo de inferencia Bayesiana que no se limita a la optimización de la probabilidad condicional o *MLE* de los parámetros del modelo, como sucede con el PLSA, sino que incorpora también un cálculo del la

posterior (*MAP*), la probabilidad predictiva que funcionará como probabilidad previa en iteraciones sucesivas, *para distintas distribuciones de probabilidad de tópicos por documentos*. Para ello, incorpora al cálculo de la probabilidad, además de los parámetros que vimos en el modelo anterior, un nuevo parámetro α, también llamado un "hiperparámetro", el cual agrega una dimensión adicional a la tarea de cálculo de la distribución de probabilidad de tópicos por corpus al calcular *una distribución de probabilidad sobre distribuciones posibles de probabilidad* de tópicos sobre documentos dados en el corpus. El hiperparámetro α ofrece la posibilidad de calcular distribuciones de probabilidad sobre previas conjugadas como *distribuciones multinomiales*, el tipo de cálculo de probabilidad que caracteriza a los conjuntos de datos de texto multidimensionales.

Al igual que en el PLSA, los tópicos son distribuciones de probabilidad sobre palabras. Es decir, cada documento distribuye los tópicos con una distribución multinomial y cada tópico comporta, a su vez, una distribución multinomial sobre palabras. A diferencia del PLSA, sin embargo, la LDA calcula aproximaciones a la distribución de probabilidad *posterior* de todos los tópicos dado un corpus de documentos, optimizando el mejor estimado para la distribución de probabilidad de las *previas* de todas las distribuciones de probabilidad posibles de tópicos sobre los documentos. Esta posibilidad la hace posible la incorporación al modelo de cálculo de dos hiperparámetros típicos de las distribuciones de Dirichlet, los hiperparámetros α y η.

En modelos de LDA su punto de partida es análogo a cualquier cálculo de probabilidad de tópicos en corpora de documentos. En primer lugar, elige una distribución posible sobre tópicos en un documento, una previa. Aquí, a manera de ilustración, obsérvese la Tabla 10, en donde se definen los tópicos con la letra griega β para armonizarla con los ejemplos que dan David Blei y sus colaboradores, suponiendo que la distribución inicial de tópicos es realizada con conocimiento experto, en donde $\beta_1 =$ "teoría del derecho" y $\beta_2 =$ "política", y en donde, como se recordará, π designa la "cobertura" o probabilidad marginal de palabras de un tópico en un documento:

Docs/Topic	β_1	β_2	⋯	β_k
Doc A	$\pi_{A1} = 0{,}4$	$\pi_{A2} = 0{,}1$	⋯	$\pi_{Ak} = 0 \ldots$
Doc B	$\pi_{B1} = 0{,}2$	$\pi_{B2} = 0{,}3$	⋯	$\pi_{Bk} = 0 \ldots$
⋮	⋮	⋮	⋮	⋮
Doc N	$\pi_{ik}0{,}0 \ldots$	$\pi_{ik}0{,}0 \ldots$	⋯	$\pi_{ik} = 0 \ldots$

Tabla 10: Coberturas de palabras de tópicos por documentos.

Después del proceso de distribución de tópicos en un documento, seguidamente se calcula para cada palabra la probabilidad de que pertenezca a un tópico de alguna distribución posible (Tabla 11).

	β_1	β_2	⋯				⋯	β_k
w_1	$z_{d,n}, w_{d,n}$	⋯						⋮
w_2	⋮							
⋮								
w_N								

Tabla 11: Cálculo de la probabilidad de que una palabra sea una palabra de tópico, en donde *w* denota la palabra observada (*word*) y *z* es el índice que denota una palabra de tópico.

Este proceso se hace para cada documento. Cada documento exhibe una distribución de tópicos que es distinta a cualquier otro. Pero, adicionalmente, como ya se ha señalado, se postula un nuevo parámetro α, definido por distribuciones posibles de distribuciones posibles de tópicos por documentos, los cuales se mezclarán con procedimientos aleatorios, estocásticos o de inferencia variacional para una mejor aproximación a la probabilidad posterior que servirá de previa en la siguiente iteración del algoritmo.

El modelo gráfico del LDA, de acuerdo con Blei (2012) es el siguiente, en donde cada elemento es una variable aleatoria (Figura 10). Obsérvese la presencia de los hiperparámetros *alpha* y *eta*, que permiten un cálculo sobre distribuciones de distribuciones de probabilidad de tópicos en un documento:

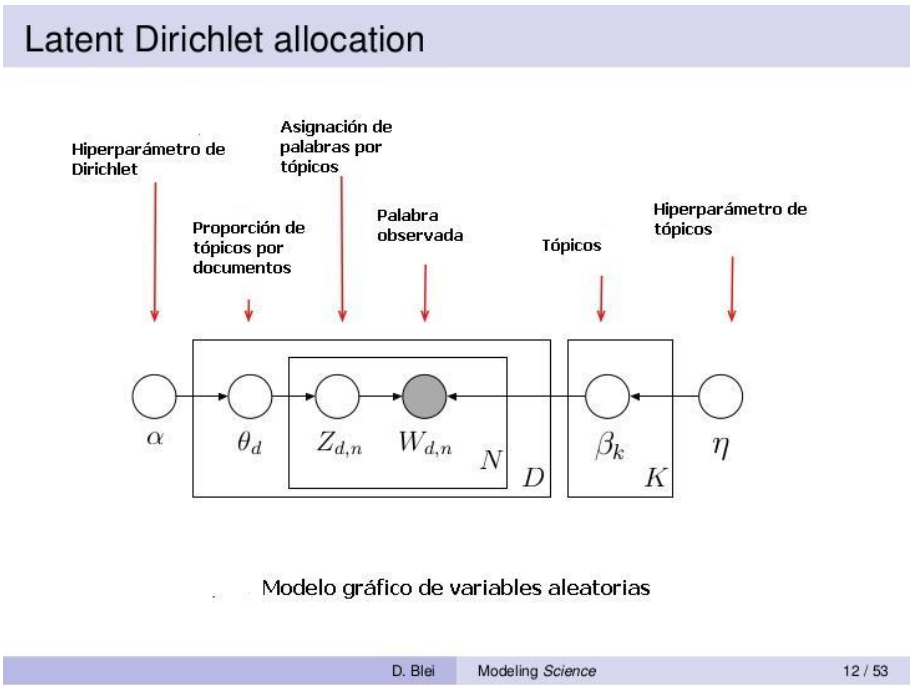


Figura 10: Modelo gráfico de la LDA en donde puede observarse los parámetros alpha y eta (adaptado de Blei, 2012, 81)

→En ese modelo, θ_d representa la distribución de los tópicos por documento y hay una variable θ_d para *cada* documento.

→$w_{d,n}$representa la palabra observada. Está en gris porque es lo único que se observa o es un dato evidente en un documento. $w_{d,n}$depende tanto de $z_{d,n}$ como de β_k.

→β_k es alguna distribución de palabras en tópicos, es decir, es algún tópico con una determinada distribución de probabilidad sobre sus palabras. Hay un número k de tópicos, que pudieran ser 100, y β_{87}, por ejemplo, es alguna distribución de probabilidad de palabras para un tópico.

→En el gráfico anterior el recuadro *D representa* el *corpus*, en donde, como veíamos, θ_d representa la distribución de los tópicos por documento, o, lo que es lo mismo, la proporción de tópicos por documento.

→El recuadro *N* representa *cada* palabra dentro del *corpus*.

→$z_{d,n}$es la variable aleatoria que representa la asignación de tópicos para cada palabra, es decir, representa la probabilidad de que un tópico dado forme parte de un documento dada la palabra que indexa. Depende de θ_d porque depende de los parámetros arrojados por la distribución de cada documento. Si parece que θ_d exhibe la probabilidad de que se ocupe de los tópicos x,y,z, entonces $z_{d,n}$ pudiera ser uno de ellos. Por otro lado hay una variable $z_{d,n}$para cada palabra.

→El parámetro que designa la *proporción* posible de tópicos en un documento es α, un hiperparámetro, mientras que η es el hiperparámetro de de tópicos posibles.

En la LDA, al inicio, lo único que vemos, en efecto, es un conjunto de palabras desordenadas en un documento. $w_{d,n}$ ofrece la probabilidad de una palabra dado z y β_k. Es decir, en la distribución conjunta que se muestra en la expresión,

$$p(w_{d,n}|z_{d,n},\beta_k) = \beta_{z_{d,n},w_{d,n}} \qquad (4:4)$$

En donde se trata de calcular la probabilidad de ver la palabra dado un tópico y dada la probabilidad de ver esa palabra bajo ese tópico. *Z* es un índice de β_k.

En el modelo gráfico puede observarse una distribución conjunta que define la siguiente probabilidad posterior y que se será transportada por el algoritmo a la siguiente iteración a través del parámetro α:

$$p(\theta, z, \beta|w) \qquad (4:5)$$

Pero para comprender cómo lo hace, es necesario examinar, primero, las propiedades formales o matemáticas de las distribuciones de Dirichlet.

4.1. Propiedades formales de las distribuciones de probabilidad de Dirichlet.

Una distribución de Dirichlet es un tipo de distribución que modela de modo eficiente distribuciones sobre distribuciones multinomiales. Como es evidente, calcular la proporción de tópicos en un corpus de documentos requerirá lidiar con tipos de distribución multinomiales, una generalización de distribuciones binomiales o de Bernoulli. Calcular, como hace una distribución de Dirichlet, distribuciones multinomiales sobre distribuciones multinomiales permite el cálculo de la posterior a través de la actualización en sucesivas iteraciones del parámetro *alpha*, que conjuga la previa de iteraciones anteriores.

De este modo, a diferencia del PLSA, la LDA, al calcular distribuciones de probabilidad sobre distribuciones multinomiales de probabilidad, ofrece un método mucho más flexible para el modelado de tópicos de nuevos documentos o documentos no observados. En este sentido, es mejor que el análisis latente de índole probabilística. En este último, se necesita entrenar de nuevo todo el corpus al incorporar nuevos documentos y, por lo tanto, nuevos tópicos. El tener que entrenar todo el modelo de nuevo supone el crecimiento lineal de los parámetros de la distribución de probabilidad y el peligro de sobreajuste del modelo, con la consiguiente deficiencia en su capacidad de generalización. En la LDA no se necesita entrenar cada corpus de nuevo, dado que cada documento presenta distintas distribuciones de probabilidad que son parametrizadas de manera independiente.

Téngase presente que un parámetro estadístico o un parámetro de población es un estimado que permite inferir la distribución de probabilidad de la población de la que se toma la muestra y que se especifica con base en uno o varios valores de la muestra, incluyendo valores aleatorios.

En tareas de inferencia estadística Bayesiana, como los que atañen a modelos de LDA, los parámetros no pueden observarse, de modo que lo que se trata es de *inferir* todo lo que se pueda de una variable aleatoria de la muestra, o de distintas variables aleatorias, para saber cómo está distribuida la población.

Entre los tipos de parámetros más importantes se encuentran los *parámetros de forma* o de figura de un gráfico en una función de densidad de la probabilidad. Este tipo de parámetros son característicos de las así llamadas distribuciones Beta y definen o parametrizan la forma de la distribución de una función de probabilidad que no puede modelarse acabadamente porque no se abarca toda la probabilidad marginal o toda la población.

Los parámetros de la forma de una distribución estiman los máximos de una distribución que no se conoce: en particular el *MLE* y *MAP*. Son parámetros adecuados para conjuntos de datos masivos en donde no se conoce la distribución de probabilidad de la población total y hay que estarla actualizando

continuamente. En contraste, esto no es necesario cuando se sabe que una población puede ser modelada con una distribución normal, por ejemplo, o una uniforme, en donde la media y la varianza se apoyan en parámetros constantes.

Los parámetros son particularmente importantes en tareas de estimado de la probabilidad que involucran tareas de cálculo. Es decir, cuando hay que inferir la forma de la distribución y no es posible modelarla a priori con una función (cosa que podemos hacer es un estimado de la velocidad de un vehículo, que "sabemos" que puede modelarse con una función sigmoide, por ejemplo).

Precisamente por esto, la distribución Beta es adecuada para modelar la probabilidad de que un vehículo espacial complete su misión en el espacio, en donde, como es obvio, no es posible suponer que su trayectoria puede modelarse con la ya mencionada sigmoide.

Su función de densidad de probabilidad es la siguiente (Cfr. Johnson y Beverlin, 2013, 1):

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad (4:6)$$

En donde la función Beta es igual a una razón de funciones Gamma (Ibíd, p. 2):

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad (4:7)$$

Y *x* es un parámetro para una distribución de probabilidad tal que $x \in [0,1]$.

Las distribuciones Beta también son adecuadas para modelar el comportamiento aleatorio de proporciones y, por lo tanto, apropiadas en el modelado de las proporciones de tópicos en documentos.

La otra propiedad importante de las distribuciones Beta, asociada a la anterior, es su capacidad para modelar el comportamiento de variables aleatorias delimitadas por intervalos de extensión finita. De nuevo, los fenómenos que estudiamos, los tópicos que caracterizan un corpus que no se domina en su totalidad, requieren el estimado de parámetros que se calculan con integrales llamadas "impropias".

La distribución Dirichlet es precisamente una generalización multivariante de las distribuciones Beta.

La capacidad de las distribuciones Beta para modelar intervalos de extensión finita permite a los cálculos de optimización en modelos probabilísticos la *conjugación* de la previa, en el parámetro α, para las sucesivas iteraciones del algoritmo, de modo que la posterior de la inferencia bayesiana sobre distribuciones multinomiales de tópicos por documentos sea también una

distribución multinomial de "posteriores" posible. Así como la distribución Beta es la previa conjugada de distribuciones binomiales o distribuciones de Bernoulli, la distribución Dirichlet es la previa conjugada de distribuciones multinomiales, en donde las distribuciones de múltiples variables (o múltiples tópicos, que es lo que nos interesa) se parametrizan en un vector α de valores reales positivos.

Supóngase, en efecto, que el vector de parámetros $\theta = (\theta_1, \dots \theta_n)$ es una distribución de Dirichlet con parámetros α, tal que $\theta = Dir(\alpha)$.[7] Y en donde el vector de parámetros de α es un vector de números reales estrictamente positivo: $\alpha = (\alpha_1, \dots \alpha_n) > 0$.

Entonces la función de densidad de la probabilidad de θ es proporcional a 1 sobre una distribución Beta generalizada de α multiplicada por los productos de i hasta n de $\theta_i^{\alpha_i - 1}$, multiplicados por la función característica que indica que θ es un miembro del simplex de probabilidad (Cfr. Frigyik, Kapila y Gupta, 2010, p. 7):

$$p(\theta) \propto \frac{1}{B(\alpha)} \prod_{j=1}^{n} \theta_i^{\alpha_i - 1} \; I(\theta \in S) \qquad (4{:}8)$$

En donde S es el conjunto de números reales tal que la suma $\sum_{i=1}^{n} = 1$.

La función o distribución $B(\alpha)$ es una distribución análoga a una distribución multinomial tal que:

$$\frac{1}{B(\alpha)} = \frac{\Gamma\alpha_0}{\Gamma\alpha_1 \dots \Gamma\alpha_n} \qquad (4{:}9)$$

En donde α_0 es la suma de todos los alphas: $\alpha_0 = \Sigma\alpha_i$

Ya se ha señalado que las distribuciones multinomiales de probabilidad se definen como generalizaciones de distribuciones binomiales. Véase, como ilustración de este punto, el siguiente ejemplo de una distribución multinomial que modela tópicos. Sea un corpus con 20 documentos, de los cuales 13 expresan el tópico "teoría del derecho" y 7 el tópico "política" y se quiere saber cuál es la probabilidad de que 6 nuevos documentos sean de uno o de otro. En este caso, los resultados posibles no son 2, como en una distribución binomial, en donde sólo existe la probabilidad de que el evento tenga o no tenga lugar, sino 12, dado que se tiene un número d de documentos y k de tópicos cuyos resultados son dos: éxito o fracaso, sí o no, es decir: si expresan el tópico "teoría del derecho" o no, o el tópico "política" o no. La probabilidad inicial *P(d)* viene dada por la muestra en la que "se sabe" (es una presunción previa de probabilidad), por los datos observados o por opinión o conocimiento experto, que, en 20 documentos, 13 expresan un tópico y 7 el otro. La distribución de probabilidad multinomial viene dada por:

[7] Cfr. "Dirichlet Distribution", disponible en https://www.youtube.com/watch?v=nfBNOWv1pgE. Cfr. Frigyik, Kapila y Gupta, 2010.

$$P(X_i = x_i, \dots X_k = x_k) = \frac{n!}{x_1!,\dots,x_k!} p_1^{x^1} \dots p_k^{x^k} \qquad (4:10)$$

En el ejemplo, se tendría, pues, si nos preguntamos cuál es la probabilidad de que, de 6 nuevos documentos, 4 sean sobre teoría del derecho y 2 sobre política:

$$P(X_1 = 4, X_2 = 2) = \frac{6!}{4!\,2!}\left(\frac{13}{20}\right)^4 \left(\frac{7}{20}\right)^2$$

$$= 0{,}32800 \qquad (4:11)$$

Es decir, hay un 32% o un 0,32 probabilidades de que esos seis tópicos presenten una distribución de probabilidad de 13/20=0,65 para un tópico y 7/20=0,35 para el otro.

En este ejemplo, la distribución multinomial deriva de un número n de ensayos independientes, que resultan en un número k de salidas mutuamente independientes y en cada ensayo las salidas k ocurren con probabilidades $(p_1, \dots p_k)$ cuya suma es igual a 1: $\sum_{i=1}^{k} p_i = 1$.[8]

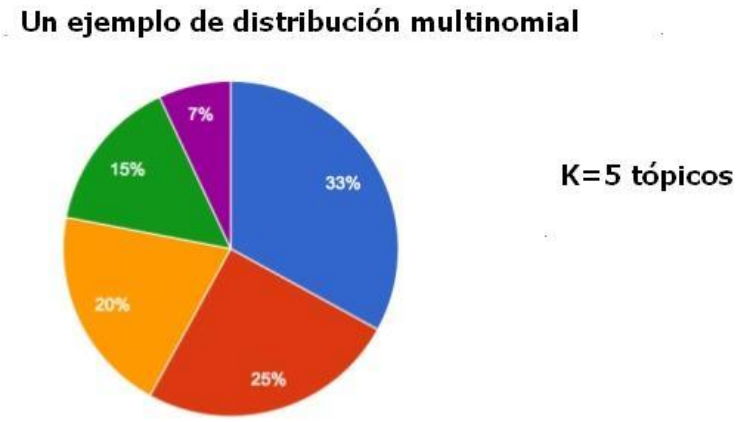


Figura 11: Gráfico de una distribución multinomial (adaptado de Sklar, 2014)

Obsérvese ahora en la Figura 11 un ejemplo gráfico de distribución multinomial con 5 tópicos o categorías. Aquí, un documento cualquiera pudiera exhibir la distribución de tópicos que se muestra en este gráfico de torta.

El punto de partida de una distribución multinomial, los datos crudos o desestructurados, son tablas de frecuencia o matrices de bolsas de palabras. Por ejemplo, filas de documentos y columnas de tópicos o columnas con palabras, cada una de las cuales distribuye una probabilidad de que en un tópico se mencionen las palabras o tópicos definidos en el encabezado de la columna. Luego las

[8]Cfr. "Introduction to the Multinomial Distribution" en el canal Jbstatistic, disponible en https://www.youtube.com/watch?v=syVW7DgvUaY. Véase también Frigyik, Kapila y Gupta (2010, 5).

probabilidades de que las palabras de un documento apunten a un tópico en vez de a otro se van distribuyendo o parametrizando con un cálculo como el que se acaba de ofrecer. Se puede suponer que distintas distribuciones de tópicos por documentos irán emergiendo poco a poco a medida que se itera el algoritmo. La clave de las distribuciones de Dirichlet es su capacidad para modelar como distribuciones multinomiales, en una previa conjugada, distintos vectores de parámetros para los documentos del corpus.

De este modo, con una tabla de frecuencias o de bolsas de palabras podemos calcular la probabilidad condicional o *likelihood* de las palabras dados los tópicos si se toman todos los valores aleatorios de las columnas a lo largo de una fila, tal que la sumatoria sea igual a 1, y se multiplican por el número *k* de columnas, categorías y tópicos (la distribución previa de la probabilidad). Con el estimado de la probabilidad condicional se designa la probabilidad de que la evidencia (las frecuencias) sea observada en las condiciones descritas por la matriz, es decir, bajo el número *k* de categorías en las que se ha dividido la distribución de tópicos de un documento. Recordemos que, en la regla de Bayes, ese estimado debe normalizarse o dividirse por la probabilidad marginal o evidencia, el número total de palabras distribuidas por tópicos en el corpus, para calcular una distribución de probabilidad posterior. No obstante, es necesario recordar que el cálculo de la condicional requiere el dominio de la probabilidad marginal, es decir, la distribución total de variables aleatorias dadas la evidencia, es decir, las palabras observadas en el tópico.

Las distribuciones de Dirichlet, al conjugar la previa para subsiguientes iteraciones de los algoritmos que calculan la distribución de probabilidad multinomial, permiten actualizar cada vez la distribución de cada tópico para el corpus de documentos. Por lo tanto, es agnóstica respecto de cuál es la distribución de probabilidad definitiva para el corpus, como puede verse ilustrado en la Figura 12.

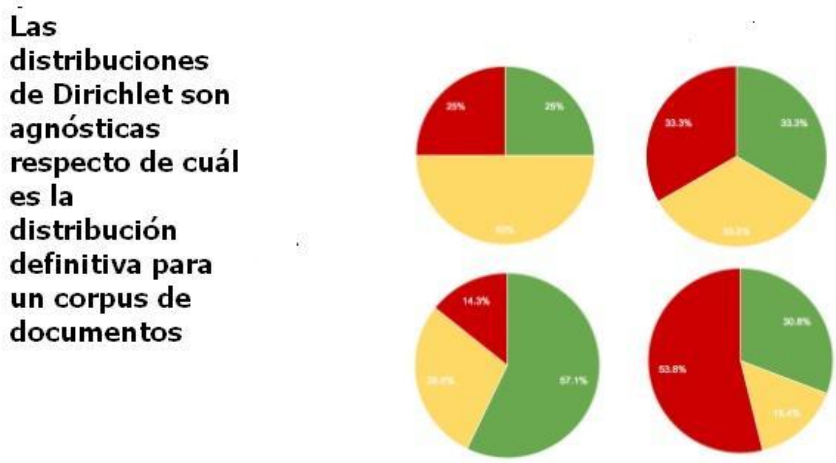


Figura 12: Distribuciones multinomiales. Adaptado de Sklar, 2014

Las distribuciones de Dirichlet representan, como dice Max Sklar (2014), nuestra incertidumbre respecto de cuál es la "verdadera" distribución de probabilidad de un corpus.

Como se puede observar en la Figura 13, el proceso de actualización, que aquí se expresa de modo intuitivo, requiere ir modificando en iteraciones sucesivas la distribución de probabilidad previa, es decir, la distribución de palabras en los tópicos a medida que van ingresando palabras representativas de tópicos:

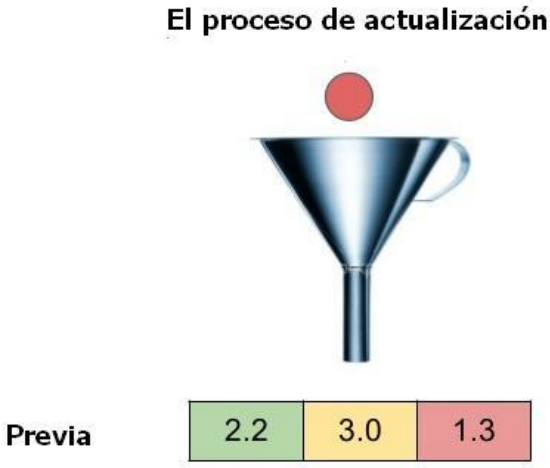


Figura 13: Actualización de la distribución previa de la probabilidad (adaptado de Sklar, 2014)

La actualización tendría como resultado una modificación de la previa o la probabilidad marginal de los tópicos, como se puede observar en la Figura 14, que expresa este proceso de modo intuitivo:

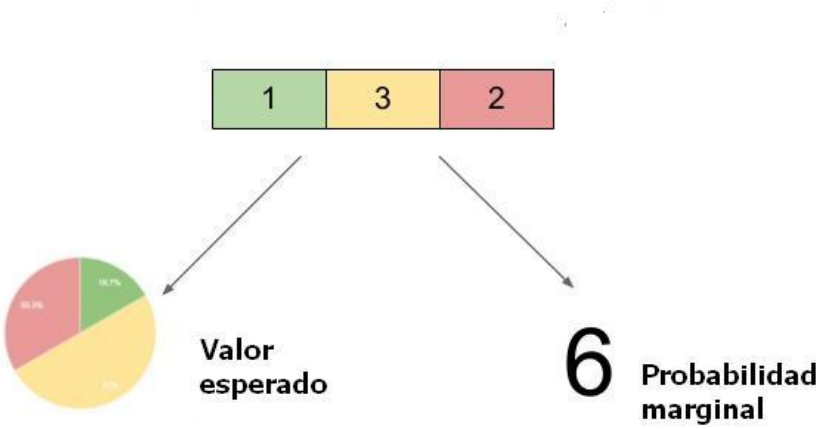


Figura 14: Representación gráfica de la actualización de la previa (adaptado de Sklar, 2014).

Así, si representamos una distribución de Dirichlet como una regla de Bayes tendríamos:

$$p(T \mid P) = \frac{p(P|T)p(T|\alpha)}{p(P)} \qquad (4.12)$$

En donde α es el hiperparámetro que distribuye, en un vector, las palabras en tópicos. El resultado es una distribución de Dirichlet, es decir, una actualización del vector de parámetros α.

El proceso de actualización del hiperparámetro α puede observarse en la Figura 15, en donde se expresa de manera intuitiva que también la distribución de distribuciones puede representarse como una distribución multinomial:

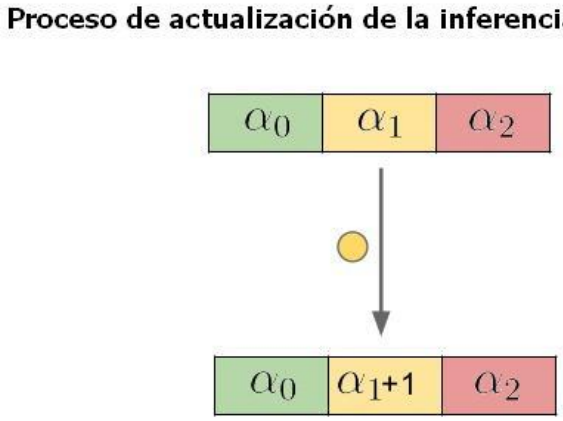


Figura 15: Representación gráfica de la actualización del hiperparámetro α (adaptado de Sklar 2014).

En el caso de las distribuciones de Dirichlet, mientras más precisa sea la probabilidad marginal y, con ella, la distribución de las previas, por supuesto,

mejor es la distribución de probabilidad de todos los tópicos en los documentos, del mismo modo que, en las distribuciones normales, mientras más cerca está la varianza de la media, más precisa es la distribución normal. Por eso, el algoritmo actualiza la distribución de la probabilidad marginal de una distribución previa con un vector de pesos. El sopesado de la distribución marginal es crucial: si no se tiene cuidado con esto, se corre el peligro de caer bajo el así llamado "efecto de Mateo" para la ley de potencias: que los términos más frecuentes terminarán sesgando la constitución de un tópico o la definición de un concepto y, con ello, se perderá una estimación precisa de los parámetros del modelo.[9]

La distribución de Dirichlet está definida por la función de densidad (Cfr. Sklar, 2014 y Frigyik et al., 2010, 12):

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=0}^{K-1} \alpha_k)}{\prod_{k=0}^{K-1} \Gamma(\alpha_k)} \prod_{k=0}^{K-1} \theta_k^{\alpha_k - 1} \qquad (4{:}13)$$

Desde el punto de vista intuitivo, la distribución Dirichlet puede tener muchas formas: parecer una distribución normal (en lo que se refiere al conjunto *S*), o ser uniforme, o exhibir kurtosis o toda clase de formas extrañas.

La Figura 16 ofrece de modo ilustrativo un conjunto de gráficos que representan, en planos de tres dimensiones, distintas funciones de densidad de la probabilidad de distribuciones de Dirichlet. Los números alrededor de las "montañas" pueden verse como probabilidades de tópicos:

[9]En efecto de Mateo dice que tener un alto valor de una propiedad hace que esa propiedad crezca: "los ricos se vuelven más ricos" o, como dice el versículo de Mateo que da nombre a este efecto: "Pues al que tiene se le dará más, y se le dará bastante; pero al que no tiene, hasta lo poco que tiene se le quitará" (Mateo, 13,12). Para un análisis del efecto de Mateo como interpretación de la ley de potencias véase Leskovec, Jure, Anand Rajaraman y Jeffrey Ullman, *Mining of Masive Data Sets*, 2010. También Merton, "The Matthew Effect in Science", *Science* 159(3810):56-63, January 5, 1968, disponible en http://www.garfield.library.upenn.edu/merton/matthew1.pdf.

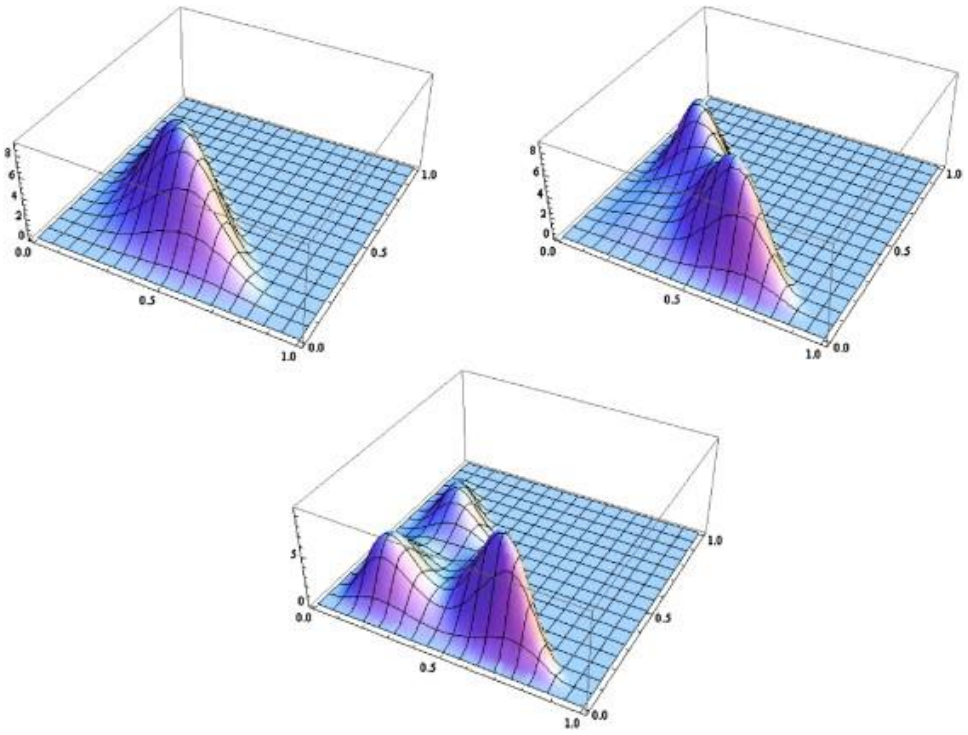


Figura 16: Representación gráfica de distribuciones de Dirichlet para tres valores de *alpha* distintos (Ongaro y Migliorati, 2012, 415).

Así, se puede interpretar cada "montaña" como una distribución de tópicos (parametrizados por un vector α) para *cada* documento del corpus.

4.2. Ejemplo de aplicación sobre un conjunto de datos de texto.

Supongamos que tenemos un documento cualquiera. Ese documento, como es natural, exhibirá múltiples tópicos. En el ejemplo que proporciona Blei (en 2012), tenemos un documento tomado de la revista Science que discurre sobre el número de genes que necesita un organismo para sobrevivir evolutivamente. Un experto, puede, simplemente a mano, resaltar, en diferentes colores, palabras que parecen apuntar a distintos tópicos: azul claro para "análisis de datos", amarillo para palabras que parecen apuntar al tópico "genética" y así sucesivamente. La idea es, desde luego, poder hacer este trabajo de modo automático, anotar un documento de modo automático.

De este modo, se concibe cada documento como si comportara una distribución de tópicos y cada tópico como una distribución sobre términos. En efecto:

→Cada tópico es una distribución sobre palabras.

→Cada documento es una mezcla de un número de tópicos que se considera presente en todo el corpus o tiene una distribución determinada sobre esos tópicos.

→Cada palabra se atribuye a alguno de esos tópicos.

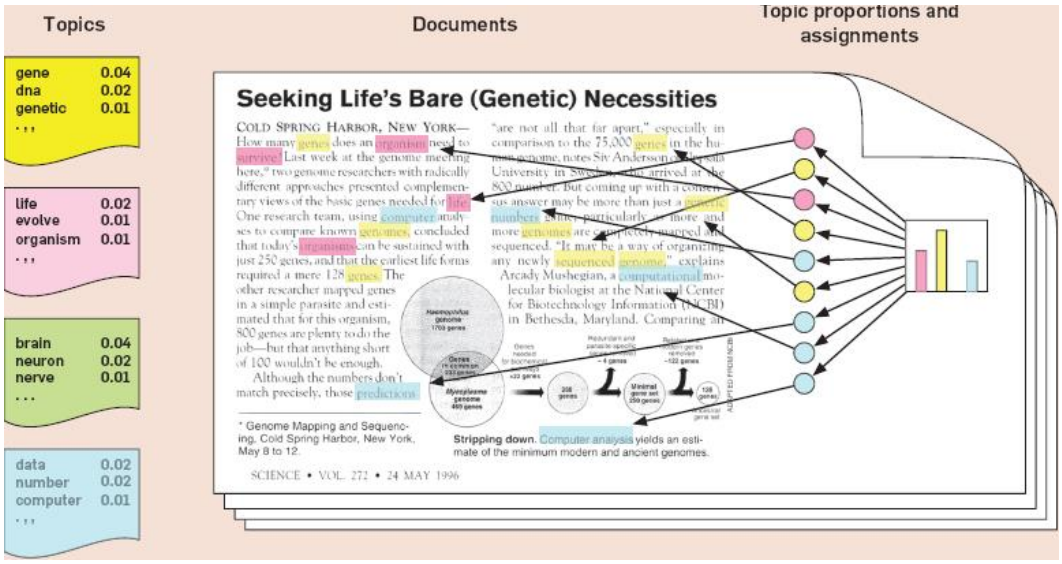


Figura 17: Ejemplo de etiquetado experto de la distribución de probabilidad previa dado por Blei, 2012, 78.

De este modo, un documento como el de la Figura 17 exige que el modelo generado por la LDA elija una distribución sobre esos tópicos y, además, asigne cada palabra a un tópico. Hasta aquí, el proceso no se distingue de un modelo de PLSA. Adicionalmente, sin embargo, la LDA genera cada documento como una *mezcla de distribuciones multinomiales,* como ya hemos señalado.

Se elige así una distribución sobre tópicos, luego para cada palabra se calcula la probabilidad de que pertenezca a un tópico de esa distribución, y luego se busca la distribución de palabras en el tópico asociada a la palabra anterior. Este proceso se hace para cada documento. Pero, al exhibir cada documento una distribución de tópicos que es distinta a cualquier otro, se posibilita la aplicación de una distribución multinomial de Dirichlet sobre los tópicos mismos a través del hiperparámetro.

El hiperparámetro α es el parámetro que designa la previa de Dirichlet o la distribución de tópicos en cada documento, que se conjuga como una previa para la siguiente iteración. Un α con un valor alto, de acuerdo con Sullivan (2017), indica que cada documento probablemente comporta una mezcla de la mayoría de los tópicos y no se restringe a uno o dos en particular. Un α con un valor bajo indica que cada documento exhibe una mezcla de pocos tópicos. β es el parámetro que designa la distribución de palabras por tópicos. Un β alto indica que un tópico tiene probablemente una mezcla de la mayoría de las palabras por documento. Un β bajo indica que el tópico esta conformado por una mezcla de pocas palabras.

Como ya se observó arriba, θ es la distribución de tópicos para cada documento. Distintas distribuciones de θs, en los modelos de LDA, se conjugarán como previas sobre las que se realizarán inferencias variacionales en iteraciones

de la posterior bayesiana. Como ya hemos observado también, el parámetro z denota una palabra de tópico.

De este modo, los parámetros del modelo son:

α es el hiperparámetro de la previa de Dirichlet que indica las distribuciones de tópicos por documentos.

β es el parámetro que indica la distribución de palabras por tópicos.

θ_m es la distribución de tópicos por documentos m.

z_{mn} es el tópico para la enésima palabra del documento *m.*

w_{mn} es la palabra observada.

Un modelo de LDA :

1. Cuenta las frecuencias de palabras en un documento.
2. Elige (por opinión experta y sucesivamente con previas "ruidosas") una mezlca de tópicos para cada documento de un conjunto prefijado de tópicos, tales como β_1 ="teoría del derecho", con un conjunto de palabras con probabilidades distintas y $\beta_2 =$ "política", con un conjunto de palabras con sus distribuciones aleatorias.
3. Se van actualizando la distribución de palabras de tópicos en los tópicos a través de, en primer lugar, la elección de un tópico desde la distribución multinomial de tópicos inicial (como la dada arriba) y, en segundo lugar, eligiendo una palabra de otra distribución multinomial de palabras por cada tópico. El algoritmo va iterando sobre las distribuciones multinomiales hasta que las palabras convergen en la previa especificada.

Recuérdese ahora la Tabla 2, que se reinterpretó en el capítulo anterior en términos probabilísticos, aunque cambiando ligeramente sus parámetros (Tabla 6):

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	1	1	1	0	0
B	3	3	3	0	0
C	4	4	4	0	0
D	5	5	5	0	0
E	0	2	0	4	4
F	0	0	0	5	5
G	0	1	0	2	2

Documentos	'Derecho'	'Humanos'	'Ley'	'Derecha'	'Izquierda'
A	0,33	0,33	0,33	0	0
B	0,40	0,30	0,30	0	0
C	0,20	0,40	0,40	0	0
D	0,10	0,20	0,40	0,15	0,15
E	0	0,10	0	0,50	0,40
F	0	0	0	0,55	0,45
G	0	0,20	0	0,50	0,30

En donde se tenía un conjunto de documentos cada uno de los cuales estaba conformado principalmente por dos tópicos:

β_1 ="teoría del derecho"

β_2 ="política"

En donde cada uno de los tópicos puede ser descrito con las siguientes palabras probables z:

'Derecho', 'humanos' y 'ley' para el tópico β_1

'Derecha' e 'izquierda' para el tópico β_2

Cada uno de las cuales comporta una probabilidad z_{mn} de ser una palabra de alguno de los dos tópicos.

De este modo, el algoritmo de LDA generará, cada vez, un nuevo documento cuya distribución previa por tópicos estará dada por el parámetro α, tal que, supóngase:

Documentos	$Dir\ (\alpha_1)$	$Dir\ (\alpha_2)$	⋯	$Dir\ (\alpha_n)$
θ_1	$\beta_1 = 0{,}55;\ \beta_2 = 0{,}45$	$\beta_1 = 0{,}58;\ \beta_2 = 0{,}42$	⋯	⋮
θ_2				
⋮	⋮			
θ_M				

Tabla 12: Distribuciones de probabilidad del hiperparámetro α

En donde contamos las frecuencias de palabras por documento, generamos un tópico basados en una distribución multinomial de las palabras en tópicos, generamos una distribución de tópicos en documentos basados en una distribución de Dirichlet previa y, finalmente, elegimos una palabra basados en la distribución previa de cada palabra por tópico, tal y como se ejemplificaba en la Tabla 11 que veíamos arriba:

	β_1	β_2	$\cdots$				$\cdots$	β_k
w_1	$z_{d,n}, w_{d,n}$	$\cdots$						$\vdots$
w_2	$\vdots$							
$\vdots$								
w_N								

Como se puede observar en la anterior Tabla 12, arriba, cada $Dir\ (\alpha_n)$ es una distribución posible de tópicos por documentos, sobre el que se va iterando en sucesivas jugadas de distribuciones multinomiales de probabilidad. El hiperparámetro que lleva la proporción de tópicos de los documentos es α. El hiperparámetro α controla la media y la dispersión de θ y es, como se ha señalado, el parámetro que designa la proporción de tópicos en un documento y conjuga, como previas, el cálculo de la probabilidad posterior de las distintas proporciones para sucesivas iteraciones.

Desde el punto de vista del funcionamiento del algoritmo, la LDA

1. Se inicia atribuyendo de manera aleatoria cada palabra de los documentos a cada tópico.

2. Para cada documento *d*:

- Asume que todas las asignaciones de sus palabras son correctas, excepto la actual.
- Calcula dos probabilidades: la proporción de palabras al tópico asignado en la previa de distribuciones de tópicos para el documento, tal que:

$$\beta_1 = p(\beta_1|\alpha)$$

Y se va actualizando la probabilidad de que la palabra *w* sea una palabra del tópico β_k, tal que:

$$z_{mn} = p(w_{mn}|z_{mn})$$

- Seguidamente, multiplica estas dos probabilidades o proporciones para asignar las palabras a los tópicos hasta convergencia, es decir, cuando los tópicos se estabilizan. Por ejemplo:

	w_1	w_2	$\cdots$			$\cdots$	w_n
θ_1	$\beta_1 z_{d,n}$	$\beta_2 z_{d,n}$	$\cdots$				$\beta_k z_{d,n}$
θ_2	$\vdots$						$\vdots$
$\vdots$							
θ_M							

Tabla 13: Cálculo de la probabilidad de las palabras típicas de tópicos en los documentos

El resultado puede visualizarse en una salida posible de un algoritmo de LDA, como la que se muestra en la Tabla 14, en donde las palabras de un corpus de documentos terminan convergiendo alrededor de tópicos que tienen sentido para el investigador, que luego los interpreta.

Tópico # 1		Tópico # 2	
'Derecho'	0,30	'Derecha'	0,15
'Humanos'	0,05	'Izquierda'	0,08
'Ley'	0,15	'…'	0,0…
'…'	0,0…		

Tabla 14: Salida posible del algoritmo de LDA

Desde el punto de vista de sus desventajas, de acuerdo con Crain et al. (2012), la LDA tiende a aprender tópicos más generales o tal vez demasiado amplios. Por esta razón, los tópicos generados por el LDA pueden volverse cada vez más difusos. Si un concepto, en efecto, comporta una serie de atributos o dimensiones y cada uno de ellos ocurre con frecuencia como un atributo de su tópico principal, la LDA tenderá a favorecer ese tópico y sus atributos, así como tenderá a subsumir atributos de otros tópicos que parecen ser instancias del concepto más general. Esta tendencia puede contrarrestarse, de acuerdo con los autores citados, con la implementación de un modelo de tópicos jerárquico.

El gran desafío de los modelos de LDA, como ya se ha señalado, es la elección del método para calcular, tanto la distribución condicional, como la distribución de probabilidad posterior. Se trata de un problema de cálculo en donde, dados la dispersión y el carácter masivo del conjunto de datos, resulta muy difícil calcular la integral de la función que ajusta los datos en el hiperplano. El enfoque probabilístico es solo una ayuda para calcular la probabilidad de que una función ajuste el conjunto de datos y, por ello, es un enfoque conceptualmente distinto al de la estadística descriptiva, en donde los "*outliers*", en un conjunto de datos pequeño y acotado, pueden dejarse fácilmente de lado sin perjuicio de la función, por decirlo así, que luego permite factorizar la matriz, como hemos visto.

Acá no: los *outliers* en un conjunto masivo de datos, por su propia cantidad, dificultarán la tarea de saber si son "*outliers*" o definen más bien una variable esencial al modelo. De este modo, es necesario incorporarlos a éste y calcular su

capacidad explicativa para generar los tópicos de los documentos, es decir, su distribución de probabilidad.

Para la importante tarea de calcular la posterior hay, pues, muchos métodos disponibles. La posterior *exacta* es, en efecto, muy complicada de computar o directamente intratable, pero algunas aproximaciones a la posterior pudieran ser de ayuda. Algunas técnicas de las mencionadas por Blei son:

→*Mean field variational methods* (métodos variacionales de campo medio) (Blei et al., 2006, 2011).

→*Expectation propagation* (propagación del valor esperado) (Minka, Lafferty, 2002).

→*Collapsed Gibbs sampling* (muestreo de Gibbs colapsado)(Griffiths and Steyvers, 2004).

→*Distributed sampling* (muestreo distribuido) (Newman et al., 2009: Ahmed et al. 2012).

→*Collapsed variational inference* (inferencia variacional colapsada) (Teh et al., 2006)

→*Stochastic Inference* (inferencia estocástica) (Hoffman et al., 2013; Mimno et al., 2012).

→*Factorization inference* (inferencia de factorizaciones) (Arora et al.,2013).

→*Amortized inference* (inferencia amortizada) (Srivastava and Sutton, 2016) (Cfr. Blei, 2017 a).

Para este propósito, el modelado de tópicos adapta distintas distribuciones alternativas para aproximarlas lo más posible a la "verdadera" o exacta distribución posterior. Las estrategias para ello son dos: la de los algoritmos basados en el muestreo (*sampling-base algorithms*) y la de los algoritmos variacionales. Los primeros coleccionan muestras para compararlas a una distribución empírica. La más conocida de estas estrategias es el muestreo de Gibbs, en la que se construye una cadena de Markov (que es una secuencia de variables aleatorias que se van concatenando de modo que la siguiente depende de la anterior), en donde ella se define para un conjunto de tópicos ocultos y lo que hace el algoritmo es correr o ejecutar la cadena por mucho tiempo, coleccionar las muestras que va recabando en el camino, y luego aproximarse a la posterior con esa "previa", o sea, esas muestras. Se trata de una estrategia muy usada en paquetes de R para el modelado de tópicos basados en LDA (Blei, 2012, 82).

Los métodos variacionales constituyen la segunda alternativa de aproximación, de carácter "determinístico". Su estrategia es proponer un conjunto

de parámetros para la distribución conjunta y luego encontrar cuál de esos parámetros se acerca mejor a la evidencia empírica. De este modo, el problema de computación se convierte en un problema de optimización. Puede decirse que ambas estrategias tratan de encontrar una estructura tópica oculta en una colección fija de documentos que sirve de guía respecto de qué es lo que hay que buscar.

Cada familia de variaciones es una distribución de $\boldsymbol{z}$. Computacionalmente, esto se implementa iniciando una familia de distribuciones $\boldsymbol{\nu}\,(init)$, que se va optimizando hasta alcanzar un parámetro $\boldsymbol{\nu}^*$, aquel que se acerca más a la verdadera posterior $p(z|x)$ y que se usará como un "proxy". La divergencia o distancia entre las dos funciones de distribución de probabilidad p y q la evalúa Blei, en particular, con el método de la divergencia o ganancia de la información Kullback-Leiber, que ya se ha examinado en el capítulo anterior. De este modo, podemos definir la convergencia de la función como una ganancia de información.

En este caso, $\boldsymbol{\nu}^*$ sería un parámetro "libre", que se iría optimizando en tanto que *proxy* de la distribución posterior. Por ejemplo, la Figura 18 ofrece un ejemplo que presenta Blei de una mezcla de distribuciones gaussianas, en la que podemos ver los gráficos de las medias de los *clusters* o agrupaciones de datos que se van alcanzando a medida que se optimiza la función de densidad. Se trata de un problema intratable si se computa cada vez, en iteraciones sucesivas, las previas. Pudiera decirse que la imagen de abajo ofrece una representación del camino que toma el parámetro ν hasta alcanzar $\boldsymbol{\nu}^*$.

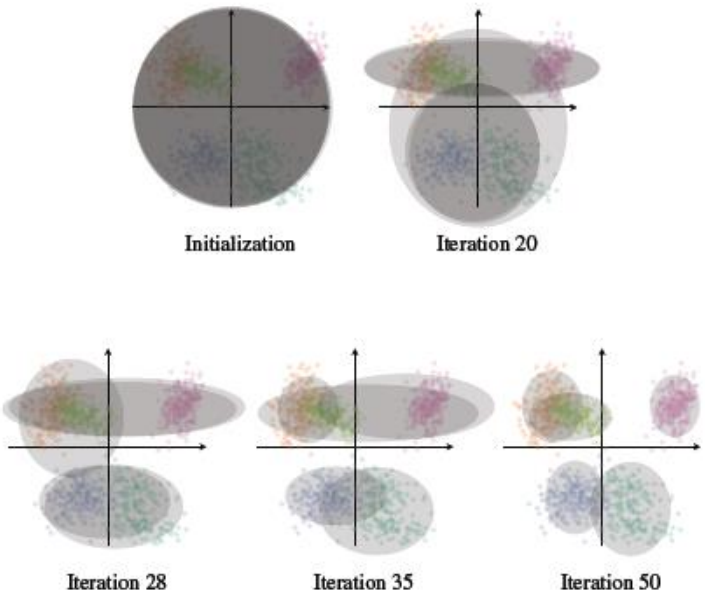


Figura 18: Optimización de la función de densidad de la probabilidad en distribuciones gaussianas (Blei et al. 2016, p.16)

La inferencia variacional adapta ideas provenientes de la estadística de la física a la inferencia probabilística. Se atribuye a Peterson y Anderson (1987) un primer uso de métodos de campo medio para ajustar una red neuronal. La idea luego fue retomada por Tommi Jaakkola, Lawrence Saul y Zoubin Charamani y generalizada a muchos modelos probabilísticos (ver Jordan et al, 1990). Por su parte, Hinton y Van Camp (1993) también usaron métodos variacionales de campo medio para el desarrollo de redes neuronales. Neal y Hinton (1993) conectaron estos métodos con EM. La inferencia variacional toca muchas áreas tales como programación probabilística, aprendizaje reforzado, redes neuronales, optimización convexa y estadística bayesiana.

De este modo, la inferencia variacional se propone, en principio, minimizar la divergencia KL entre $q(\beta, \mathbf{z}; \mathbf{v})$ (donde el punto y coma se lee "para algún parámetro ν) y la posterior real $p(\beta, \mathbf{z}|\mathbf{x})$. Esta divergencia se define como:

$$KL(q(\mathbf{z}; \mathbf{v}^*))\|p(\mathbf{z}|\mathbf{x})) \qquad (4.14)$$

Sin embargo, en los modelos de LDA el cálculo de la divergencia KL se implementa de modo que se incorpore al algortimo el hiperparámetro alpha para el cálculo de la inferencia Bayesiana de carácter predictivo. Se verá esta implementación en detalle en el capítulo siguiente.

4.3. Interpretación y evaluación de los métodos de modelado de tópicos.

Hemos visto a lo largo del presente trabajo tres métodos para la reducción de la dimensionalidad de conjuntos de datos de texto: el LSI, el PLSI y el LDA. Lo que comparten en común todos estos métodos es la referencia a un espacio semántico latente. Es decir, refieren las relaciones de los términos y los documentos a un tal espacio. Desde el punto de vista intuitivo, lo que se sugiere es que estas dimensiones latentes son equivalentes a conceptos o tópicos que dan sentido a los documentos.

La manera más evidente de utilizar los resultados del modelado de tópicos consiste en examinar las asociaciones entre los términos que conforman el tópico.

El LSA ordena los términos conforme al coeficiente que corresponde a un atributo dado en el espacio semántico. En los modelos probabilísticos (el PLSA y la LDA), los términos son correctamente atribuidos por la probabilidad de generar un término dado un tópico (Cfr. Crain et al., 2012, 148).

Los enfoques para evaluar estos modelos son variados: en primer lugar, está la capacidad de ajuste de los modelos al conjunto de datos de prueba, a fin de examinar su capacidad de generalización. Pero otras estrategias, incluyendo el análisis llevado a cabo por un ser humano, pueden ser adicionalmente necesarias o inevitables, de acuerdo con los autores. Examinamos ahora algunas de ellas:

4.3.1. Ajuste a los datos de prueba.

En el caso del LSI, se puede calcular el error ℓ_2 introducido por la aproximación de los documentos de prueba, o documentos que se han reservado aparte, sobre el espacio semántico. Los modelos probabilísticos pueden ser probados calculando la probabilidad de generar esos documentos de prueba con el modelo desarrollado. Esta probabilidad se puede calcular con una técnica llamada "Perplejidad":

$$\exp(-\frac{1}{N}\sum_{d=1}^{M}\sum_{n=1}^{N_d}\log p(w_{dn}|model)) \qquad (4:15)$$

Un número de "perplejidad" con un valor de 100, por ejemplo, indica que las probabilidades que arroja el modelo son equivalentes a escoger una palabra aleatoriamente de un léxico de 100 palabras. De este modo, valores cada vez menores indican que el modelo se adapta mejor al conjunto de datos de prueba. Existen, de acuerdo con Crain et al., diferentes maneras de computar esta probabilidad, pero la más recomendada es el método de izquierda a derecha, en donde la probabilidad de generar cada *token* en el documento se encuentra condicionada por todos los *tokens* anteriores en el documento, de modo que se hace evidente las interacciones de todos los *tokens* en el documento.

Otra manera de evaluar un modelo es examinar sencillamente su desempeño en una aplicación. Por ejemplo, se ha probado el desempeño del LDA con métricas estándar, o creadas *ad hoc*, para evaluar la recuperación de información en documentos.

Los modelos de modelado probabilístico tienen que servir realmente para ayudar a comprender los documentos cuyos tópicos reflejan. Pero, a veces, las medidas de evaluación están en contradicción con esta aspiración, en la medida en que modelos con una mejor medida de "perplejidad" son los más difíciles de interpretar: una cosa es un buen ajuste y otra un ajuste *significativo*. Muchas veces ambos están en contradicción en este tipo de tareas (Cfr. Crain et al., 2012, 150). Cualquier modelo puede ajustarse mejor a los datos que uno significativo en sentido semántico. Así pues, una manera de encontrar un modelo significativo es sometiéndolo a prueba por parte de usuarios que pueden identificar qué término no pertenece al pool de términos propios de un tópico: la evaluación por parte de un ser humano resulta, en este sentido, como se ha dicho, muchas veces inevitable.

Capítulo 5. Aplicaciones de las técnicas LSA, PLSA y LDA sobre tres conjuntos de datos de texto tomados de la red social Twitter.

Los siguientes capítulos, 5 y 6, están dedicados a las aplicaciones de los tres tipos de algoritmos. Describe de manera detallada y paso por paso la creación de tres aplicaciones sobre tres conjuntos de datos tomados de la red de microblogging Twitter, o de tres muestras de documentos de tweets, que hemos sometido al escrutinio de los tres tipos de técnicas: el LSA, el PLSA y la LDA. Se ofrecen tablas y figuras detalladas con los resultados que hemos obtenido, es decir, con los tópicos que han arrojado cada una de estas muestras. Al mismo tiempo, se ofrecen los pseudocódigos y descripciones de las herramientas que hemos usado para la aplicación de los algoritmos. Finalmente, también describimos las distintas bibliotecas de métodos que han sido utilizadas en la aplicación. Los parágrados (5.2), (5.3) y (5.4) están dedicados, respectivamente, al análisis de los resultados obtenidos para las muestras A, B y C. En el capítulo siguiente (6) se presenta una interpretación de los resultados obtenidos.

Para la aplicación de los algoritmos y métodos se ha utilizado el lenguaje de programación Python y su biblioteca de métodos de aprendizaje automático Scikit-learn (Pedregosa et al., 2011). Se han usado las versiones de Python 3.6.2 y de Scikit-learn 0.19. Algunos algoritmos de esta versión de Scikit-learn se ejecutaron en la plataforma de Python en la Nube: https://www.pythonanywhere.com/. Para la preparación de los documentos de texto, como veremos, se ha usado la biblioteca de métodos de Python *Natural Language Toolkit* (NLTK)(Bird et al., 2009). El editor de texto de Python que se ha usado es, en la mayoría de los casos, *Sublime*.

Para ejecutar el algoritmo de descomposición en valores singulares SVD Truncado, el análisis semántico latente, se ha usado el ambiente de distribución de bibliotecas de Python *Anaconda* y la versión de Python 2.7. En las páginas que siguen se describirán las distintas bibliotecas y métodos de Python que se han utilizado, así como los algoritmos usados en este trabajo. Igualmente se ofrecerán capturas de pantalla tomadas en los distintos momentos en los que se realizaron las distintas tareas.

5.1. Métodos y algoritmos usados en este trabajo.

De acuerdo con la documentación de Scikit-learn para la implementación, que hemos usado nosotros, de un algoritmo para la descomposición de matrices en valores singulares, el algoritmo SVD Truncado aplicado sobre conjuntos de datos de texto retorna las matrices como espacios semánticos de baja dimensionalidad.

Como ya se ha indicado en el capítulo 2, dedicado al Análisis Semántico Latente o LSA, el algoritmo, aplicado sobre una matriz X produce una aproximación de rango bajo tal que:

$$X \approx X_k = U_k \Sigma_k V_k^T \qquad (5{:}1)$$

Luego de esta operación, queda $U_k \Sigma_k^T$ como el conjunto de entrenamiento transformado con un número k de componentes, que se expresa en la implementación con el parámetro n_components, o tópicos, que en la aplicación hemos fijado, siguiendo distintos ejemplos que recomiendan ese número, en un máximo de 30 tópicos. Igualmente, se ha seguido la recomendación de los desarrolladores de Scikit-learn de aplicar el algoritmo sobre matrices transformadas conforme a un sopesado de términos-documentos *tf·idf*. En este sentido, esta recomendación concuerda con la literatura que se ha estudiado en el capítulo 1.

En relación con los algoritmos de índole probabilística, la aproximación bayesiana variacional o Bayes variacional (VB por sus siglas en inglés, *variational Bayes*) ofrece, tanto para el PLSA como para la LDA, una solución cuando la distribución de la probabilidad marginal es intratable computacionalmente: en vez de inferir las variables latentes a través de la marginalización directa de la distribución conjunta, se usa un *proxy* de una distribución mucho más simple y se realizan las inferencias a través de un proceso de optimización (Cfr. Crain et al. 2012, 146). Este "proxy" se conoce, como ya se ha señalado en los capítulos 3 y 4, como la divergencia de Kullback-Leibler.

Ésta:

→Convierte la inferencia de la máxima probabilidad condicional o *MLE* en un problema de optimización.

→Propone una familia variacional de distribuciones para las posibles variables latentes.

→Ajusta los parámetros variacionales para que se acerquen a la verdadera posterior, a través de la divergencia de Kullback-Leibler.

En el presente trabajo, siguiendo la implementación del PLSA en la bilbioteca de métodos de Python Scikit-Learn (Pedregosa et al., 2011), el análisis semántico latente de índole probabilístico se concibe como una factorización de una matriz no negativa X (la matriz de tweets/documentos-términos que se usa) con la función objetivo definida por la divergencia Kullback-Leibler generalizada.[10]

La función implementada[11] en el algoritmo de Scikit-Learn que hemos utilizado es:

[10]Véase la documentación sobre la implementación del método PLSA o factorización de matrices no-negativas con la divergencia de Kullback-Leibler de Scikit-Learn enhttp://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py.
[11] Véase http://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation.

$$d_{KL}(X,Y) = \sum_{i,j}(X_{i,j}\log\left(\frac{X_{i,j}}{Y_{i,j}}\right) - X_{i,j} + Y_{i,j}) \qquad (5:2)$$

La divergencia KL utiliza uno de los parámetros de forma Beta, que define una divergencia entre dos funciones de probabilidad Beta, una aproximada y la real. En la divergencia KL, β=1 (Févotte e Idier, 2010).[12]

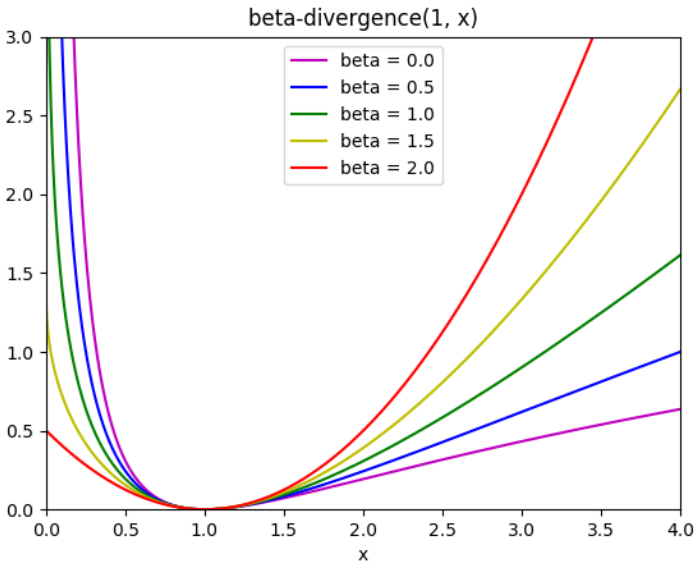


Figura 19: Representación gráfica de las divergencias Beta, que en KL optimiza la función objetivo alrededor de Beta=1 (documentación de Sklearn)

El algoritmo de Scikit-Learn implementa un solucionador "mu", o Multiplicative Update, que optimiza rápidamente cualquier divergencia Beta.

Por su parte, en relación con el tercer algoritmo, la atribución latente de Dirichlet, también se ha usado la implementación de Scikit-Learn para la LDA.[13] En donde, recordemos, en contraste con el PLSA, la optimización de la función de aproximación a la distribución posterior se realiza tomando en cuenta además los hiperparámetros α y η, como se observó en el capítulo 4. De este modo, la posterior en funciones de la LDA puede formularse como:

$$p(z,\theta,\beta\,|w,\alpha,\eta) = \frac{p(z,\theta,\beta|\alpha,\eta)}{p(w|\alpha,\eta)} \qquad (5:3)$$

Debido al carácter mayormente intratable de la probabilidad marginal, se utiliza aquí un método de inferencia variacional en donde se aproxima la

[12]Févotte, Cedric y Jérome Idier, "Algorithms for nonnegative matrix factorizacion with the beta divergence"en arXiv.org, cs, 1010.1763, Cornell University, 2010, disponible en https://arxiv.org/abs/1010.1763.
[13]Véase también http://scikitlearn.org/stable/modules/decomposition.html#latentdirichletallocation.

distribución real con parámetros variacionales obtenidos de modo empírico $p(z, \theta, \beta | \lambda, \phi, \gamma)$, los cuales se optimizan maximizando el ELBO o *evidence lower bound*, la cota inferior de la evidencia.

Ésta se define como:

$$\mathcal{L}(\boldsymbol{v}) = \mathbb{E}_q[\log p\,(\beta, \mathbf{z}, \mathbf{x})] - \mathbb{E}_q[\log q\,(\beta, \mathbf{z}, \boldsymbol{v})] \qquad (5{:}4)$$

Es decir, el valor esperado del logaritmo de la probabilidad conjunta de todos los tópicos y las variables ocultas y observadas, comparado con la entropía de la distribución variacional, el valor esperado de q.

Dado que en modelos de LDA la divergencia KL siempre se considera intratable, maximizar la ELBO es equivalente a minimizar de KL. La ELBO arroja sólo máximos locales en espacios de puntos no convexos y representa también una cota inferior para el logaritmo de la evidencia o marginal desconocida: $\log p(\mathbf{x})$.

A diferencia de KL, que es básicamente una medida de distancia, en la ELBO el primer término de q $\{\mathbb{E}_q[\log p\,(\beta, \mathbf{z}, \mathbf{x})]\}$ depende del cálculo de la máxima probabilidad posterior estimada (MAP: *maximum a posteriori estimation*), que luego se compara con el cálculo del MAP de la distribución variacional.

Al mismo tiempo, ELBO también prefiere que el segundo término $\{\mathbb{E}_q[\log q\,(\beta, \mathbf{z}, \boldsymbol{v})]\}$ sea *difuso*.

Al respecto, Blei escribe:

"¿Sobre cuales valores de $\mathbf{z}$ esta función objetivo $q(\boldsymbol{z})$procurará colocar su masa? El primer término es una distribución condicional esperada; promueve densidades que colocan su masa en configuraciones de las variables latentes que explican los datos observados. El segundo término es la divergencia negativa entre la densidad variacional y la previa, que *promueve densidades cercanas a la previa. Por lo tanto, el objetivo variacional refleja el balance usual entre la distribución condicional y la previa*" (Blei et al., 2016, 7).[14]

La implementación de LDA en Scikit-learn, que hemos aplicado a los conjuntos de datos, utiliza el algoritmo de Bayes variacional "Online", el cual converge de manera garantizada a un óptimo global, actualizando pequeños conjuntos de datos seleccionados ("*mini batch data points*") de modo aleatorio o estocástico (Hoffman, Blei y Bach, 2010, 4-5).De acuerdo con Hoffman et al., este es el mejor enfoque y el más eficiente cuando los datos están llegando continuamente al algoritmo, pero no sólo entonces.

[14] El subrayado es nuestro.

Algoritmo "Online". Bayes variacional para LDA:

Definir $\rho_t \triangleq (\tau_0 + t)^{-k}$

Inicializar λ aleatoriamente

Para $t = 0$ hasta ∞ **hacer**

Paso *E*:

Inicializar $\gamma_{tk} = 1$ (la constante es arbitraria)

Repetir

Establecer $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log \theta_{tk} + \log \beta_{kw}]\}$

Establecer $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$

Hasta que $\frac{1}{K}\sum_k |cambio\ en\ \gamma_{tk}|$ <0,00001

Paso M:

Computar $\tilde{\lambda}_{kw} = \eta + Dn_{tw}\phi_{twk}$

Establecer $\lambda = (1 - \rho_t)\lambda + \rho_t\,\tilde{\lambda}$

Fin

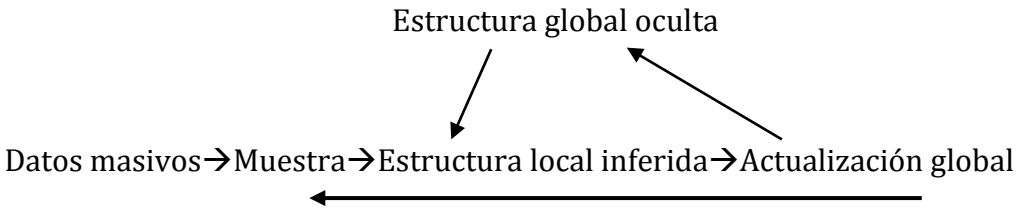
En la descripción de su uso en la implementación que hemos usado, Scikit-learn, el algoritmo está caracterizado como *learning_method* o método de aprendizaje por defecto para actualizar los componentes o tópicos. Especialmente útil si el conjunto de datos o atributos es grande. En relación con la implementación del algoritmo, los autores de la implementación en Scikit-learn señalan que, en cada actualización del algoritmo EM, se usan muestras pequeñas (*mini-batch*) de los datos de entrenamiento para actualizar incrementalmente las variables de los componentes.[15]

Aquí, un parámetro de declive del aprendizaje implementado en Scikit-learn controla la tasa aprendizaje en el método "Online". Se usó un parámetro por defecto 0,7, que garantiza convergencia asintótica. En la literatura este parámetro se denota κ. Por su parte, el parámetro de compensación de los pesos en iteraciones tempranas del algoritmo se denota τ_0, y, por defecto, es igual a 10 en la implementación que hemos usado. El tamaño de las pequeñas muestras o "*mini-batches*" es, por defecto, 128. Las iteraciones máximas del algoritmo son, por defecto, 10. Nosotros hemos usado 5, dado que el conjunto de datos no es muy grande y siguiendo la recomendación dada por el ejemplo disponible en la página de Scikit-learn.

[15]Véase página de Scikit-learn:
http://scikitlearn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocatio.html#sklearn.decomposition.LatentDirichletAllocation
para detalles sobre la implementación "online" de este algoritmo, que converge rápidamente a un óptimo.

Así pues, en el algoritmo "Online" se escala la inferencia variacional para conjuntos masivos a través de la inferencia variacional estocástica, en donde el algoritmo no pasa cada vez por el conjunto total de datos masivos sino simplemente por pequeñas muestras"ruidosas" de éste. Lo esencial del enfoque de este algoritmo no es que los datos lleguen a los algoritmos en "streaming", como pudiera sugerir su nombre "online", sino su uso de la inferencia variacional estocástica.

La SVI (inferencia variacional estocástica, *stochastic variational inference*) tiene el siguiente flujo de computación:



La inferencia estocástica reemplaza el descenso por el gradiente, un método muy usando en redes neuronales bayesianas, por estimados más "ruidosos y más variados", más "baratos" (*cheaper*) del conjunto de datos. Los estimados se hacen sobre muestras locales que, eventualmente, pudieran dar una idea de la distribución de probabilidad global de los tópicos. Desarrollado por Robbins-Monro, este algoritmo de aproximación estocástica ha posibilitado, según Blei, el aprendizaje automático.

En efecto, al respecto podemos leer en Hoffman, Blei y Bach (2010):

> "Algoritmos de optimización estocástica optimizan una función objetivo usando estimados ruidosos de su gradiente. Aunque no hay [en él] una computación explícita del gradiente, el algoritmo [Online] puede ser interpretado como un algoritmo para computar un gradiente natural estocástico" (5).

Una ventaja muy importante de la SVI es, como decíamos, la de que la convergencia está garantizada para óptimos locales. De hecho, ha posibilitado el moderno aprendizaje automático. Es muy importante en el desarrollo de esta moderna disciplina.

En resumen:

→Se toma un documento de muestra

→Se estiman los parámetros variacionales locales usando los tópicos actuales.

→Se especifican tópicos intermedios o no definitivos desde esos parámetros locales.

→Se actualizan los tópicos a través de un promedio sopesado entre tópicos actuales e intermedios.

Este procedimiento hace posible empezar con una serie de tópicos "basura" y mejorar la determinación de tópicos con el algoritmo estocástico a medida que se ejecuta. Se empieza con un documento o con una pequeña muestra del corpus de documentos y a medida que se avanza el modelo es cada vez más preciso en la determinación de los tópicos relevantes. La precisión o la convergencia se caracteriza por una suerte de "estabilización" de los tópicos o las palabras más importantes en el corpus de documentos. Su rendimiento es, de hecho, espectacular para conjuntos masivos de datos. Sus medidas de "perplejidad", en donde un número más bajo es mejor, aventajan a los algoritmos clásicos de inferencia variacional.

Este es el tipo de algoritmo preferido por Blei para conjuntos enormes de documentos, en el orden de los millones. La inferencia variacional estocástica puede aplicarse a distintos modelos, tales como modelos de mezcla bayesianos, modelos de series de tiempo, modelos factoriales y de factorización de matrices, regresión de multinivel, modelos de mezcla mixta LDA, etc. La idea, como siempre, es tratar de aproximarse lo mejor posible a la probabilidad posterior verdadera.

Finalmente, para concluir este parágrafo, es menester señalar, en relación con el problema de cómo evaluar estos modelos, que las distintas implementaciones de Scikit-learn no ofrecen una manera de evaluar la "perplejidad" en el desempeño de este y otros algoritmos de índole probabilística. La razón se encuentra en la documentación de Scikit-learn que especifica las funciones y métodos que se han usado en la implementación de un algoritmo de atribución latente de Dirichlet.[16] Allí, puede leerse que, en versiones anteriores, la perplejidad de la muestra era definida como:

$$\exp(-1 * log - likelihood\ per\ word) \qquad (5:5)$$

No obstante, la posibilidad de este cálculo ha sido descartada por los programadores de la versión más reciente, que hemos usado nosotros, 0.19 en el momento de redacción de este trabajo, dado que "el argumento *doc-topic-distr* (distribución de tópicos por documento) se ignora porque el usuario ya no tendría acceso a una distribución no normalizada" (véase Figura 20). Esto es así porque los algoritmos que implementan métodos probabilísticos van actualizando cada vez el cálculo de la probabilidad condicional hasta encontrar un óptimo global que ya no se puede comparar con la matriz desestructurada original.

[16] Véase http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html.

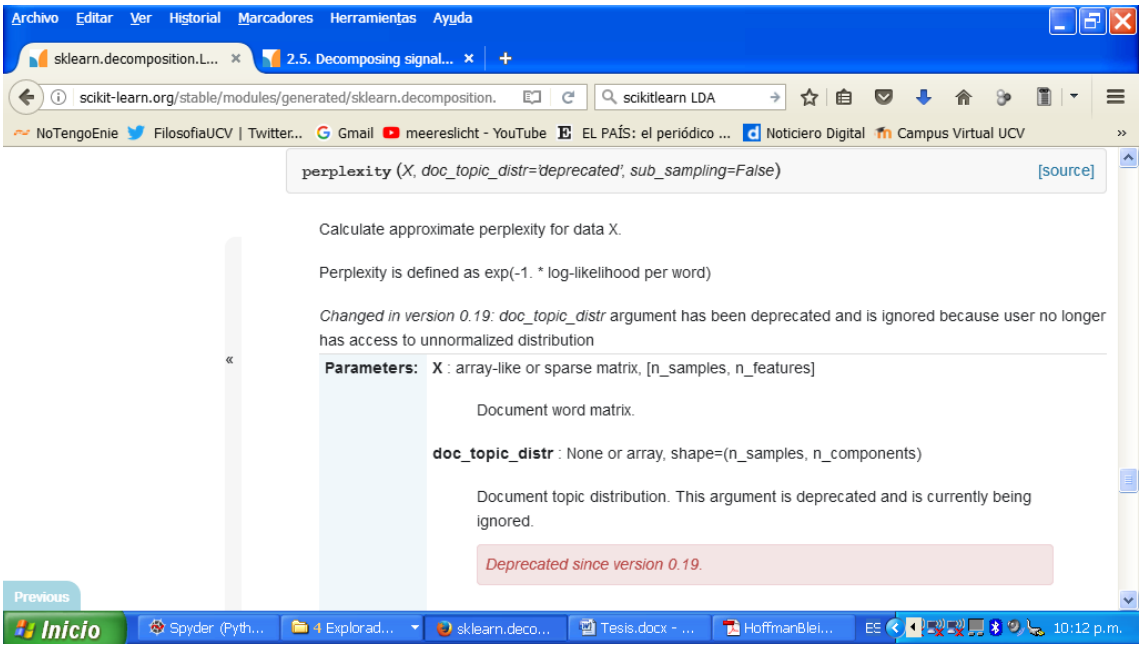


Figura 20: Descarte del método doc_topic_distr en la versión 0.19 de Scikit-learn

Por esta razón, en nuestras aplicaciones nos hemos acogido a la posibilidad de evaluar los resultados arrojados por los tres tipos de algoritmos tomando en cuenta no una medida de perplejidad, sino el sentido, evaluado por un experto, de la semántica en la determinación de los distintos tópicos, una posibilidad de evaluación que en la literatura relevante se presenta, muchas veces, como inevitable para este tipo de modelos (Cfr. Crain et al., 2012), como se expuso al final del capítulo anterior.

5.2. Aplicación sobre un primer conjunto de datos de texto (Muestra A)

5.2.1. Análisis Semántico Latente (LSA) de la muestra A.

Las aplicaciones de los algoritmos se iniciaron tomando tres muestras de tweets que giraban alrededor de la palabra clave "*derechos*", palabra de fuerte carga conceptual relacionada con la ética, la filosofía moral y filosofía de derecho, áreas de desempeño profesional de quien esto escribe. Esta palabra se usó como "operador" para explorar sus connotaciones en el lenguaje natural u ordinario de los usuarios de Twitter. En las páginas que siguen se describirá cómo se recuperaron estos tweets y cómo fueron tratados en el estudio.

En primer lugar, para la construcción del modelo de LSA, en las tres muestras se aplicó un algoritmo de descomposición de una matriz de bolsas de palabras en valores singulares (SVD) sobre un conjunto de datos de texto extraído de la red social Twitter.

El algoritmo que se utilizó fue el "TruncatedSVD" o "Descomposición en valores singulares truncado", un algoritmo que se conoce también como "Análisis semántico latente", como se recordará (Pedregosa et al., 2011 y Aggarwal, Charu, 2012, p. 43).

El SVD Truncado se encuentra implementado para Python en la colección de bibliotecas y métodos para Python, ya citada, Scikit-learn.

Veamos, de forma detallada, el proceso de análisis semántico latente del conjunto de datos y sus resultados:

1. En primer lugar, como se observa en la Figura 21, se creó en la página del API de Twitter, una aplicación:

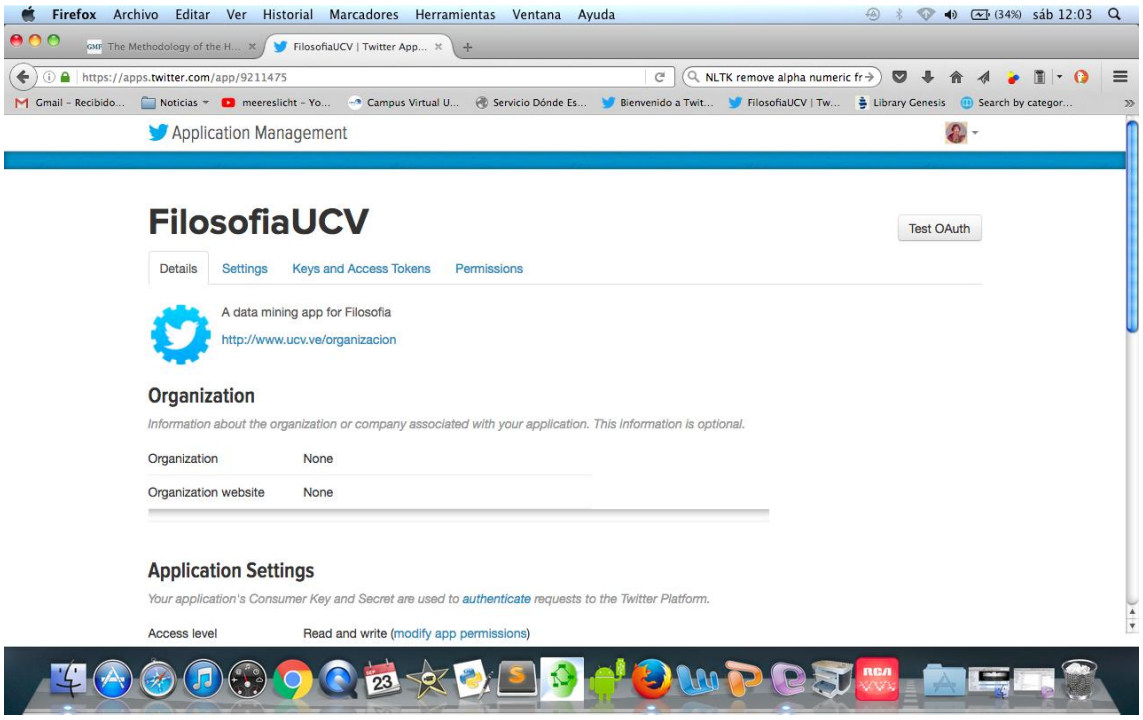


Figura 21: Aplicación creada en Twitter

Esta aplicación crea las credenciales y claves que nos permiten acceder a la base de datos de Twitter. Entre las distintas opciones que la plataforma de Twitter ofrece a los desarrolladores para la exploración de su base de datos, destacan dos: los REST APIs y los Streaming APIs. La diferencia fundamental entre ambas es que mientras que la primera recupera tweets que ya forman parte de la base de datos, la segunda transmite a la aplicación los tweets que van entrando al así llamado "firehose" de Twitter en tiempo real, lo que posibilita el desarrollo de aplicaciones de investigación muy interesantes, tales como, por ejemplo, la evolución y cambio de los sentimientos positivos o negativos alrededor de una marca o un personaje de manera instantánea.

Aunque los APIs de Streaming de Twitter ofrecen solo el 1% del firehose de manera gratuita, los tweets así recuperados abren la posibilidad de preguntarse qué están diciendo en ese momento los usuarios de Twitter sobre un determinado tema. Se ha optado, pues, por esta segunda opción.

2. De este modo el día 24/9/2017 se capturaron, en una conexión en tiempo real que se extendió durante 30 minutos, 1148 tweets usando como palabra clave o "trackeador" el concepto "derechos", es decir, tweets que mencionaban el término "derechos" en las emisiones de los usuarios. Como puede observarse en la Figura 22, las instrucciones del script de Python utilizaron la biblioteca y los métodos de Tweepy, los cuales comportan una función que permite rastrear (*track*) un determinado concepto o palabra en un tweet y recuperar sólo aquellos que lo mencionan, de entre los que llegan al firehose de Twitter en ese momento.

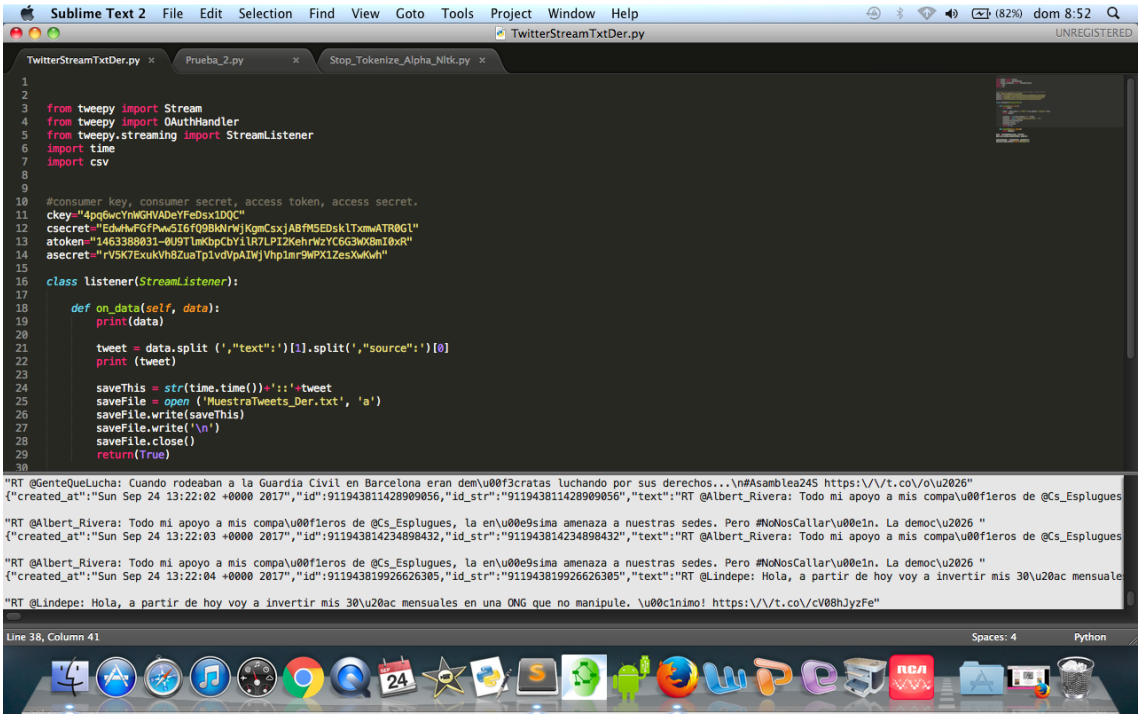


Figura 22: Recuperación de tweets con la biblioteca de métodos Tweepy de Python.

Así, usando la biblioteca de métodos de Python Tweepy, se logró recuperar de la base de datos de Twitter, y salvar en un documento con formato txt, el ya mencionado número de tweets.

3. Como se puede observar en la Figura 23, dada la naturaleza particular de la herramienta Twitter, los tweets son un conjunto muy "sucio" de datos de texto. Este es uno de los mayores desafíos que se plantean al investigador que quiere hacer análisis semántico de los conjuntos de datos de texto que ofrece Twitter. El usuario de Twitter, que está restringido a muy pocos caracteres, utiliza la herramienta para dar opiniones breves y golpes de efecto comunicativo, en textos apresurados, formularios o abreviados, en donde no se suelen respetar las reglas habituales de la redacción de textos. El usuario de Twitter no se preocupa por la

pervivencia de sus ideas, ni se imagina que nadie va a reprocharle la mala ortografía, el recurso a un "emoticón" o las palabras incompletas.

Por otro lado, la empresa Twitter tampoco está muy preocupada en ofrecer al desarrollador un filtrado de caracteres especiales o extraños. Esto dificulta enormemente la tarea del investigador, sobre todo si no es un anglohablante.

De este modo, los tweets recuperados inicialmente por la aplicación se ven así (Figura 23):

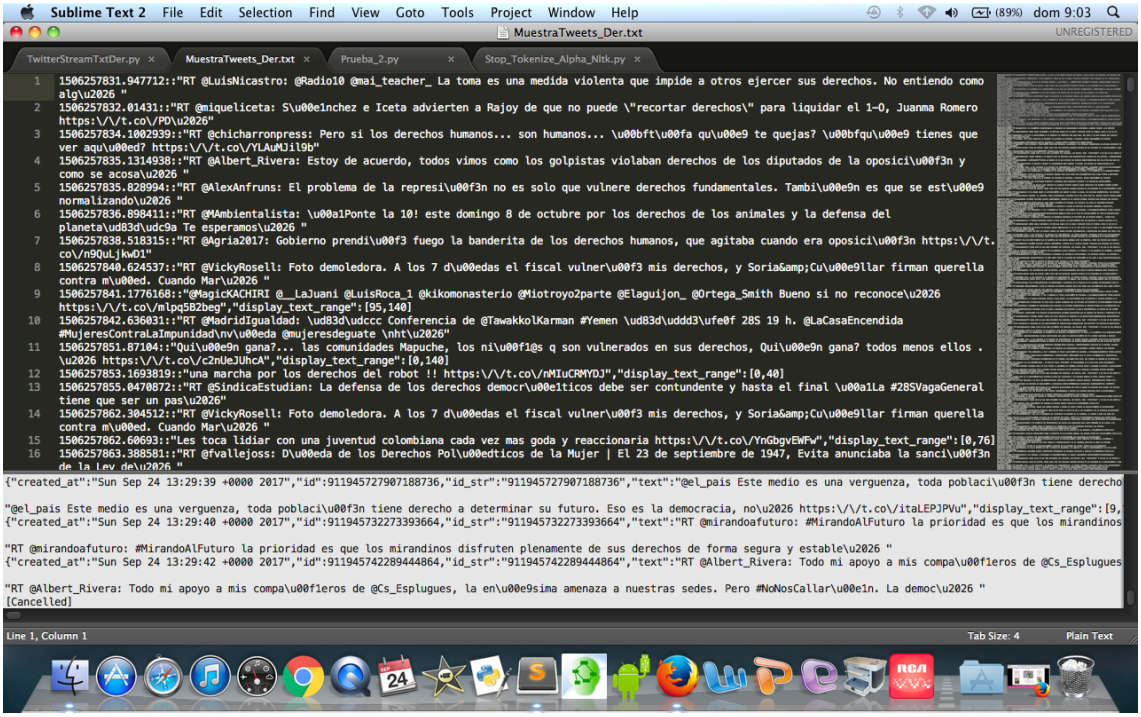


Figura 23: Muestra de tweets recuperados.

3. El siguiente paso es procesar los tweets para hacerlos aptos para el análisis semántico. Para ello (véase Figura 24), se utiliza un script que pasa el documento inicial por tres filtros sucesivos: uno para las "stopwords" o palabras comunes de poco peso semántico (tales como artículos determinados e indeterminados, preposiciones, etc.), otro filtro para puntuaciones y caracteres extraños (tales como la arroba o los *slash*) y otro para los caracteres numéricos. En algunos de los documentos sucesivos que van limpiando el documento inicial hemos usado el codificador de Python "latin1", que procesa un poco mejor que el habitual "utf-8", las oraciones en español. Se ha utilizado el catálogo de stopwords que ofrece NLTK, así como se han agregado algunos signos y términos en un catálogo personalizado.

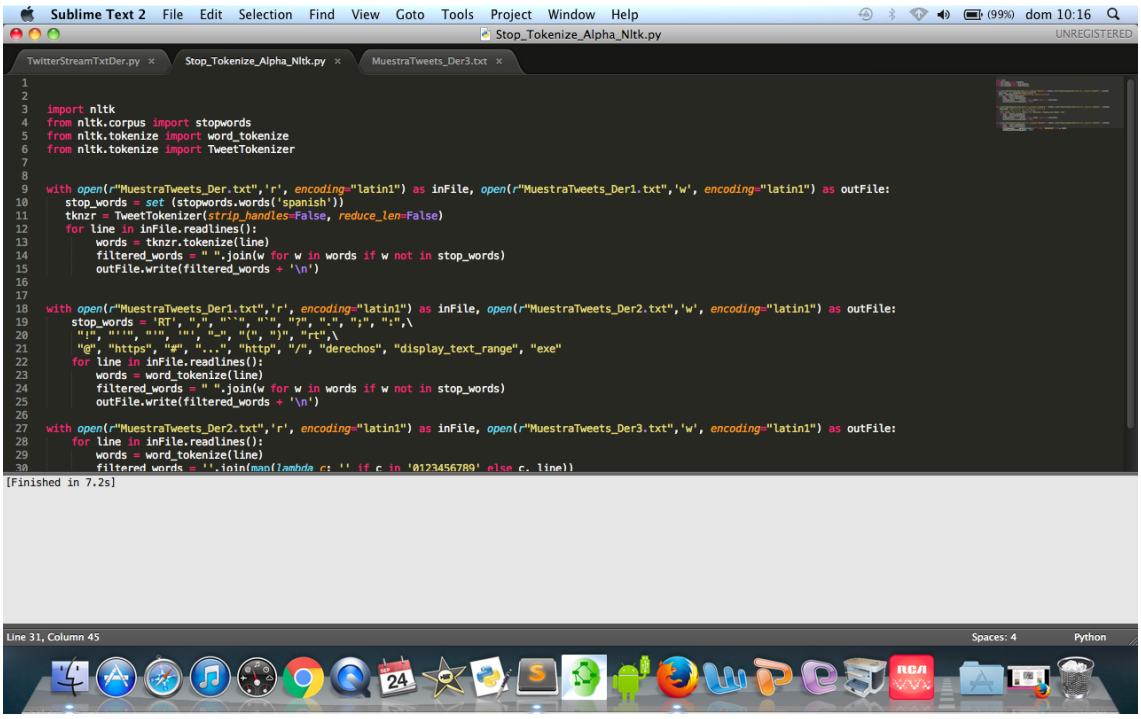


Figura 24: Preprocesamiento de tweets con la biblioteca de métodos de Python NLTK.

Como resultado de esos distintos procesos de filtrado, el conjunto inicial de tweets adquiere un formato mucho más idóneo para su análisis semántico latente (Figura 25). En efecto:

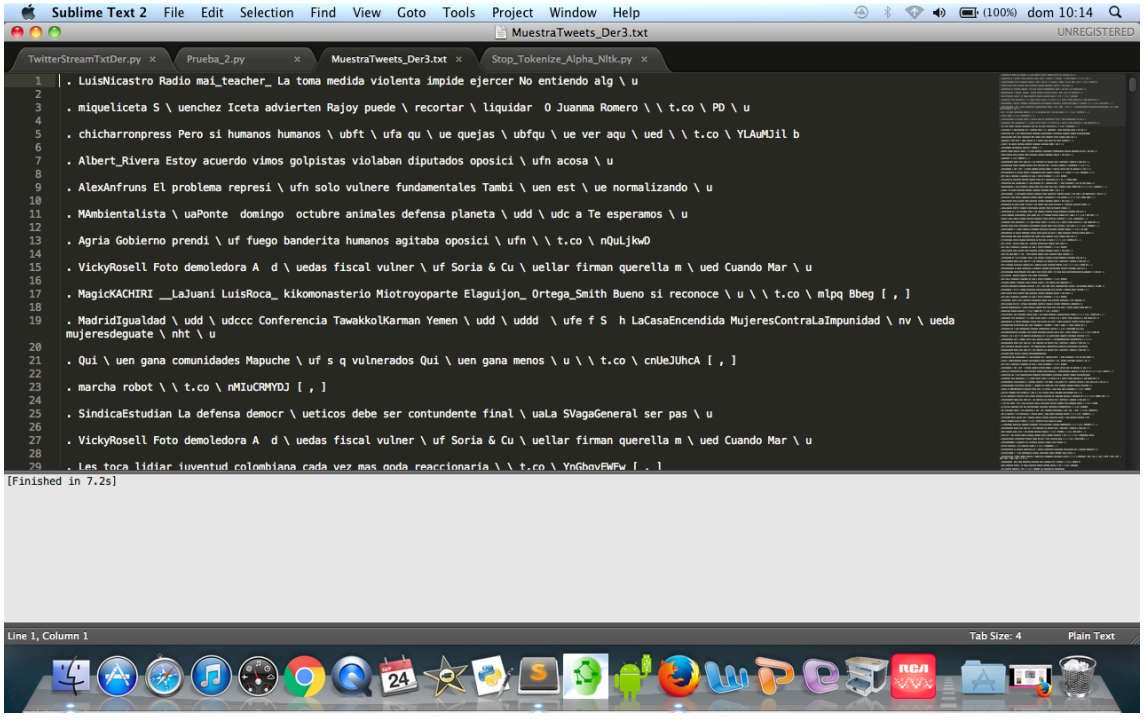


Figura 25: Muestra de tweets ya procesados.

La biblioteca de métodos de Python que se ha utilizado para esta tarea de procesamiento fue NLTK, *Natural Language Tool Kit*. Como ya se ha señalado, se usó su corpus de "stopwords" en español y tokenizadores para palabras y tweets.

4. Una vez en posesión de un documento limpio pasamos al procesamiento del conjunto de datos de texto para su análisis semántico latente. En pseudocódigo el script pudiera tener la siguiente estructura:

Comienzo

Importar bibliotecas de métodos de Scikit-Learn (Python). Feature_extraction.text

Leer archivo de texto con los tweets

Crear matriz de bolsa de palabras X (método sklearn_CountVectorizer)

Sopesar matriz *tfi·df* (método sklearn_TfidfVectorizer)

Procesar matrix con el algoritmo LSA (sklearn_TruncatedSVD)

Escribir los componentes de la SVD como Conceptos o Tópicos principales de la matriz X.

Fin

El primer paso es, pues, etiquetar cada tweet del conjunto de datos de texto anterior a fin de que Scikit-learn pueda identificar cada tweet como un documento independiente para una matriz de bolsas de palabras. Esta fue una de las mayores dificultades que se enfrentaron al hacer las aplicaciones, dado que Scikit-learn no identifica un tweet como un documento, dada la brevedad del texto. Hubiera sido posible ejecutar un script que subía a la memoria del programa los tweets como documentos separados por comas, pero se optó por etiquetar cada tweet de un archivo txt que pudo, entonces, ser leído por el programa como un compendio de documentos individuales.

5. El paso siguiente es, como ya adelantábamos en el análisis teórico de los métodos de modelado de tópicos basados en la factorización de matrices, la creación de una matriz de bolsa de palabras. Esta fue realizada con el método de la biblioteca Scikit-learn "feature_extraction.text: CountVectorizer", la cual crea vectores de palabras o atributos y cuenta su frecuencia en matrices dispersas. Con la función X.shape fue posible también estimar las dimensiones de la matriz creada. Para el corpus en cuestión: 1148 documentos y 13565 palabras o atributos.

6. El siguiente paso es, como ya se ha indicado también, el sopesado *tf·idf* a través del método "*Tfidfvectorizer*", el cual ofrece aquellas palabras o atributos más representativos de cada documento.

7. Ya en posesión de una matriz de bolsa de palabras sopesada con el algoritmo *tf·idf*, el siguiente paso es el análisis semántico latente o la descomposición de la matriz sopesada en valores singulares (Figura 26):

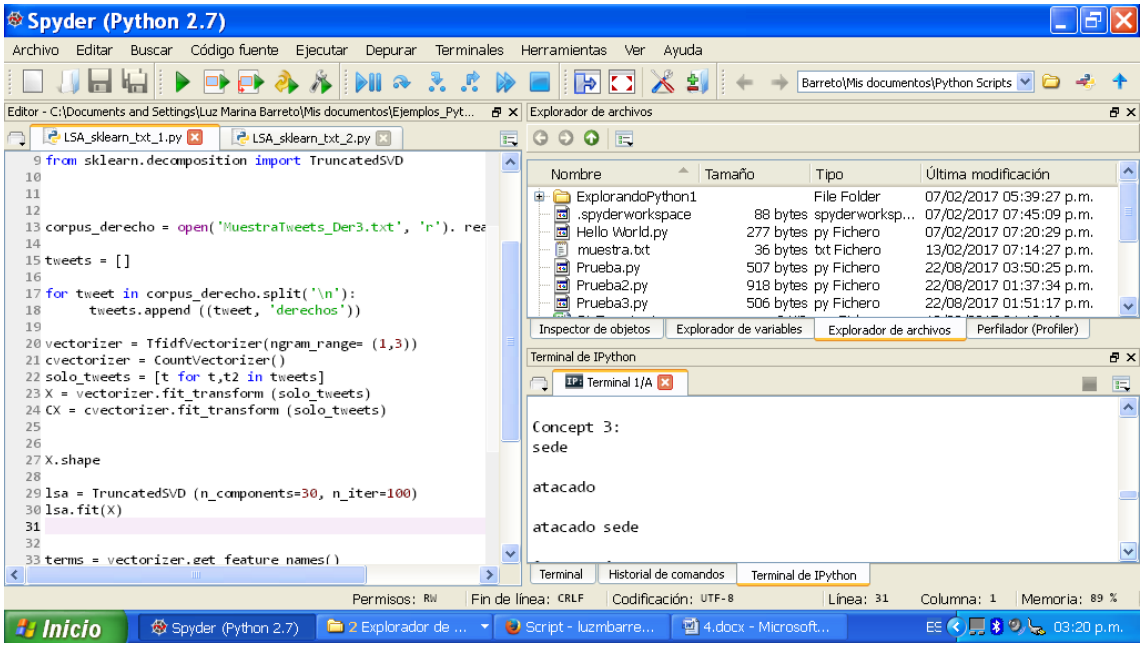


Figura 26: Aplicación del algoritmo SVD Truncado para el LSA, ejecutado en el ambiente de bibliotecas para Python *Anaconda*.

8. Resultados. Los conceptos resultantes de la aplicación del algoritmo LSA se transcriben en la Tabla 15. Se decidió, siguiendo ejemplos en la literatura, que para la matriz X un número de 30 conceptos pudiera ser adecuado, y en 100 iteraciones, un número adecuado dado el relativamente pequeño número de documentos. Para hacer más amigable la lectura de los resultados obtenidos en los tres conjuntos de datos, para los tres métodos hemos destacado las palabras más prometedoras, desde el punto de vista de una reconstrucción de la semántica, en negritas.

Concepto 0	Concepto 1	Concepto 2	Concepto 3	Concepto 4	Concepto 5
amenaza **amenaza**, **sedes**, **amenaza sedes** pero	**acosa**, **acuerdo** vimos, **acuerdo** vimos **golpistas**	est claro claro lado	**Sede** **atacado** **atacado sede**	ufn la **humanos**	ufn noche est uen
Concepto 6	**Concepto 7**	**Concepto 8**	**Concepto 9**	**Concepto10**	**Concepto11**
Ufn est uen uen	co co xxqse co xxqse rh	co lado uf	Co **diputados** la	est uen est **diputados**	lado **libertades** ufn
Concepto12	**Concepto13**	**Concepto14**	**Concepto15**	**Concepto16**	**Concepto17**
defensa no queda	der **libertades** **defensa libertades**	co xkypiqth **democ** co democ co xkypiqth	est uen **queda** libertades	**acuerdo** **defensa** der	**acuerdo** la est
Concepto18	**Concepto19**	**Concepto20**	**Concepto21**	**Concepto22**	**Concepto23**
Esos **importa** **diputados**	la **noche** **sede**	***antiespecista*** no *antiespecista* *antiespecista* **no veganismo**	la **golpistas** esos	__lajuani __lajuani luisroca_ **kikomonasterio** __lajuani luisroca_	uen la todo
Concepto24	**Concepto25**	**Concepto26**	**Concepto27**	**Concepto28**	**Concepto29**
_avila adacolau _avila todo	**diputados** uen _avila	todo diputados todo apoyo	**albert_rivera** co xkypiqth democ co	todo compa uferos cs_esplugues uferos cs_esplugues	pero albert_rivera todo

Tabla 15: Salida del algoritmo de LSA en la muestra A

5.2.2. Analisis semántico latente probabilístico o PLSA de la muestra A:

1. En segundo lugar (Figura 27), se observan los tópicos obtenidos con una técnica de indexado semántico latente de índole probabilístico o PLSA, que calcula la divergencia Kullback-Leibler:

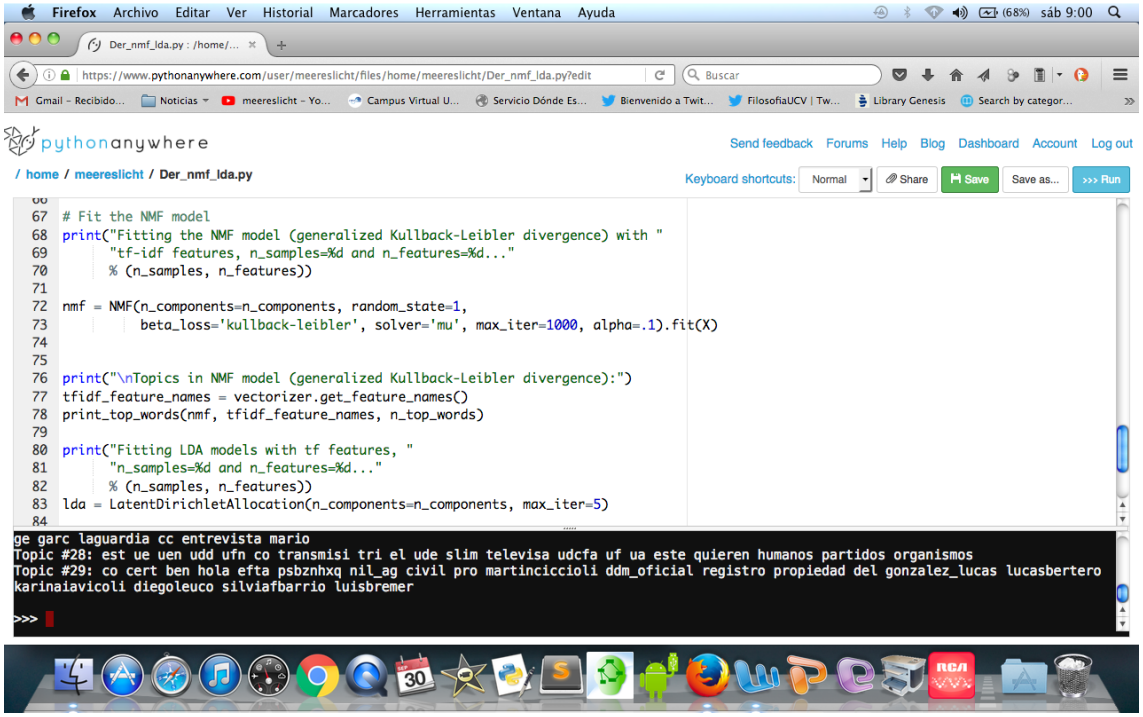


Figura 27: Aplicación de las implementaciones de los algoritmos PLSA y LDA en Scikit-learn.

2. Su algoritmo tendría la siguiente forma:

Comienzo

Importar bibliotecas de métodos de Scikit-Learn (Python). Feature_extraction.text y sklearn.decomposition

Leer archivo de texto con los tweets

Crear matriz de bolsa de palabras X (método sklearn_CountVectorizer)

Sopesar matriz tfidf (método sklearn_TfidfVectorizer)

Procesar matrix con el algoritmo NMF (divergencia de Kullback-Leibler generalizada o PLSA)

Escribir la salida como Conceptos o Tópicos principales de la matriz X.

Fin

3. Salida del algoritmo y tópicos recuperados:

"Fitting the NMF model (generalized Kullback-Leibler divergence) with tf-idf features, n_samples=1144 and n_features=13000...

"Topics in NMF model (generalized Kullback-Leibler divergence, PLSA):

Topic #0:

pero todo uferos **compa** uferos compa uferos cs_esplugues todo **apoyo** apoyo cs_esplugues compa uferos cs_esplugues uferos cs_esplugues uesima **democ sedes** pero **nonoscallar** sedes pero sedes cs_esplugues uesima uen la pero nonoscallar uen uesima **amenaza** sedes uesima amenaza

Topic #1:

ued uf uedas **fiscal soria** vickyrosell cuando cuando mar vulner uf soria soria cu uellar soria cu ued cuando mar uf soria cu uf soria **foto demoledora** uedas vulner uf vulner foto demoledora uedas fiscal vulner uedas fiscal

Topic #2:

golpistas ufn **albert_rivera diputados oposici acuerdo** oposici ufn acuerdo vimos golpistas oposici ufn acosa diputados oposici ufn ufn acosa albert_rivera estoy **violaban** golpistas violaban golpistas violaban diputados **estoy acuerdo** vimos estoy acuerdo estoy violaban diputados oposici violaban diputados

Topic #3:

ufa espa **reclamar expresarse romper** espa espa ufa reclamar **derecho decidir** decidir espa ufa derecho **reventar permitir violencias** espa ufa reventar ufa reventar permitir ufa reventar ufa reclamar derecho ufa reclamar violencias reventar permitir reventar romper expresarse romper

Topic #4:

uen est uen est lado esos **importa queda** los importa **roben soviets** esos **patriotas** soviets esos soviets uen **radicales** los est uen radicales lado est uen roben der roben **queda claro lado paulavazqueztv** queda claro paulavazqueztv queda

Topic #5:

ufos ddhh matibagnato **vida** ue der **mataron cagaron** vida vida cumplea ufos org ddhh ufos org **hijo mariano** ddhh **hablan** cumplea ufos org **cumplea** ufos cumplea org cagaron vida cumplea org ddhh org ddhh hablan

Topic #6:

no **libertades defensa** sede defensa libertades **esta noche** cs_esplugues uedmetro defensa uedmetro sede cs_esplugues no sede cs_esplugues ciutadanscs esta **ciutadanscs** cs_esplugues **no cederemos** ciutadanscs esta noche no cederemos cent esta noche esta noche **atacado** uedmetro defensa libertades

Topic #7:

ueda ufn ufos **mujer derechos ley hoy voto** pol ueda derechos pol ueda derechos uedticos mujer derechos pol uedticos pol uedticos mujer derechos pol uedticos pol uedticos ufn ley ley voto ufos promulgaci ufn

Topic #8:

libertad ufn **amnistiaespana reuni deben garantizar libertad expresi** ufn expresi **manifestaci** ufn manifestaci expresi ufn ufn reuni ufn libertad expresi reuni ufn manifestaci reuni ufn ufn reuni ufn manifestaci ufn ufn manifestaci expresi ufn reuni deben garantizar libertad

Topic #9:

derecho votar renuncia porlagoma votar derecho renuncia **venezuelanoseabstiene porlagoma** votar votar derecho venezuelanoseabstiene votar derecho renuncia derecho renuncia venezuelanoseabstiene porlagoma derecho **renuncia** uedamos **deber** co delapuentejuan deber teoverasrd und teoverasrd und uedacomohoy ufola **cambio** espa ufola

Topic #10:

co uf si nhttps co **pueblo parte** sin nhttps **derechos** co **pais pablo_iglesias_ independencia hablando** en **sitios** par **favor** irene_montero_ uflo **nonoscallaran**

Topic #11: **defensor pol humanos** uedtico defensor humanos **niexcusas** pol uedtico uedtico **joven** uedtico joven **cristiano** humanos **elsalvador** santa pol uedtico joven santa ana niexcusas **santa ana** santa humanos

elsalvador joven cristiano niexcusas co elsalvador santa joseluissafie joseluissafie pol

Topic #12: **gente venezuela** eso sep **cambiar reclamar** uedso **viva pepa gente** venezuela uedso **abusadores** venezuela uedso eso cambiar sep venezuela sep venezuela uedso viva reclamar eso cambiar reclamar eso viva pepa uedso abusadores viva gente **acostumbra** pepa

Topic #13:

derechos ser humanos derechos humanos un ue **procurador** derechos humanos procurador derechos procurador ser ue **presidente** ufmodo humanos **firme** uedcilmente ser ufmodo **poder** un procurador firme firme dif ser ue ufmodo un procurador derechos

Topic #14:

queremos ufoles **espa** ufoles espa vaya well **entendido** co djxrvbj well vaya **catalanes** espa ufoles entendido ser espa ufoles ser espa vaya catalanes **queremos** vaya catalanes well vaya ufoles entendido co djxrvbj queremos ser ufoles entendido co **queremos ser espa** djxrvbj

Topic #15:

mujeres en pol uedticos uedticos pol uedticos mujeres pol uedticos mujeres ezeiza vivo ezeiza **chicas porlosderechosdetodas** ezeiza chicas pol vivo desde ezeiza **en vivo** desde vivo desde en vivo desde desde ezeiza desde ezeiza chicas uedticos mujeres porlosderechosdetodas

Topic #16:

es **personas todas pueblos nuevo respete** nuevo **pacto es necesario** nuevo necesario nuevo pacto necesario pacto respete pueblos **constituyente** necesario nuevo es necesario **todas personas** decidi **sentidcomun** sentidcomun es **sentidcomun es necesario pueblos todas personas**

Topic #17:

ue **solo lucha** ueda sino ufn **sur** pa ueds sur bogot **sur bogot** ue pa ueds bogot ue **puede ciudadan** se bogot ufn sino **gran ciudadan** Ueda

Topic #18:

la los **marcha animales nueva familia** uedas **primera** la familia **maldonado** familia maldonado **convoca** la familia maldonado

maldonado convoca familia maldonado **polic octubre** marcha los polic uedas **excedieron** polic uedas excedieron **violaron primera march**

Topic #19:

asambleas asambleas co **cuando luchando** dem **civil** ufcratas dem ufcratas dem ufcratas luchando civil **barcelona** ufcratas luchando ufcratas luchando asambleas luchando asambleas co luchando asambleas **gentequelucha rodeaban guardia** rodeaban gentequelucha cuando rodeaban **guardia civil Barcelona guardia civil**

Topic #20:

ufn ues **humanos civiles** pol uedtica **represi** ufn pol uedtica ufo represi quieren **habla no vulneraci** ufn vulneraci la social **organizaciones civiles** martinhonoriega **guerra negaci**

Topic #21:

uedticos pol uedticos pol **llaman observadbinario** observadbinario bin **ciudadano** eso bin **uegimen** uedticos empre llaman **abstencionista** pol uedticos empre empre ciudadano **destruido** eso llaman eso llaman abstencionista llaman abstencionista ciudadano observadbinario bin eso destruido pol

Topic #22:

humanos derechos derechos humanos los los derechos los derechos humanos una **violaciones** ufn **impuestos** humanos co **viola vulnera sindicales** ikeaspain **avanzado respeta** el **bajar atenta**

Topic #23:

mirandoalfuturo segura forma prioridad mirandinos disfruten segura **estable** disfruten **plenamente** mirandoalfuturo prioridad plenamente plenamente forma plenamente forma segura mirandinos disfruten plenamente disfruten plenamente forma mirandoalfuturo prioridad mirandinos estable **mirandinos disfruten mirandinos prioridad** prioridad mirandinos **forma segura estable** disfruten

Topic #24:

el ue est **ppopular** ued **gobierno** est ue **libertades catalanes violencia** ufe **javiermaroto** ppopular javiermaroto **fundamentales** libertades catalanes ueis la si alexgimirizaldu javiermaroto aqu ued

Topic #25:

udd **ahorapodemos** udc vez **cada** cada **vez** ude udd udc ueticos **democr** ueticos udd ude democr puede recortar puede ude udd **vivimos** vivimos **momento asfixian** ueticos **correa** cada correa cada vez udcf vivimos

Topic #26:

defender uen tambi **tambi** uen ufn me **catalanas** iunida **catalanes catalanas** tambi iunida me **orgullosa** orgullosa defender iunida me siento **me siento** orgullosa siento orgullosa defender siento orgullosa siento **me siento orgullosa defender catalanes** vosotros

Topic #27:

humanos co **crecimiento derechos** humanos co **industriaguate** uemara **industria guatemala manifiesta** manifiesta **procurador derechos** manifiesta procurador manifiesta guatemala manifiesta procurador uemara industria humanos co jja uemara industria guatemala industriaguate uemara industriaguate industria guatemala manifiesta industria guatemala industria jja kzkt uemara

Topic #28:

todos ufos hace ufa **podemos luchan catalu** ufa no ued **defendiendo** catalu **nderechos nada cuando robaban mujeres hoy laborales** expositoorteg ni sigue

Topic #29:

democracia siempre mientras ueis uelogo di uelogo di **laborales** mismos **derechos** toda **democracia** derechos ufol ufablica **estado espa** ufol estado espa rep ufablica rep espa ufol dedic

5.2.3. Atribución latente de Dirichlet (LDA) de la muestra A:

Por último, se observan los tópicos obtenidos con la atribución latente de Dirichlet:

1. Pseudocódigo para esta aplicación:

Comienzo

Importar bibliotecas de métodos de Scikit-Learn (Python). Feature_extraction.text y sklearn.decomposition

Leer archivo de texto con los tweets

Crear matriz de bolsa de palabras X (método sklearn_CountVectorizer)

Sopesar matriz tfidf (método sklearn_TfidfVectorizer)

Procesar matrix con el algoritmo LDA (atribución latente de Dirichlet)

Escribir la salida como Conceptos o Tópicos principales de la matriz X.

Fin

2. Salida del algoritmo:

"Fitting LDA models with tf features, n_samples=1144 and n_features=13000.../home/meereslicht/.local/lib/python3.6/site-packages/sklearn/decomposition/online_lda.py:532: DeprecationWarning: The default value for 'learning_method' will be changed from 'online' to 'batch' in the release 0.20. This warning was introduced in 0.18. DeprecationWarning)

"Topics in LDA model:

Topic #0:

ufn **libertad expresi reuni deben manifestaci garantizar autoridades estatales amnistiaespana catalanas** ht ued **constituci** co el **ahorapodemos garantice nueva convivencia**

Topic #1:

democracia venezuela respeto gente sep **siempre di** eso uelogo **cambiar pepa** uedso **abusadores viva reclamar acostumbra siendo** el **laborales mismos**

Topic #2:

ueis co **estado** ufol **espa** qu dedic **mientras** ue **planeta libertades parece hab** aqu **pues pisotea catalanas** xavierantich baj de

Topic #3:

pol uedticos bin **observadbinario eso ciudadano llaman rajoy empre abstencionista destruido puede culpa** uenchez **recortar iceta advierten uegimen sino criminal**

Topic #4:

ufos **hoy hace** ni **mujeres podemos todos onu nada argentina protege** aleli **votar** ucs **tendremo nderechos adoctrinamiento xenofobia** ubfquien **odio**

Topic #5:

humanos co **derechos defensor ana** pol uedtico **niexcusas santa elsalvador joven joseluissafie cristiano quieren procurador** ny **mientras industria manifiesta** uemara

Topic #6:

bien cada as que ued **ciudades somos** ueda ma ufana **trabajo reforma sigue iniciamos** ufa **zaragozaencomun defienden defensa policia muchachos**

Topic #7:

ufn ue co ueda ues **solo** el **humanos** uf **sur** sino **bogot** ufo **ciudadan represi lucha gran petrogustavo ciudadanos deja**

Topic #8:

uen est los **lado** esos **claro der importa queda roben radicales patriotas soviets paulavazqueztv** co **tambi** ufn tur **nunca** de

Topic #9:

pa ueds pp **justicia vulneran** ufa **activistas mejor mismos social psoe mierda conocer incluye civiles** uedses **castigan naciones unidas espa**

Topic #10:

udd udc co ude **marcha** la los **animales familia condena primera maldonado nueva convoca edgardorovira excedieron** march **violaron**

Topic #11:

ued uf **cuando fiscal vickyrosell soria** uedas cu **vulner foto** mar uellar **firman querella demoledora** ufos **ddhh vida matibagnato** ue

Topic #12:

lo udn para ufanico **constitucionales** co **seguro congreso** mxcdmx **mexicano lugar laborar** patr **brindarte tienes publico_**es **vulnerar materias matriz contenidos**

Topic #13:

agarzon co **parte trata nazis espa** ufoles **ahora marianorajoy ddhh alfahispania** mapihema **ernestomonino llena somos rosachavarri espera retornados violan garantizar**

Topic #14:

pasa **catalunya fuego** hanspetert **civismo realmente** lasextatv penjatcat **pueb gabrielrufian** harold_root_ oposici nquljkwd prendi agria **banderita agitaba** uf **gobierno instituciones**

Topic #15:

vez **cada momento ahorapodemos** ues **correa** ueticos **democr asfixian vivimos** udd **aprieta** udcf les mas tal **toca reaccionaria colombiana** yngbgvewfw

Topic #16:

ufa espa **derecho romper reclamar decidir violencias reventar expresarse permitir fasc** mmunera **renuncia votar porlagoma venezuelanoseabstiene** ue si co **proteger**

Topic #17:

Vot **carta ciudadanos** uedan sanci **septiembre** veo **utampoco** useg **votaste** upu **imponerme puedes** condedegondomar ampl **rehenes conceden** hizo **kirchnerismo** alfredodarrigo

Topic #18:

catalanes co ser **queremos** ufoles **espa** djxrvbj well **vaya entendido sido millones golpe base animales** cosa ufablicos **pisoteados** jaumeclotet **protegiendo**

Topic #19:

partir invertir mensuales ucnimo uac voy **hola** co **manipule ong** lindepe cvhjyzfe **obligaciones golpistas acuerdo diputados vimos estoy violaban albert_rivera**

Topic #20:

defensa ufn no es ce **causa humanos** co rol **oculta documentaci imposible** a **casos exista** ff **ministro vicios libertades govern**

Topic #21:

co **alta** es **defienden personas** uen **vosotros mensaje experiencia conflictos envianos mivecino** ufasica vmqunqsti ubftenes **contanos** cn **comentario ejercicio** qui

Topic #22:

co **ser debe toda** la **democr** ueda **mismos pueblo** ueticos ufn **perder privilegios legalidad apoya provoca sus violencia mismas ver**

Topic #23:

no **derechos** ufn **libertades ley pol mujer** ueda **sede** esta ufos uedticos **voto atacado noche ciutadanscs** uedmetro **cederemos** cent cs_esplugues

Topic #24:

La **albert_rivera pero todo** compa uferos **apoyo** uen cs_esplugues **democ sedes** uesima **amenaza nonoscallar acuerdo golpistas diputados** ufn **violaban estoy**

Topic #25:

co en ues pol **asambleas mujeres** ufn uedticos ezeiza **chicas vivo desde porlosderechosdetodas civiles** uedtica **luchando** dem **cfkargentina barcelona** wiz

Topic #26:

defender me **catalanas orgullosa** iunida **catalanes tambi** uen **siento derecha habla** melnick cr hablan andresad **fascista milicos democracia ejercer** Rutilio

Topic #27:

> **sociales idea bajar una avanzado impuestos implicancia sola desarticular servicios micronauta independencia cuidado** co las das **nderechos votofemenino** si **consulta**

Topic #28:

> com ens pepbros **pagar espanyols** cul donin volem ells pel dropo **catalans** els fer **sempre** ser **callar govern mirthalegrand** lanacion

Topic #29:

> **nuevo todas constituyente** es **pueblos respete personas gobierno necesario pacto sentidcomun decidi libertades ppopular** el ufe **marianorajoy catalanes** est **garantizando**

5.3. Aplicación sobre un segundo conjunto de datos de texto (Muestra B)

El día viernes 29 de septiembre de 2017, siguiendo el mismo procedimiento anteriormente descrito, se inició la recuperación de un conjunto de datos en streaming desde el API de Twitter. La captura se inició a las 13:45 y se extendió hasta las 14:34, para un total de 3415 documentos o tweets, y un número de 32407 atributos, cuya semántica giraba, como el primer conjunto de datos de texto, alrededor de la palabra “derechos”.

5.3.1.Análisis semántico latente o LSA del segundo conjunto (muestra B):

El análisis semántico latente o LSA arrojó los siguientes conceptos (Tabla 16):

Concepto 0	Concepto 1	Concepto 2	Concepto 3	Concepto 4	Concepto 5
individuales fundamentales las **medidas referendumcatalan** medidas referendumcatalan	xsalaimartin **espa** espa ufa	Espa **constituci** tan	**colisiona fundamentales estrategia** colisiona estrategia colisiona fundamentales	**expertos onu** expertos **ind**	**humanos** qu **consejo** derechos
Concepto 6	**Concepto 7**	**Concepto 8**	**Concepto 9**	**Concepto10**	**Concepto11**
ufn **espa** **constituci**	ufn uen co	**Espa** **humanos** ufa	ufn uen est uen **violando**	**gobierno** **violaciones** **violando**	ufa uen ue
Concepto12	**Concepto13**	**Concepto14**	**Concepto15**	**Concepto16**	**Concepto17**
Si **humanos** est	est uen **humanos** **debe**	Si **constituci** ufn ufa	**espa** xsalaimartin uen	**constituci** ufn xq constituci	**espa** ufa qu si
Concepto18	**Concepto19**	**Concepto20**	**Concepto21**	**Concepto22**	**Concepto23**
fundamentales **constituci** ufn **espa** ufa	**fundamentales** **onu** xsalaimartin	ufn espa ante **violando**	**onu** **constituci** **espa** ufa	**constituci** qu **fundamentales**	**espa** ufa **fundamentales** xsalaimartin
Concepto24	**Concepto25**	**Concepto26**	**Concepto27**	**Concepto28**	**Concepto29**
_infolibre directo **_infolibre** **medidas**	**_albertosuarez** los _albertosuarez _albertosuarez los medios	_el_**patriota farc_epaz** intlcrimcourt _el_patriota farc_epaz _el_patriota	las xsalaimartin **fundamentales**	**violar** qu ue **espa** ufa	**individuales** **espa** ufa **violar**

Tabla 16: Salida del algoritmo LSA en la muestra B

5.3.2. Analisis semántico latente probabilístico o PLSA para el segundo conjunto de datos (muestra B):

El análisis semántico latente de índole probabilística o PLSA para el segundo conjunto de tweets arrojó el siguiente resultado, para un número de 3415 documentos y alrededor de 32400 atributos, en 1000 iteraciones.

Salida del algoritmo:

"Fitting the NMF model (generalized Kullback-Leibler divergence) with tf-idf features, n_samples=3415 and n_features=32400...Topics in NMF model (generalized Kullback-Leibler divergence)":

Topic #0:

> las **onu parecen violar relatores referendumcatalan** parecen violar **individuales** parecen violar individuales violar individuales **fundamentales** referendumcatalan parecen violar referendumcatalan noticiasonu relatores noticiasonu **relatores onu** relatores onu las relatores onu onu las medidas medidas referendumcatalan las medidas referendumcatalan las **medidas** medidas

Topic #1:

> **obligaciones todas** nsi sos nsi sos **izquierda** izquierda **tenes** izquierda **grieta** nsi **sos** grieta nsi **tenes todas obligaciones** la sos tenes nsi sos izquierda tenes sos **izquierda verdadera** la verdadera tenes todas verdadera grieta nsi tenes nsi sos

Topic #2:

> si ufn **debe humanos violaciones** violaciones humanos xsalaimartin **silencio catalunya** ufa espa ufa humanos espa ufa ufn violaciones humanos silencio comisi ufn violaciones debe plantearse si si formar espa ufa catalunya silencio comisi ufn xsalaimartin ante silencio

Topic #3:

> uen ue est uen ufn est **avisa espa** ufa qu tan si qu ue **violando onu** avisa ufa uen violando ufn espa onu est uen violando xsalaimartin tan **sagrada** ufola

Topic #4:

> uedas **libertades** para **defienden** co para defienden **barcelona** para defienden **catalanes** defienden barcelona libertades defienden barcelona cannanan barcelona libertades cannanan para defienden libertades catalanes barcelona libertades catalanes cannanan para yotambiensoyfcse **yotambiensoyfcse** co uedas yotambiensoyfcse libertades catalanes uedas

Topic #5:

> **ahora ncuando cambiaron** ufn **constituci** ufn constituci **pago espa** ufasalealacalle **ufasalealacalle protestona** espa deuda ncuando ufn anteponer protestona ahora ufasalealacalle ahora espa ufasalealacalle ahora ufasalealacalle ahora ncuando **cambiaron constituci** constituci ufn **anteponer pago deuda cambiaron constituci** ufn

Topic #6:

> udd udc udd udc udc uddf udc uddf udc uddf udc uddea uddea udc uddf udc uddf udd udc co uddf udc udd udc udc **hispaniaspain** udc udc udc udc uddea co udd udc co udc uddea udc uddf udc uddea uddea

Topic #7:

> ues la ueds pa **derecho** pa ueds ufn **mejor** est est ues **demostraci** ufn **demostraci cadenas** pa ueds est **decadencia** ufn decadencia pa la mejor demostraci ues cadenas la mejor ues **cadenas presidenciales**

Topic #8:

> **defender vida asambleaecuador merecen adultomayor vida digna** defender adultomayor defender adultomayor prioridad digna merecen vida digna **prioridad** asambleaecuador merecen **prioridad** prioridad asambleaecuador merecen vida asambleaecuador merecen asambleaecuador merecen vida **solbuendia** defender adultomayor solbuendia solbuendia defender adultomayor prioridad

Topic #9:

> co nhttps nhttps co **vamos** ufn ud ben cert **personas** la **deoctubremarchaporlosanimales** sino **autor** ufana **sexuales** que **garantizar** ma ufana **trabajadores** puede de

Topic #10:

> **bandera defensa** espa **siempre defensa** ufa ufa bandera ufa bandera siempre siempre udoles siempre defensa **espa** espa ufa defensa espa defensa espa udoles espa ufa bandera espa udoles **cocochiquitin** espa ufa cocochiquitin cocochiquitin espa vkx hzejft vkx

Topic #11:

> se **onu** ueda **rajoy fundamentales estrategia** rajoy estrategia **colisiona** rajoy estrategia htt **onu advierte** colisiona fundamentales ve **gobierno rajoy estrategia** estrategia **colisiona fundamentales advierte** gobierno

rajoy onu advierte gobierno advierte gobierno rajoy **tableroglobal** se tableroglobal

Topic #12:

marcha octubre derechos uedculo **animales** art art uedculo uedculo publimetrocol derechos animales **publimetrocol plataformaalto** marcha derechos yoleopublimetro plataformaalto yoleopublimetro este marcha **derechos animales** art uedculo publimetrocol ufn yoleopublimetro edici ufn ufn marcha

Topic #13:

ue udd ufn ude **transmisi** ufn transmisi ude udd est ue uen est udd ude tri udcfa udd udcfa ua busca **transmisi** ufn **partidos record_mexico** ufa **slim**

Topic #14:

humanos derechos derechos humanos **venezuela** pa uedses uedses pa **onu consejo** consejo derechos humanos consejo derechos humanos co vtvcanal venezuela consejo venezuela consejo derechos **respaldo** ufn derechos ufn humanos en **ratifican onu**

Topic #15:

los no si **mujeres** pero los **derechos humanos viola pueden** es **ser zurdos derecha** pero si mujeres **no musulmanes** musulmanes **pueden abusar** pueden abusar **pueden abusar mujeres pero si blanco**

Topic #16:

siempre **tenemos concepto izquierda** ufanica **cambiar siempre** tenemos siempre tenemos cambiar cambiar concepto fargosi siempre tenemos fargosi siempre fargosi ufanica **propuesta** edgerome **impecable** tenemos cambiar tenemos cambiar concepto propuesta edgerome impecable fargosi **impecable concepto izquierda atrasa**

Topic #17:

ufn ued as as ued **regresi reelecci** regresi ufn **utilidades** reelecci ufn vicentetaianoec ufn **indefinida** indefinida **reelecci** ufn **indefinida** xq ufn indefinida regresi ufn xq **pensaron limitaron utilidades** pensaron as ued limitaron ued limitaron utilidades

Topic #18:

yotambiensoyfcse polic yotambiensoyfcse co yotambiensoyfcse co nkyddltko ueda **libertades** vsyrafagas vsyrafagas yotambiensoyfcse co

vsyrafagas yotambiensoyfcse **guardia polic** ueda guardia civil nkyddltko **catalu** ufa **defensa civil** yotambiensoyfcse apoyo yotambiensoyfcse apoyo polic co nkyddltko polic uedas

Topic #19:

el ue uf **quieren ser** est est ue **organismos humanos kirchnerismo** kirchnerismo organismos organismos humanos uf **desaparicion maldonado** maldonado uf **desaparicion kirchnerismo** organismos humanos humanos **quieren** el kirchnerismo **organismos** el **kirchnerismo desaparicion**

Topic #20:

mundo hacen espa urnas su manolo_coco ufasalealacalle espa ufasalealacalle mundo **exigiendo** su **sociedades** sociedades **catalana** sociedades catalana **hacen** hacen **avanzar** mundo mundo exigiendo mvtdiasoct sociedades mvtdiasoct sociedades catalana mvtdiasoct ufasalealacalle mvtdiasoct sociedades ufasalealacalle **mvtdiasoct** manolo_coco urnas espa

Topic #21:

ley defensa defendiendo ciudadanos yotambiensoyfcse protegiendo siempre protegiendo siempre siempre protegiendo **defendiendo** defendiendo **ciudadanos** defensa protegiendo defendiendo ciudadanos protegiendo defendiendo yotambiensoyfcse co yotambiensoyfcse co hldjti ley yotambiensoyfcse co ley yotambiensoyfcse hldjti defensa ley yotambiensoyfcse defensa ley defendiendo ciudadanos

Topic #22:

ueda ni ufez ufo **derechos** ni derechos ueda **nacional protecci** ni ufez nacional ueda nacional derechos ufn nacional derechos ni co nacional derechos derechos ni ufez protecci ufn ufas ufos **gobierno**

Topic #23:

uf **democracia** son **unidos gobierno garantes libertades jaarreaza denunci** uf **estados unidos** denunci **venezuela** estados **yotambiensoyfcse** garantizan derechos libertades son yotambiensoyfcse yotambiensoyfcse **garantizan garantes democracia dios** garantizan **derechos** garantizan

Topic #24:

ufa **espa** ufa espa **onu ddhh catalu mismo** catalu ufa lo xq **expertos violar** uedctima las **medidas** ufa co **parecen** violar ind **preocupantes** xq violar ind onu ddhh espa

Topic #25:

ufn **hoy** ue **selecci** selecci ufn uexico **personas** ufn selecci ufn ufn selecci ufn **nacional** nacional selecci ufn nacional **transmisi** ufn selecci ufn mexicana voy transmisi **honestidad** el honestidad ue **haciendo** selecci **haciendo selecci** ufn

Topic #26:

humanos pol defensor uedtico pol uedtico **joven libertad** uedtico joven **cristiano** uedtico joven pol uedtico joven **democracia** joven cristiano **joseluissafie** pol uedtico joven cristiano defensor joseluissafie pol **joseluissafie cristiano defensor defensor humanos retweeted** cristiano

Topic #27:

televisa ufos ufn tri **transmisi** ufn transmisi **slim** udn **lucha ofrece** ufn tri ue transmisi ufn tri todo **sociales** mdd nbc **trabajar** ofrece mdd mismo

Topic #28:

ufn uf **favor vulneraci** vulneraci ufn **protege represi** represi ufn **pueblo asociaci** ufn ufn protege cam ingres ingres uf ingres uf **inmigrantes** inmigrantes inmigrantes **banderas** inmigrantes banderas **remeras protege** cam favor **dreamers** asociaci

Topic #29:

fundamentales gobierno onu fundamentales co onu gobierno **estrategia colisiona** estrategia colisiona fundamentales estrategia colisiona colisiona fundamentales onu gobierno **estrategia gobierno** estrategia gobierno estrategia **colisiona aviso onu** gobierno colisiona fundamentales co la aviso onu pide la onu publico_es

5.3.3. Atribución latente de Dirichlet para el segundo conjunto de datos (muestra B):

Para el mismo número de documentos y atributos en 5 iteraciones (iteraciones recomendadas):

Salida del algoritmo:

“Topics in LDA model:

Topic #0:

humanos pol respeto defensor co uedtico **democracia joven** refer **joseluissafie cristiano** uendum **retweeted** niexcusas toda uflo **legalidad** mano jefebebesv esto

Topic #1:

ues pa **derecho** ueds la **mejor** est **cadenas atentan demostraci decadencia bmarmoldeleon presidenciales** co **nadie** ufan **parte** ir dar ma

Topic #2:

ufn **salud igualdad trabajar contra obligados defiende paz veo partido racismo tomar corrupci padres discriminaci temas revolufashion periodistas amor educaci**

Topic #3:

ued as **mujer** cu **marchadiversidad conciencia beatrizandrino proceso toma faltan ucderechos** re **televisivos** forma **pedimos pasar plan** manumarlasca uedrse yrsco

Topic #4:

los co no **mujeres humanos** si pero la **viola derecha** es **pueden zurdos** hog arihagadol **abusar piropea musulmanes blanco** am

Topic #5:

derechos co **humanos** ufn en **venezuela** ufos de **canciller** una **jaarreaza** el **vencancilleria denunci estados unidos maniobra fallida** ufas **septiembre**

Topic #6:

ahora espa ufasalealacalle constituci ufn **pago protestona deuda anteponer cambiaron sociale ncuando hacen mundo urnas** su **trabajadores** manolo_coco **catalana avanzar**

Topic #7:

uf co nhttps ufn ufez **laborales gracias** la **vamos favor asociaci vulneraci colombia protege argentina** misi un el **seguridad remeras**

Topic #8:

> **mismo** lo ven uedctima **primero** hay eu ddhh uedctimas **autor victimario** desp **sexualmente abusado winston_dunhill online nslim carlos**_ponz mddd **ofrecen**

Topic #9:

> **izquierda** sos nsi **tenes** la fargosi **obligaciones** todas **verdadera grieta concepto cambiar tenemos** ufanica **siempre propuesta** edgerome **impecable atrasa fiscal**

Topic #10:

> **memoria catal urgente** esma **arte** zgkptsitd **pagina elmundoes respeta lenin profesora guerra merece piquete planta** aqu **informativo** juan **critica represores**

Topic #11:

> **defensa siempre** co **bandera** yotambiensoyfcse cocochiquitin **defendiendo** udoles **ley libertad ciudadanos protecci protegiendo** ufn hzejft vkx cannanan hldjti **deberes primera**

Topic #12:

> co me ue tv **tema violan** yo_soy_asin pasan **bajo instituciones luz** porque salida actuaciones darse uylzjeawl **lgtb luchadora deber fallecimiento**

Topic #13:

> **frente servicios contrato abierta** ueticos **ceofanb comienzan profesionales clubes** espnsutcliffe **virtudes complejo reconocemos** uedadelasecretaria **ministerial deshonrosa cobarde postura negociaci** uenime

Topic #14:

> **personas** hoy co ue ufn **gente** con **sociales** van voy **defensores** da dos **diversidad sexuales comit marcha** sesi yo **pensar**

Topic #15:

> ueda solo **trabajo** ue **guardia civil nunca velar polic igual seguir** ufanico **participaci luchan nacional manera sin ciudadanos colegio** estos

Topic #16:

> **catalu** ud **democracia** yotambiensoyfcse ufa **sino apoyo** por que **libertades** co **derechos civiles son garantes constitucionales** udf **dios garantizan defienden**

Topic #17:

> **nuestros votar ganan cuerpos usaban hijos nsabemos perfecta ilegalizaci maquinaria votarem** dir **esconder traslamanta marchamos** mar decir aterrizar ht compah

Topic #18:

> udc udd uddf uddea co **recortar yotambiensoyfcse** vez **popular partido hispaniaspain** bien deuda ccifuentes vendr ue luego dnnt subversivos_ due

Topic #19:

> co **catalanes libertades** uedas para **defienden** yotambiensoyfcse **barcelona** cannanan ser **animales deoctubremarchaporlosanimales** jusapoljscat txarito nkyddltko vsyrafagas ueda seguro chiguiroestudio **vulneran**

Topic #20:

> **onu fundamentales** las **violar medidas parecen individuales** noticiasonu **relatores referendumcatalan gobierno estrategia colisiona** se **rajoy advierte** htt **expertos** ve ueda

Topic #21:

> **estudiantes** te **justa entonces colectivo patria leches igualitaria banderitas dejaros feliz cumplea mbachelet alcalde** alcaldia_ss espnmx fmlnoficial bonhamled **manos importa**

Topic #22:

> ufn **derechos** ni **nacional selecci octubre marcha animales** ufo art uedculo hoy ued ue el as xq yoleopublimetro publimetrocol plataformaalto

Topic #23:

> udd ufvenes ufn **exigen total solidaridad cientos** mi **grande concentraci** sismo **padre universidad salamanca** pobladores javiera_oliv lalegua gerard ouisse **amenazas**

Topic #24:

> espa ufa si ufn xsalaimartin **humanos debe comisi violaciones catalunya silencio** ante **plantearse formar** qu **onu** ue tan uen **avisa**

Topic #25:

> **gobierno consejo humanos** onu co uedses **respetar vtvcanal violaci** pide **respaldo arreaza ratifican eeuu rechaza** dice ue **sanciones cuba califica**s

Topic #26:

> **pueblo voto miedo** os **brutal video** gif link dejo udamate antoniobanos_ enga huelgaalsaqueo siguen **amnistiaespana** dejen **entregaste** volver jus pe

Topic #27:

> **defender vida asambleaecuador adultomayor merecen digna prioridad solbuendia** tambi ex uesicos **presidente congreso justicia** jorge garc laguardia cc **entrevista** mario

Topic #28:

> est ue uen udd ufn co **transmisi** tri el ude **slim televisa** udcfa uf ua este **quieren humanos partidos organismos**

Topic #29:

> co cert ben hola efta psbznhxq nil_ag **civil** pro martinciccioli ddm_oficial **registro propiedad** del gonzalez_lucas lucasbertero karinaiavicoli diegoleuco silviafbarrio luisbremer

5.4. Aplicación sobre un tercer conjunto de datos de texto (Muestra C)

El día 7 de octubre de 2017, siguiendo el mismo procedimiento anteriormente descrito, se dio comienzo a la captura de un conjunto de datos en streaming desde el API de Twitter, utilizando los mismos procedimientos descritos arriba. La captura se inició a las 17:27 y se extendió hasta las 18:10, para un total de 2020 documentos o tweets y 20881 atributos, cuya semántica giraba, como los anteriores conjuntos de datos de texto, alrededor de la palabra “derechos”.

5.4.1. LSA de la muestra C:

El análisis semántico latente de la muestra C arrojó los siguientes resultados (Tabla 17):

Concepto 0	**Concepto 1**	**Concepto 2**	**Concepto 3**	**Concepto 4**	**Concepto 5**
aqu ued **catalu** catalu ufa **murcia** est ue **pasando**	Pidi **aprovecharte** aprovecharte **represi**	**Espa** espa ufa lleva	**derechos humanos** seg **estrasburgo** humanos seg	**bandera** si lo	**constitucional** ueda pidi
Concepto 6	**Concepto 7**	**Concepto 8**	**Concepto 9**	**Concepto10**	**Concepto11**
ue en **derec**	ufn **derec** lleva	en lo si rosadiezglez	**espa** ufn bandera	ue si **bandera**	ueda ued aqu
Concepto12	**Concepto13**	**Concepto14**	**Concepto15**	**Concepto16**	**Concepto17**
ue ueda est	ueda ufa est	ue **derec** **rosadiezglez**	ued ufa est	**caritas pontifex**_es caritas caritapontifex_es **debemos**	aqu est aqu ued
Concepto18	**Concepto19**	**Concepto20**	**Concepto21**	**Concepto22**	**Concepto23**
est est ue nhttps	nhttps aqu aqu ued	_espatricia mi **patria** _espatricia mi **_espatricia**	esto **represi** represi ufn	**murcia** represi **represi** ufn	**gente** nhttps grancocolio
Concepto24	**Concepto25**	**Concepto26**	**Concepto27**	**Concepto28**	**Concepto29**
gente _lai_lai_ _lai_lai_ si	_lai_lai_ si _lai_lai_ **murcia**	**Murcia** est ue esto	est ue **palos** **luchan**	est ue grancocolio _lai_lai_ si	**luchan** aqu ued nhttps

Tabla 17: Salida del algoritmo de LSA en la muestra C

5.4.2. PLSA de la muestra C:

Fitting the NMF model (generalized Kullback-Leibler divergence) with tf-idf features, n_samples=2020 and n_features=20881...Topics in NMF model (generalized Kullback-Leibler divergence):

Topic #0:

ued est ue **gente murcia** est ue esto ueda **catalu palos** grancocolio **luchan** ufa **polic** ueda polic ued catalu ued catalu ufa ue **pasando** aqu ufa murcia palos gente

Topic #1:

en ufn **represi represi** ufn **derec pablo_iglesias_** pidi **lugar** en lugar uendole **gente renuncie** uendole gente uendole jordigraupera pablo_iglesias_ jordigraupera **fracasada renuncie** renuncie derec **aprovecharte** ufn fracasada pidi ufn fracasada

Topic #2:

humanos ufn humanos humanos co ufn ufn humanos co tumaco onu **masacre tumaco ocurre contexto** tumaco ocurre masacre tumaco **desprotecci** ufn masacre **onu** masacre tumaco ocurre masacre tumaco ocurre ocurre contexto desprotecci ufn humanos onu ocurre contexto desprotecci contexto

Topic #3:

ufa lo **espa** si **lleva** espa ufa **protege viva** rosadiezglez lo si rosadiezglez lo si si lleva **bandera** lo si lleva lleva bandera lleva bandera **constitucional** protege viva protege viva constitucion rosadiezglez lo ufa protege ufa protege viva

Topic #4:

libertades queremos autoritario olgarodriguezfr queremos **independizarnos queremos independizarnos gobierno recorta** olgarodriguezfr muchos olgarodriguezfr **muchos queremos** gobierna **libertades** gobierna muchos muchos queremos recorta libertades gobierna autoritario recorta **muchos queremos independizarnos gobierno autoritario recorta** independizarnos gobierno autoritario recorta libertades autoritario recorta **libertades**

Topic #5:

uf el ufan nhttps **humanos** nhttps co intent **gobierno intent** uf seg ufan co seg ufan **estrasburgo legislar derechos** humanos uf legislar uf legislar derechos seg ufan estrasburgo humanos seg ufan legislar derechos legislar

Topic #6:

hoy ufa ufana espa espa ufa **defendiendo** sido ma ufana udd **barcelona** ma **unidad madrid** ufana barcelona defendiendo ufana barcelona **unidad espa** ma ufana barcelona unidad espa ufa udd udce udce

Topic #7:

ufn pol co **nueva tsj** uedticos ufn **asegura nueva rond** ufn ufn asegura rond pol uedticos **sentencia rector luis rond** ufn asegura tsj **restringe** pol tsj restringe **rector luis emilio** restringe pol restringe

Topic #8:

ufn **constituci** constituci ufn **naturaleza lenin** ufanica **mundo somos garantiza** garantiza naturaleza ufanica mundo ufanica mundo garantiza ufn ufanica **somos consecuentes constituci** ufn ufanica mundo somos consecuentes mundo garantiza naturaleza mundo garantiza lenin somos consecuentes consecuentes constituci ufn

Topic #9:

no **todos macri** mas uniciudadanaar **reconoce pocos congreso vota** congreso vota congreso todos **mayorias** vota congreso reconoce mayorias reconoce mayorias vota **no reconoce mayorias** no reconoce pocos no pocos no reconoce todos uniciudadanaar mayorias mayorias vota

Topic #10:

rousepaez **catalanes quitado sentimos** cat quitado cat **mentir** rousepaez lndesafiocat lndesafiocat rousepaez lndesafiocat **vasta** vasta lndesafiocat vasta mentir **coaccionado** coaccionado **humillado quitado mentir** coaccionado vasta mentir coaccionado humillado quitado cat catalanes lndesafiocat vasta vasta mentir

Topic #11:

co ue los de **pueden personas mayores** uedan que ud se tambi ufas uedhayrumbo tambi uen ser **siempre** pi va pero

Topic #12:

por **respeto humanos nuevos** respeto humanos **graduandosdepaz egresan unestachira** co czp unestachira unestachira co por respeto **respeto humanos** egresan graduandosdepaz por respeto humanos egresan **nuevos** humanos **egresan profesionales** profesionales unestachira nuevos profesionales unestachira nuevos profesionales profesionales unestachira co

Topic #13:

ueis **amenac vamos muerte quit** por **dejar** vamos dejar quit quit ueis vamos dejar dejar quit ueis dejar **quit** muerte vamos **nacionalidad albert_rivera** por amenac quit ueis nacionalidad ueis muerte ueis muerte vamos muerte vamos dejar ueis nacionalidad

Topic #14:

ueda uen ufos **llunacatalana** ntambi ntambi uen ueda hablar **constitucional hablar** hace ufos hace hablar **quiero** yo constitucional **recortaba** pnique hace ufos constitucional recortaba nahora quer est uen

Topic #15:

humanos violaciones para oct **violaciones** humanos oct co **paguen vota** oct co vota vota oct violaciones humanos vota humanos vota oct uedmenes uedmenes violaciones uedmenes violaciones humanos paguen cr uedmenes para paguen paguen cr para paguen cr cr uedmenes

Topic #16:

la ufn **mujer una defender fundamentales josepuntcat evitar pues desnuquen** una mujer josepuntcat defender evitar desnuquen evitar desnuquen una **una mujer arrastrada** mujer arrastrada la mujer arrastrada una mujer desnuquen **fundamentales** evitar **desnuquen la cabeza sedici**

Topic #17:

ue **derechos humanos** ufn derechos humanos qu **mud** qu ue **violaci** ufn violaci ue **expediente** expediente humanos cne humanos cne co mud **abrir** sepaque mud abrir expediente **violaciones derechos** violaciones derechos ue expediente violaciones sepaque mud

Topic #18:

derecho pa ueds pa ueds nporque **amo respeto estado tierra** porque nporque reinasonia ueds nporque ueds nporque amo respeto estado pa ueds nporque respeto estado derecho porque quiero pa porque quiero tierra nporque respeto quiero pa ueds

Topic #19:

los ufn **futbolistas proteger** ufn proteger proteger co los futbolistas **mexicanos** mexicanos mexicanos **fundan** mexicanos fundan **asociaci**

futbolistas mexicanos los futbolistas futbolistas mexicanos fundan fundan fundan asociaci ufn ufn proteger co fundan asociaci asociaci ufn proteger asociaci asociaci ufn

Topic #20:

ue **liliantintori lucha mundo seguimos denunciando impedir nada** res liliantintori nada impedir liliantintori nada ue sigamos mundo **crisis humanitaria** ue sigamos denunciando sigamos sigamos denunciando **sigamos denunciando** mundo impedir ue sigamos mundo crisis nada impedir

Topic #21:

libertad fascistas sociales extra visto labordeta npues todavia visto nvamos ufais ufais piten ufais **piten canto** npues **todavia habeis libertad** labordeta libertad labordeta npues piten piten canto npues labordeta npues piten canto libertad os extra os

Topic #22:

defender debe mexicano miedo desaparecer miedo **futbolista** miedo futbolista mexicano desaparecer co futbolista mexicano defender mexicano defender mexicano defender debe futbolista defender debe desaparecer futbolista mexicano defender debe debe desaparecer co debe desaparecer **toda democracia humanos**

Topic #23:

ues **mujeres** el **quieren feminismo** ve ser pelo ueda **deber ocuparse** el feminismo ues ucstas mujeres ues ucstas garzageugenio el quieren pintar pelo quieren pintar ocuparse ues ocuparse ues ucstas mujeres quieren

Topic #24:

udc udd **defensa vida** udd udc **derechos marcha octubre** ufana ueda **animales** ma ue udc udc udd udc udc derechos animales **comunidad** ma ufana ude natal

Topic #25:

mejor ueds **puede** pa ueds pa **uniciudadanaar vivir esperanza** puede vivir vivir pa esperanza puede ucvamos **devolverle argentinos** uniciudadanaar ucvamos uniciudadanaar ucvamos devolverle vivir pa ueds ucvamos devolverle **ucvamos** pa ueds mejor ueds mejor puede vivir pa puede vivir

Topic #26:

ufa **policia proteger guardiacivil dignidad vamos** zoidoji ufa **libertades ciudadanos** libertades **ciudadanos integridad** espa libertades integridad **espa** ufa jugando integridad espa nos jugando integridad nos jugando nos **jugando** integridad ufa libertades jugando **proteger** dignidad

Topic #27:

si ufn **derechos trabajo** menos no uelogo di di uelogo **futuro tenes** que **banderas quiere trabajo tenes techo macri vota** tenes techo tenes techo **comida techo** comida **derechos**

Topic #28:

humanos trabajadores ufn habla uf as **defensor** defensor humanos si **derechos** defensor humanos los ued los derechos los derechos humanos ex si polvial_goboax derechos humanos eso **ahora**

Topic #29:

laborales gobierno pueblo colombia reclama ufn **vulneraci** vulneraci ufn ufn **animales** esta **justicia** partidopacma pueblo **reclama propio gobierno** propio **santander** pueblo reclama santander pueblo santander rogerdhio rogerdhio esta colombia

5.4.3. LDA de la muestra C:

Fitting LDA models with tf features, n_samples=2020 and n_features=20881.../home/meereslicht/.local/lib/python3.6/site-packages/sklearn/decomposition/online_lda.py:532: DeprecationWarning: The default value for 'learning_method' will be changed from 'online' to 'batch' in the release 0.20. This warning was introduced in 0.18. DeprecationWarning)Topics in LDA model:

Topic #0:

udc udd co **animales defensa laborales mejor colombia** ud **derechos mexicanos marcha puede** ma ufana **jugadores vida domingo** ues amfpromx

Topic #1:

co ufn pol luis tsj rond uedticos **rector sentencia nueva asegura restringe emilio** de **pueblo** ufos **ley** qu **mismo golpe**

Topic #2:

> el ueda ues uen **mujeres quiero hablar** quer **ahora** hace **estado** ufos **quieren recortaba llunacatalana constitucional** yo **deber feminismo** nahora

Topic #3:

> ueds pa uedas **vivir pueden** di **argentinos devolverle** ucvamos **esperanza conoce social** uelogo **encima carro paz deben aborto** abalosmeco ximopuig

Topic #4:

> son uedan **fortalecen instituciones** he **respetar** dos **injusticia** nace uexico su **entrenando discuten ccsoficial invencible exigen frecuentemente** ieew invencibles qkvxwz

Topic #5:

> nni ua **personas sociales** ni siempre co igua es rey **libertad dictaduras** carpereal **religi nazis nderechos imperialistas fascistas** dir ufoles

Topic #6:

> **derecho exigir cuando argentina mud pero reclamar restituyan esperar entonces** uctenemos viv **podremos thayspenalver comprendamos** vac **sicario desarmado soldado campesino**

Topic #7:

> por ueis ufn **constituci naturaleza somos lenin** ufanica **garantiza mundo vamos muerte consecuentes dejar** quit nuevos albert_rivera **amenac nacionalidad primera**

Topic #8:

> **futuro tenes ricos perdes comida techo dan mano conmovedora inspiracional botero** arte obra **josefinasalomon** deja periodistasu leal edgar esqueda **periodista**

Topic #9:

> al the in of are ifueputa **santos candela palacio** ech not police dar **catalonia** this **spanish people time charm offensive**

Topic #10:

> ces_ec **iguales** esantos anouka mnalvear wjbh svftc **clases pisoteando joder** savaje jorgeearcinieg **capitalismo rapido volar restituido asusta** roisinblit la_imposible **retrocesos**

Topic #11:

> **gobierno gobierna libertades queremos vota autoritario independizarnos** olgarodriguezfr **recorta muchos** no **macri todos** aliciacastroar para **congreso reconoce mayorias pocos si**

Topic #12:

> **proteger trabajadores futbolistas gracias fundan egresan** graduandosdepaz **guardiacivil** czp **profesionales unestachira** unes__tachira af **policia** zoidoji ex nhttps **dignidad presencia** ufa

Topic #13:

> **mayores** co **adultos derech** con **democr** uedhayrumbo pi **cambiarderumbo encuentro sociedad** ntros ufera **nuestra bachelet libre** cra npara sanisidro malecholakian

Topic #14:

> oct cr uedmenes ivdxwoalwt **ademocratica hacer gran foro** misi **fiesta animalista** ufnnevadoaragua **invita mira** juanarcones rd **dijo debemos** netflixes trato

Topic #15:

> ufan **intent** seg **estrasburgo legislar publico_**es ttddrlwme **violaci** te **corrupci realizamos compa imagino invitamos hablar** eval **activista funcionarios** ning **voy**

Topic #16:

> **lucha** uda solo ufmo este santamariamonic **sacrificio** monica lograste jos dem claro udn quiere campa ufablica henrycucalon **banderas huelga guerra**

Topic #17:

> **uniciudadanaar** uen **catalanes** rousepaez **quitado sentimos** cat si **trabajo mentir** lndesafiocat vasta **coaccionado humillado tambi** la pais **mierda** las **peor**

Topic #18:

ue nhttps ued est aqu ueda **murcia** esto **gente catalu luchan** ufa **palos** grancocolio polic **pasando protecci** eso **luchando mundial**

Topic #19:

nporque porque tierra mexicano libertade amo miedo desaparecer futbolista ya podr asi **civil declaraci unidos** jam **colaborador futbol arrebatarnos quitarnos**

Topic #20:

ciudadanos lgbt garantizar menos libertades defensores salud integridad ufa nos **banderas constitucionales existe** xdelucas cu estos **plenamente cuentan** espa **alguien**

Topic #21:

libertad extra **canto visto sigue** os npues labordeta **habeis** ufais **piten** nvamos **todavia** dictad fermont **dictadura** zmloc **den ejemplo** nfsk

Topic #22:

la **defender** ufn **mujer parte fundamentales** una **evitar** un **pues trump desnuquen arrastrada sedici** josepuntcat **cabeza** va da **tiempo** ueda

Topic #23:

paguen octubre igualdad violan ufo fe **docentes** ufas uefrica sud **jubilaciones cancelaciones navarro** olayadotel **escuelas vicios mala construcci injustas plan**

Topic #24:

no **ser mas vulneraci plumazo** ucf **caza** barquerosb partidopacma **congreso presentado consideramos tirano ninguna pasa frente partidos** der **entienden** elizwar

Topic #25:

co **humanos** ufn **derechos masacre tumaco onu contexto desprotecci ocurre** los **respeto violaciones asociaci debe noticia cne expediente** uc **mud**

Topic #26:

> uf ue **liliantintori derechos defensor mundo libertad seguimos impedir habla denunciando crisis nada** argiro **viviendo humanitaria sigamos** res ind **humanos**

Topic #27:

> as que **humanos luchar comunidad** hay robo ubb hijos **privilegios agenda llaman leyes** est rayovirtual lekaconk **comiendo hilo** ufores via

Topic #28:

> en ufn **gente derec represi pablo_iglesias_ lugar** pidi jordigraupera **fracasada aprovecharte** uendole **renuncie poder imagen catalunya** sur **organizaci reputaci** uba

Topic #29:

> **espa** ufa si lo **bandera** lleva **protege** viva rosadiezglez **constitucional constitucion** hoy udd **defendiendo** udce ufana **sido barcelona** ma **unidad**

Capítulo 6. Interpretación de los resultados.

El siguiente capítulo está dedicado al análisis de las salidas que se han obtenido en la aplicación de cada uno de los algoritmos en las tres muestras de tweets. Mientras el capítulo anterior ofrecía las salidas obtenidas por los algoritmos, en este capítulo se reconstruyen los datos observados en vistas del objetivo inicial que nos habíamos propuesto al iniciar este trabajo: la reconstrucción automatizada de la semántica en los conjuntos de datos de textos que conforman el corpus de documentos tomados de Twitter.

Como puede observarse, las técnicas que se acaban de aplicar sobre los conjuntos de datos de texto revelan una semántica implícita o, lo que es lo mismo, los tópicos implícitos alrededor de un concepto particular, en este caso, el de "derechos".

Los resultados obtenidos son enormemente interesantes para el científico social. Revelan las concepciones y definiciones semánticas implícitas que subyacen al uso que hizo cada uno de los usuarios cuyos tweets fueron recuperados para este conjunto de datos de texto en particular. En un sentido, los tweets son una suerte de encuesta de opinión que revela lo que este grupo de usuarios de Twitter en particular piensa del concepto "derechos": cuáles son las asociaciones, implicaciones, sentidos sugeridos y semántica implícita en el uso de este concepto particular.

Desde un punto de vista general, las aplicaciones de los algoritmos objeto del presente estudio en los tres conjuntos de datos arrojaron dos tipos de resultados claramente discernibles: 1) por una parte, ofrecen una visión reducida de los tipos de problemas, preocupaciones o dimensiones que los usuarios de Twitter tienen en el momento de emitir sus tweets. 2) Por la otra, ofrecen una visión general de la semántica que define el uso del término "derechos", y que es transversal a las preocupaciones o dimensiones que definen los tópicos o dimensiones reducidas, en los usuarios que emiten sus tweets en el momento de la captura.

Estos dos tipos de resultados no se ofrecen con independencia el uno del otro, sino que ambos constituyen el rendimiento esperado de la aplicación de un algoritmo de modelado de tópicos sobre un conjunto determinado de datos de texto. La distinción es, por tanto, más bien analítica. Ambos permiten al investigador dos tipos de tareas diferentes, dependiendo de su foco de interés.

Puede, por ejemplo, en primer lugar, reconstruir los tópicos que caracterizaron una determinada transmisión de tweets. En los tres conjuntos de datos se pudo observar que el tema de Cataluña es prominente para los tres conjuntos de datos, incluso estando separados por varios días. Igualmente importantes en los tres conjuntos de datos son las referencias a la situación

política de Venezuela, la situación política de Argentina y distintas referencias a crímenes, derechos de minorías y situaciones sucedidas a activistas en la región. No obstante, la preeminencia del tema de Cataluña resulta notable, habida cuenta de que la aspiración a la independencia de una nación no es, desde el punto de vista filosófico, un problema relacionado con la defensa de los derechos humanos.

Por esta razón, es importante observar que las aplicaciones de modelado de tópicos rinden un segundo resultado analíticamente distinto al que se acaba de mencionar. Se trata de la reconstrucción de la semántica de un término, en este caso, "derechos". La presente investigación ha arrojado un contexto semántico enormemente rico para la comprensión de la palabra "derechos" en el ámbito Iberoamericano.

En efecto, alrededor de la palabra "derechos" se tejen sentidos de una gran diversidad. Se asocia, como en su significado convencional, a los derechos humanos individuales. Pero también al derecho de autodeterminación de los ciudadanos de una nación (Cataluña y España), a los derechos económicos y sociales (una extensión de los derechos que se desarrolla en los años 60 del siglo XX como una prologanción de la Declaración de los Derechos Universales de 1948, y que se vincula con la filosofía política de la izquierda latinoamericana). Pudimos observar que este último sentido es muy importante para los tuiteros del ámbito argentino. La palabra "derechos" está muy vinculada también al derecho a la vida y a la seguridad en nuestra región. La semántica de la palabra se encuentra también asociada a la idea de "gozar de libertades", en particular en relación con actores como las mujeres, la comunidad LGBT (comunidad lesbiana, gay, bisexual y transexual) y los defensores de los derechos de los animales, un sentido muy reciente de la comprensión del derecho que se ha incorporado cómodamente al ámbito comunicativo de la región. La idea del derecho a la libertad también es prominente en los tweets de usuarios vinculados con Venezuela. También resulta muy importante su vinculación con el "derecho a la vida", un sentido sofisticado desde el punto de vista filosófico, que fue central para los autores de la Carta de los Derechos de la ONU en 1948 y cuya presencia en los tópicos obtenidos evidencia cómo los conceptos e ideario filosófico de la mencionada Carta se han permeado al mundo comunicativo cotidiano o a la cultura política de la nación.

También en un sentido general pudo observarse la presencia activa de usuarios de Twitter preocupados por el tema de los "derechos" en distintos países de América Latina, llamando la atención la relativa ausencia de tweets significativos de países como Paraguay, Uruguay y Ecuador. De Centro América el país más activo es México.

6.1. Interpretación del primer conjuto de datos de texto (muestra A):

La primera muestra fue tomada la mañana de del 24 de septiembre de 2017, una semana antes del pautado Referendo Independentista de Cataluña. Pudo

observarse en los conceptos arrojados por el LSA (Tabla 18) que uno de los usuarios más activos durante el período de tiempo en el que se recuperaron los tweets (de 8:27 am a 8:57 am, hora local), fue Albert Rivera, el líder del partido de centro derecha español Ciudadanos, quien repudió en distintos tweets, que eran retuiteados repetidamente, un ataque a una sede de diputados anti-independentistas. El repudio a ese ataque, pues, queda asociado en el conjunto de datos a la palabra "derechos", cuya semántica se redefine en ese momento como defensa de libertades políticas.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	albert_rivera, kikomonasterio, Avila
Instituciones	amnistía, defensor, guardia, policía, presidente, procurador, sindicato
Verbos	acosa, atacado, apoyo
Sustantivos	acuerdo, amenaza, antiespecie, defensa, democracia, diputados, golpistas, humanos, libertades, sedes, veganismo
Adjetivos/Adverbios	no, todo, pero

Tabla 18: Semántica LSA de la muestra A.

Similarmente, se obtuvieron los siguientes resultados para el algoritmo de PLSA:

Aspectos de la salida del algoritmo del PLSA. Concepto “derechos”	
Países/Nombres	albert rivera, Barcelona, Bogotá, Cataluña, El Salvador, España, Guatemala, Miranda, Pablo Iglesias, Soria, Venezuela
Instituciones	amnistía, defensor, guardia, policía, presidente, procurador, sindicato
Verbos	abstenerse, acosar, apoyar, atacar. callar, ceder, deben, defender, destruir, disfrutar, excederse, expresar, expresar, garantizar, mirando, permitir, queremos, reclamar, reprimir, respetar, reventar, robar, roben, rodeaban, romper, sentir, ver, violar, votar, vulnerar (total: 30)
Sustantivos	abstención, abusadores, acuerdo, amenaza, asamblea, cambio, chicas, ciudadano, claro, compañeros, constituyente, cristiano, defensa, democracia, derechos, diputados, fiscal, foto, fundamentales, futuro, gente, golpistas, guerra, humanos, impuestos, independencia, joven, laborales, ley, libertades, lucha, manifestación, mujer, mujeres, noche, oposición, pacto, patriotas, personas, políticos, promulgar, pueblos, régimen, renuncia, sedes, sentido común, soviets, violencia, voto, vulnerar (total: 51)
Adjetivos/Adverbios	demoledor, entendido, estable, hoy, importar, necesario, nuevo, orgulloso, plenamente, popular, prioridad, todo

Tabla 19: Semántica PLSA de la muestra A.

La Tabla 19, que expresa la salida del algoritmo de análisis semántico de índole probabilística, PLSA, arroja resultados mucho más diversificados y ricos que los obtenidos durante la aplicación del método de factorización de matrices LSA. Pudieron identificarse 7 menciones a países diferentes, dos españoles (España y Cataluña) y el resto latinoamericano, entre ellos Venezuela.

También se mencionaron 7 instituciones públicas relevantes en la semántica del concepto de “derechos” y 30 verbos y 51 sustantivos asociados a dicha semántica.

Mucho más elocuentes son los resultados de la Tabla 20, tabla que recoge los resultados del algoritmo LDA, entre los cuales podemos reconocer 35 verbos y 68 sustantivos asociados a la semántica del concepto de “derechos” en el mismo conjunto de datos. Como observaremos también en el segundo conjunto de datos, los resultados del algoritmo LDA son mucho más precisos en el sentido de que son

más diversos y "elocuentes" en la recuperación de la semántica de la palabra que estamos examinando.

Para el filósofo que trabaja sobre la semántica del concepto de "derechos", como en el caso de quien esto escribe, los resultados no revelan mayores sorpresas, aunque sí llama la atención que la comprensión del concepto en el ámbito del lenguaje ordinario, en el "mundo de la vida" (para utilizar la expresión clásica de la fenomenología que se refiere a las comunicaciones cotidianas), incorpora rápidamente los desarrollos académicos más sofisticados.

Por ejemplo, la mención a los "derechos de los animales", un desarrollo teórico de la ética aplicada relativamente reciente en el mundo hispanohablante (véase, por ejemplo, Mosterín, 1998), se ha incorporado rápidamente al lenguaje del "planeta Twitter". Aquí es importante tener presente que el usuario de esa red social no necesita más que un texto breve de pocos caracteres para explicar el sentido de la expresión "derecho de los animales", con lo que está seguro de que su frase no despertará extrañeza o perplejidad: un claro indicativo que el sentido de la expresión "derechos de los animales" ya ha "bajado" al lenguaje ordinario en el mundo de la vida.

También es interesante para el teórico de la ética observar que la semántica asociada a la Declaración de los Derechos Económicos y Sociales de 1966 (bastante posterior a la Declaración de la ONU sobre Derechos Humanos de 1948), también se ha incorporado a la semántica de los "derechos", con términos como "laborales", "industria", "impuesto", "trabajo", todos alejados de la semántica tradicional en torno a los Derechos Humanos Universales.

Como dato curioso, la Declaración de los Derechos Humanos del Islam, promulgada en el Cairo en 1990 en un intento por ofrecer una contrapartida no liberal a la filosofía de los DDHH occidentales, se encuentra totalmente ausente de la semántica de los conjuntos de datos, limitándose a la mención de "musulmanes" e "inmigrantes" toda referencia a los contenidos asociados al Islam. Esto pudiera ser un indicativo de un relativo fracaso de los promotores de esa declaración en publicitar sus contenidos más allá de los países con mayoría musulmana, así como la pervivencia de la semántica que ha definido los DDHH en el Occidente moderno a lo largo de los últimos siglos.

Aspectos de la salida del algoritmo del LDA. Concepto “derechos”	
Países/Nombres	Albert Rivera, Argentina, Barcelona, Cataluña, Colombia, El Salvador, España, kirschnerismo, México, Rajoy, sur Bogotá, Venezuela, Zaragoza
Instituciones	autoridades, defensor, estado, estatales, gobierno, govern, instituciones, ministro, Onu, podemos, policia, procurador
Verbos	acostumbra, adoctrinar, advertir, agitar, apoyar, asfixiar, brindar, callar, cambiar, castigar, conocer, convocar, defender, enviar, exceder, expresión, firman, garantizar, imponer, incluir, iniciar, invertir, manifestar manipular, perder, pisotear, proteger, provocar, reclamar recortar, reventar, roben, violar, votar, vulnerar (total:35)
Sustantivos	abstencionista, abusadores, activistas, acuerdo, amenaza, animales, bandera, ciudadano, ciudades, civiles, comentario, constitución, constitucional, consulta, contenidos, convivencia, cristiano, culpa, democracia, derecha, diputados, documentación, ejercicio, experiencia, femenino, golpe, humanos, idea, implicancia, impuestos, independencia, industria, joven, justicia, laboral, laborales legalidad, libertad/es, lucha, lugar, marcha, matriz, mensaje, milicos, muchachos, mujeres, obligaciones, odio, oposición, patriotas, persona, planeta, privilegios,querella, radicales, reforma, régimen, rehenes, represión, retornados, sociales, soviets, trabajo, vicios, violencia, violencia, voto, xenofobia (total: 68)
Adjetivos/Adverbios	cada momento, civismo, criminal, demoledor, destruido, facista, hoy, imposible, mejor, mientras, millones, mismos, nazis, nueva, nunca, orgullosa, público, reaccionaria, realmente, siempre, también, tampoco.

Tabla 20: Semántica LDA de la muestra A.

Para el analista filosófico también pudiera ser interesante identificar y etiquetar constelaciones de tópicos. Como ejemplo de este posible etiquetado, tomaremos, en el primer conjunto de datos de texto, la salida del algoritmo con los resultados más “elocuentes”, el LDA.

A la par de los contenidos convencionalmente asociados al concepto de derecho, tales como libertad, democracia, Constitución, etc., encontramos también las siguientes posibles constelaciones de tópicos. No se presupone un orden de importancia:

1. Independencia de Cataluña. Tópicos: 2, 6, 13, 14, 18, 19, 20, 24, 26, 28, 29.

2. Derechos laborales: 1, 27 (¿Venezuela?)

3. Régimen criminal: 3 (¿Venezuela?)

4. Derechos de las mujeres: 4, 23, 25 (¿Argentina?)

5. Represión, lucha ciudadana: 7 (¿Bogotá?)

6. Derechos de los animales: 10

7. Constitución, congreso mexicano: 12

6.2. Interpretación del segundo conjunto de datos de texto (muestra B):

En la Tabla 21 se puede observar los resultados del LSA para el segundo conjunto de datos de texto, que constaba, como se recordará, de 3415 documentos.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	Cataluña, España
Instituciones	Onu
Verbos	debe, violando, violar
Sustantivos	constitución, expertos, fundamentales, gobierno, humanos individuales, medidas, referendum, violaciones
Adjetivos/Adverbios	----

Tabla 21: Semántica LSA de la muestra B.

El algoritmo PLSA arroja 18 verbos y 48 sustantivos alrededor de la semántica del término "derechos".

No obstante, aquí las ventajas y el rendimiento en elocuencia del algoritmo LDA son palpables. Se puede observar que la atribución latente de Dirichlet (Tabla 23) ofrece una visión mucho más amplia y precisa de la semántica implítica en el segundo conjunto de datos de texto, arrojando un léxico de casi el doble de palabras asociadas al concepto inicial "derechos": 84 vs. 48. También es capaz de ofrecer un contexto más rico para este léxico, al invocar un número mayor de instituciones, adjetivos y adverbios asociadas a las palabras.

Aspectos de la salida del algoritmo del PLSA. Concepto “derechos”	
Países/Nombres	Argentina, Arreaza, Barcelona, Bolivia, Carlos Slim, Cataluña Colombia, Cristina Kirschner, Cuba, Dios, Ecuador, España Estados Unidos, México, Mundo, Rajoy, Venezuela
Instituciones	gobierno, musulmanes, ONU, Televisa, guardia civil
Verbos	abusar, advertir, anteponer, avisar, cambiar, colisionar, defender, denunciar, exigir, garantizar, limitar, merecer parecer, proteger, ratificar, tener, transmitir, violar (total: 18)
Sustantivos	adulto mayor, asamblea, aviso, bandera, cadenas pres., ciudadanos, concepto, constitución, cristiano, decadencia democracia, derecho, derechos animales, desaparición, dreamers, estrategia, fundamentales, grieta, honestidad, humanos, individuales, inmigrantes, izquierda, joven, ley, libertades, marcha animales, mujeres, obligaciones, pago deuda, partidos, personas, prioridad, pueblo, reeleccion indefinida, referendum, regresión, relatores, represión, sexuales, silencio, sociedad, urnas, utilidades, verdad, vida digna, violaciones, vulneración (total: 48)
Adjetivos/Adverbios	ahora, impecable, sagrada, siempre

Tabla 22: Semántica PLSA de la muestra B.

Aspectos de la salida del algoritmo del LDA. Concepto "derechos"	
Países/Nombres	Argentina, Arreaza, Carlos Slim, Cataluña, Colombia, Cuba, EEUU, España, MBachelet, Mundo, Venezuela
Instituciones	Ceofanb, colegio, consejo, gobierno, guardia civil, Hispania ONU, organismos, Partido Popular, Televisa, UniSalamanca
Verbos	anteponer, atentar, avanzar, avisar, colisiona, defender, denunciar, exigir, ganar, ilegalizar, luchar, merecer, piropear, proteger, rechazar, recortar, tomar, usar, violar votar (total: 20)
Sustantivos	adulto mayor, amenazas, animales, autor, bandera, ciudadanos, civil, colectivo, concepto, constitución, contrato, corrupción, cristiano, cuerpos, decadencia democracia, denuncia, derecha, derechos civiles, deshonra, discriminación, diversidad, estudiantes, fallecimiento, fiscal, frente, fundamentales, garantes, gente, grieta, hijos, humanos, igualdad, individuales, individuales, instituciones, izquierda, joven, ley, libertad, luchadora, maniobra, marcha, medidas, memoria, miedo, mujer, nacional, obligaciones, online , pago deuda, partido, patria, personas, piquete, presidencial, profesionales, profesora, propiedad, propuesta, protección, pueblo, racismo, referendum, registro, relatores, represores, respeto, salida, salud, sanciones, selección, servicios, sexual, silencio, solidaridad, trabajo, urnas, verdadera, víctima, victimario, vida digna, violaciones, vulneración (total: 84)
Adjetivos/Adverbios	abierto, absurdo, cientos, contra, frontal, justo, laborales, obligados, prioridad, siempre, total, urgente

Tabla 23: Semántica LDA de la muestra B.

Si, de nuevo, identificamos constelaciones de palabras para discernir tópicos más amplios o generales, encontraremos los siguientes, tomando en cuenta la salida del algoritmo más elocuente LDA.

Tomando en cuenta la salida de este algoritmo, para el segundo conjunto de tweets es posible identificar los siguientes tópicos asociados al concepto de "derechos", al lado de los más convencionales tales como libertad, ciudadanía, estudiantes, democracia, etc., sin presuponer un orden de importancia:

1. Cadenas presidenciales: tópico 1 (¿Venezuela?)

2. Derechos laborales: 2, 7, 9 (¿Argentina?)

3. Diversidad sexual: 3, 8, 14 (¿México?)

4. Derechos de la mujer: 4, 12 (¿España?)

5. Derechos humanos violados por sanciones de EEUU: 5, 25 (Venezuela).

6. Derecho a la independencia de Cataluña. De lejos, la semántica más prevalente para este conjunto de datos: 6, 10, 15, 16, 17, 18, 19, 20, 24, 26.

7. Derechos animales: 22 (¿España?)

8. Vida digna, adulto mayor: 27 (Ecuador)

6.3. Interpretación del tercer conjuto de datos de texto (muestra C):

En la Tabla 24 se observan los resultados para el tercer conjunto de datos de texto, que constaba de 2020 documentos y 20881 atributos, de la aplicación del algoritmo para el LSA.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	Cataluña, España, Estrasburgo, Murcia
Instituciones	Caritas, Pontifex
Verbos	Luchan
Sustantivos	bandera, constitucional, derechos, gente, humanos, palos, patria, represión
Adjetivos/Adverbios	"Pasando de algo"

Tabla 24: Semántica LSA de la muestra C.

Por su parte, el algoritmo PLSA arroja 44 verbos y 53 sustantivos alrededor de la semántica del término "derechos".

Sin embargo, como se ha observado en los otros conjuntos de datos de texto, aquí las ventajas y el rendimiento en elocuencia del algoritmo LDA son también importantes. Se puede constatar que la atribución latente de Dirichlet (Tabla 26) ofrece un mejor escrutinio de la semántica implítica en el tercer conjunto de datos de texto, arrojando un léxico de más del doble de palabras asociadas al concepto inicial "derechos": 129 vs. 53. También ofrece un contexto más rico para este léxico por su número mayor de instituciones, adjetivos y adverbios asociadas a las palabras.

Aspectos de la salida del algoritmo del PLSA. Concepto “derechos”	
Países/Nombres	Argentinos, Barcelona, Catalanes, Cataluña, Colombia, España, Estrasburgo, Lenin, LilianTintori, Luis E. Rondón, Macri, Mexicanos, Murcia, Pablo Iglesias, Santander,Tumaco Unes_Táchira
Instituciones	congreso, gobierno, guardia civil, ONU, TSJ
Verbos	amo, arrastrar, asegurar, coaccionar, defender, dejar denunciar, desaparecer, desnucar, devolver, egresar, evitar, fundar, garantiza, habéis, hablar, humillar impedir, independizarnos, jugando, legislar, luchan, mentir, ocuparse, ocurrir, paguen, proteger, querer, quitar, reclamar, reconocer, recordar, recortar, renuncie, restringir, seguir, somos, tenes, vamos, visto, vivir, votar, vulnerar (total: 44)
Sustantivos	animales, asociación, autoritario, bandera, cabeza, comida, consecuentes, constitución, contexto, crisis, democracia, desprotección, dignidad, esperanza, estado, expediente, fascistas, feminismo, futbolistas, futuro, gente, humanitaria, humanos, integridad, justicia, laborales, libertad, libertades, marcha, masacre, mayoría, miedo, muerte, mujeres, mundo, nacionalidad, naturaleza, octubre, palos, personas, profesionales, pueblo, rector tsj, represión respeto, sentencia, sociales, techo, tierra, trabajo, unidad, vida, violaciones (total: 53)
Adjetivos/Adverbios	fracasado, fundamentales, mayores, mejor, nuevos, pocos, siempre, también, todos, vasta

Tabla 25: Semántica PLSA de la muestra C.

Aspectos de la salida del algoritmo del LDA. Concepto "derechos"	
Países/Nombres	Albert Rivera, Argentina, Argentinos, Bachelet, Barcelona, Botero, Catalana, Colombia, Estrasburgo, Lenin, Lilian Tintori, Luis E. Rondón, Macri, Mexicanos, México Murcia, Santos, ThaysPeñalver, Trump, Tumaco Unes_Táchira
Instituciones	CNE, congreso, gobierno, guardia civil, MUD, Netflix ONU, TSJ
Verbos	aprovechar, arrastrada, arrebatar, asegurar, cambiar, coaccionar, comiendo, comprendamos, conocer, cuentan, debemos, defendiendo, dejar, denunciando, desaparecer, desnuquen, devolverle, discuten, egresan, entrenando, esperar, exigen, fortalecen, fundan, garantizar,gobierna, habeis, hablar, hacer, humillar, imagino, impedir, independizarnos, invitar, joder, legislar, lograste, luchando, mentir, ocurre, paguen, perdés, pisoteando, podremos, proteger, puede, queremos, quiero, quitar, realizamos, reclamar, reconoce, recortar, renuncie, respetar, restituir, restringir, seguimos, sentimos, sigamos, somos, tenemos, tenes, vamos, vamos, visto, viviendo, vivir, volar, vota, voy (total: 71)
Sustantivos	aborto, activista, adultos, agenda, animales, arte, asociaciones, campesino, cancelaciones, capitalismo, carro, caza, ciuadadanos, clases, colaborador, comida, comunidad, consecuentes, constitución, construcción, corrupción, crisis, deber, declaración, defensa, defensor, defensores, democracia, desprotección, dictadura, dictaduras, dignidad, docentes, domingo, ejemplo encuentro, escuelas, esperanza, expediente, fascistas, feminismo, fiesta, foro, funcionarios, fútbol, futbolista, futbolistas, futuro, gente, golpe, graduandos, guerra, hijos, huelga, humanos, igualdad, imagen, imperialistas, individuo, injusticia, instituciones, integridad, invencible, jubilaciones, jugadores, laborales, leyes, lgbt, libertad, libertades, lucha, mano, marcha, masacre, mayoría, miedo, muerte, mujer, mujeres, mundo, nacionalidad, naturaleza, nazis, noticia, octubre, ofensiva, oficial, organización, país, partido, paz, periodista, personas, plan, plumazo, poder, policía, presencia, primera, privilegios, profesionales público, pueblo, religión, reputación, respeto, retroceso, rey, ricos, robo, rumbo, sacrificio, salud, sentencia, sicario, sociales, sociedad, soldado, techo, tirano, trabajadores, trabajo, trato, unidad, vía, vicios, violación, violaciones, vulneación (total: 129)

Adjetivos/Adverbios	ahora, animalista, autoritario, civil, claro, conmovedora, constitucional, desarmado, fracasada, frecuentemente, gran, hoy, humanitaria, iguales, imposible, injustas, inspiracional, libre, mala, mayores, mejor, muchos, mundial, nada, nuevos, peor, plenamente, pocos, rápido, siempre, social, solo, todavía, todos

Tabla 26: Semántica LDA de la muestra C.

Aquí es posible también identificar, desde el punto de vista del análisis filosófico, palabras alrededor de tópicos posibles. Tomando en cuenta los resultados más elocuentes que se han obtenido, aquellos arrojados por el algoritmo LDA, se puede ahora señalar algunos de ellos:

1. Defensa de derechos laborales, marcha (¿Colombia y México?), comida, techo, ricos, obra de arte inspiradora, Argentina, vivir, aborto, paz, esperanza, derechos sociales, salvaje, joder, pisotear, capitalismo, iguales, asustar, retroceso, docentes, cancelar jubilaciones, Navarro, ¿Venezuela?, lucha, bandera, huelga, guerra,0, 3, 5, 8, 10, 16, 23.
2. Restitución de derechos, TSJ, Venezuela, ley, golpe, México: fortalecer justicia en instituciones, derechos campesinos (¿Venezuela y Argentina?), Chile: derechos adultos mayores, democracia, sociedad, cambiar el rumbo, mexicanos, libertades, miedo, desaparecer, futbolista, arrebatarnos, quitarnos, humanos, masacre, expediente, respeto, violaciones, derecho, desprotección, luchar, humanos, libertad, hijos, privilegios, agenda, leyes 1, 4, 5, 13, 19, 25, 27.
3. Derechos de las mujeres, feminismo, Cataluña,mujer, libertades, defender, evitar, desnucar, arrastrar, 2, 22.
4. Argentina: contra imperialistas, nazis, fascitas, autoritario, Macri, reconocimientos, mayoría, gobierno, libertades, libertad, dictadura, crisis humanitaria, defender, libertad, Venezuela 5, 11, 21, 26.
5. Cataluña: amenaza a la nacionalidad, constitución, trabajo, coaccionado, humillado, mentir, Murcia, luchando, gente, policía, España, represión, renunciar, poder, imagen, Pablo Iglesias, Barcelona, defiene, constitución 7, 17, 18, 27, 28.
6. Derechos de minorías: Fiesta animalista, invita, Venezuela, Aragua, Ciudadanos lgbt, defensores, salud, integridad, libertades,14, 20.

Capítulo 7. Conclusiones

El objetivo de nuestro trabajo ha sido el estudio de tres técnicas de modelado de tópicos o reducción de la dimensionalidad en conjuntos de datos de textos que, desde el punto de vista filosófico y teórico, suponen la recuperación de la semántica de matrices dispersas de bolsas de palabras. Estas tres técnicas se desarrollan en el marco de dos paradigmas generales que son muy importantes en la literatura del análisis estadístico de datos: el paradigma frecuentista y el paradigma Bayesiano que lidia con la incertidumbre sobre cantidades desconocidas.

Se expuso que, de las tres técnicas de modelado de tópicos que se estudiaron, las dos primeras, el LSA y el PLSA, ofrecen un análisis estadístico del conjunto de datos que se apoya en un conteo preciso de las frecuencias de términos en documentos al interior de conjuntos de datos acotados.

En este sentido, la primera de las técnicas de modelado de tópicos de tipo frecuentista, el LSA, rinde una factorización de la matriz que la descompone en valores singulares, los cuales se convierten en los parámetros para la agrupación de nuevos documentos y términos en los tópicos que ha identificado el sistema. La segunda técnica, el PLSA, por su parte, rinde otro tipo de parámetros: aquellos que definen una distribución de probabilidad para tópicos de los términos y documentos que se incorporan en el sistema. De este modo, ambas técnicas son técnicas de estadística descriptiva que buscan generalizar a conjuntos de datos no observados la matriz que se ha analizado en la búsqueda de las frecuencias de términos en los documentos que conforman el corpus.

Finalmente, la tercera de las técnicas, la LDA, se desprende del paradigma frecuentista anteriormente descrito en la medida en que es considerada por la literatura una técnica de inferencia estadística sobre cantidades desconocidas. La LDA debe apelar al cálculo de la distribución de probabilidad posterior en conjuntos de datos cuyos parámetros van “desapareciendo” por pérdida de la posibilidad de un dominio de la distribución de la probabilidad marginal, a causa del crecimiento exponencial de los datos.

Por esta razón, la LDA se aleja del paradigma frecuentista y representa un intento de calcular probabilidades en condiciones de *incertidumbre*. Este tipo de condiciones, tal y como reconoce de modo creciente la literatura especializada, será cada vez más prevalente a medida que crecen los conjuntos de datos de texto presentes en la Web. De allí su importancia y la necesidad de comprenderla de modo adecuado. A medida que se incorporan nuevos datos a los sistemas de modelado de tópicos, como es natural, va desapareciendo la precisión de los parámetros que rigen la distribución de probabilidades de los tópicos. Esto hace que sea cada vez más difícil generalizar modelos cuyos datos de entrada crecen exponencialmente.

Las distribuciones de Dirichlet resuelven este problema al permitir un cálculo de probabilidades sobre distribuciones de probabilidad a través de una técnica matemática que permite la conjugación de la previa en distribuciones multinomiales, de modo que se conserve a lo largo de todas las iteraciones la familia exponencial a la que pertenece la distribución. Paralelamente, los modelos de LDA proponen distintos métodos para facilitar la inferencia variacional de los datos de entrada y, con ello, la creación de modelos con una mayor capacidad de generalización. A este fin, hemos observado que la LDA hace un uso completo de la inferencia Bayesiana, un tipo de cálculo de probabilidades que se aleja de la estadística descriptiva por su uso del cálculo de la posterior en la regla de Bayes, con lo que entra en un área de análisis más especulativo que no se encuentra exento de polémica por el carácter más subjetivo que comportan los cálculos Bayesianos.

En nuestro trabajo, hemos estudiado con detalle este y otros aspectos que definen, y delimitan teóricamente, las tres técnicas que hemos mencionado. Se analizaron los aspectos más resaltantes de sus algoritmos y se aplicaron a tres conjuntos de datos de texto, que se modelaron en matrices de términos-documentos, tomados de la red social Twitter.

Al capturar los tres conjuntos de datos de texto, la palabra clave ha sido la palabra “derechos”, un ámbito de análisis semántico y filosófico del cual, quien esto escribe, se ha ocupado a lo largo de su carrera académica. Los documentos de Twitter o tweets se eligieron, entonces, conforme a si mencionaban en el texto esa palabra. La intuición que guió el calibrado de los distintos programas de código que se usaron para la aplicación de Twitter ha sido que los tweets que mencionaban la palabra “derechos”, al ser reducida su dimensionalidad con las tres técnicas que se estudiaron, arrojarían un contexto conceptual o una serie de tópicos característicos que podían iluminar la semántica implícita o latente en el uso de la palabra “derechos” por parte de los usuarios de los tweets capturados en “streaming”, todos provenientes del ámbito Iberoamericano.

Los resultados analizados en los capítulos 5 y 6 de este trabajo demuestran que esa intuición, apoyada por la literatura sobre el tema, es correcta y que los documentos estudiados realmente ofrecen interesantes hallazgos respecto al significado del concepto de “derechos” para los usuarios de la red social Twitter en el momento de la captura. En efecto:

1. Un primer hallazgo ha sido la coincidencia del significado semántico de la palabra “derechos” con los resultados ofrecidos en los distintos estudios de este tipo de conceptos que se encuentran en el ámbito de la reflexión filosófica de carácter analítico, el tipo de análisis que caracteriza el desempeño profesional de los filósofos y que persigue una fundamentación racional del concepto. Esta coincidencia sugiere que muchos de los aspectos teóricos más importantes que se

encuentran en los esfuerzos de fundamentación de la noción de "derechos" han sido exitosamente incorporados a la cultura política de la región. En este sentido, la tarea educativa de las Naciones Unidas pudiera calificarse, en un primer nivel, como una tarea lograda: la aguda conciencia, por parte de aquellos que reclaman sus derechos en nuestros países de habla hispana, respecto de lo que involucra realmente apelar a los derechos, constituye, en lo personal, uno de los hallazgos más importantes de nuestro trabajo.

El discurso sobre el derecho entendido como derecho a la vida, la libertad, la seguridad, la libertad de expresión, pero también el derecho a una vida digna, seguridad laboral, condiciones de vida razonables (vivienda y bienestar), derecho de los animales, derechos de las mujeres, pero también de minorías, etc., se ofrece de modo elocuente a partir de los modelos estudiados y abren al profesional de la filosofía y de las ciencias sociales, que quiere fundamentar su análisis con evidencia empírica, un campo de trabajo amplio e importante.

Las herramientas que hemos estudiado, pues, iluminan la semántica que define el uso de muchos conceptos filosóficos, y de otras disciplinas cuyos conceptos se expresan en el lenguaje ordinario, y lo hacen de maneras inéditas hasta hace relativamente poco tiempo.

2. El segundo hallazgo ha sido que, al contrario de lo que podría pensarse, el significado semántico de la noción de "derechos", hasta donde quien esto escribe alcanza a ver, tal y como es usado en contextos de lenguaje ordinario al interior de estos tres conjuntos de documentos de Twitter, no revela ninguna novedad respecto de lo que ordinariamente se entiende por "derechos" en el discurso filosófico especializado. Probablemente una exploración más profunda y con conjuntos mayores pudiera revelar novedades en la semántica del concepto en el futuro. Pero no ha sido así para este estudio. De este modo, aunque la semántica reconstruida por los tres tipos de algoritmos ha sido rica y diversificada, no es particularmente novedosa y se mantiene en un nivel elevado de convencionalidad.[17] Por esta razón, podemos decir, como ya lo hicimos en el trabajo, que el análisis de contenidos semánticos latentes que ofrece el modelado de tópicos automatizado es una suerte de pequeña encuesta de opinión sobre el estado actual de la opinión pública respecto de un concepto.

3. En tercer lugar, pudimos constatar que, tal y como ya lo adelantamos en nuestro análisis de la literatura sobre las diferencias entre los modelos de PLSA y LDA, que, de acuerdo con Crain et al. (2012), la LDA tiende a aprender tópicos más generales o demasiado amplios, lo que los hace más difusos. Como ya hemos adelantado unas páginas atrás, ello tiene las siguientes consecuencias: que los modelos de LDA, según pudimos apreciar también nosotros, son muy buenos para

[17] Usamos aquí el concepto de lenguaje moral convencional en el mismo sentido que Lawrence Kohlberg (1981).

analizar transversalmente la semántica de un concepto (lo que era la intención inicial del presente trabajo), pero menos buenos si se trata de identificar cuáles son los tópicos que realmente caracterizan a un conjunto de datos. De este modo, se pudo notar que resulta más fácil identificar esos distintos tópicos en los modelos de PLSA que en los modelos de LDA, mientras que, inversamente, es más nítida la semántica de un término en los modelos de LDA que en los modelos de PLSA.

De este modo, es posible que un periodista, un científico social o un investigador en redes sociales que quiera averiguar de qué se está hablando, de qué temas o tópicos se ocupa la gente en un momento dado, obtenga mejores resultados con un modelo de PLSA. Pero, por contraste, un filósofo que quiera averiguar qué entiende la gente por un determinado concepto que se usa de manera transversal en tópicos a lo largo de distintos documentos, tal vez obtenga mejores resultados o resultados más precisos con un modelo de LDA. Por esta razón, nosotros hemos usado los resultados arrojados por el algoritmo de LDA para nuestro análisis hermenéutico de la semántica encontrada.

4. Un cuarto hallazgo de nuestro trabajo es que, si bien los modelos de LSA arrojan resultados relativamente confiables, los modelos frecuentistas y de inferencia Bayesiana que apelan al cálculo de probabilidades representan una ventaja enorme respecto de los primeros, que apelan al cálculo de valores singulares en matrices factorizadas. Tal vez por esta razón, los modelos probabilísticos son cada vez más populares en la literatura de las ciencias de la computación, quedando como aspecto polémico en la discusión sobre la futura prevalencia de estos modelos el peso de la subjetividad del investigador que distribuye la probabilidad previa, un tipo de subjetividad típico de la inferencia Bayesiana.

Queda abierta la interrogante respecto de si esa subjetividad es un aspecto inherente a la filosofía que subyace a los modelos Bayesianos y, en este sentido, tal vez los filósofos puedan realmente iluminar, en el futuro, esa discusión con su aporte particular.

Referencias y bibliografía

Abu-Mostafa, Yaser, 1987, "The Vapnik-Chervonenkis Dimension: Information versus Complexity in Learning", en *Neural Computation* 1, 312-317, MIT. Disponible en http://www.work.caltech.edu/pubs.html.

Abu-Mostafa, Yaser, Magdon-Ismail, Malik y Hsuan-tien Lin, 2012, *Learning from Data*, AMLbook.com.

Abu-Mostafa, Yaser, 2012, "Training versus Testing", Videoconferencia disponible en https://www.youtube.com/watch?v=SEYAnnLazMU, Caltech.

Ahmed, Amr et al., 2012, "Scalable Inference in Latent Variable Models", en Actas del *WSDM*, Seattle, WH. Disponible en: http://www.cs.cmu.edu/~jegonzal/papers/ahmed_scalable_inference_in_latent_variable_models.pdf.

Aggarwal, Charu (ed.), 2011, *Social Network Data Analytics*, DOI 10.1007/978-1-4419-8462-3_1, Springer Science+ Business Media.

Aggarwal, Charu y ChengXiang Zhai (Eds.), 2012, *Mining Text Data*, Springer.

Aggarwal, Charu, 2015, *Data Mining. The Textbook*, Springer.

Arora, S., et al., 2013, "A Practical Algorithm for Topic Modeling with Provable Guarantees", en *Proceedings of the 30th International Conference on Machine Learning,* Atlanta, disponible en: http://cs.nyu.edu/~dsontag/papers/AroraEtAl_icml13.pdf.

Asuncion, A., Welling, M., Smyth, P., Teh, Y., 2009, "On smoothing and inference for topic models". En *Uncertainty in Artificial Intelligence*.

Ay Yeung, Albert, 2010, "Matrix Factorization: A Simple Tutorial and Implementation in Python" en http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/.

Bernardo, José, 2003, *Bayesian Statistics*, en Viertl, R. (ed), *Probability and Statistics*en la *Encyclopedia of Life Support Systems*, Oxford, Unesco.

Bird, S., Klein, E y E. Loper, 2009, *Natural Language Processing with Python*, O'Reilly.

Blei, D., Ng, A., Jordan, M., 2003, "Latent Dirichlet allocation." *J. Mach. Learn. Res. 3.*

Blei, D y M. Jordan, 2006 a, "Variational Inference for Dirichlet Process Mixtures" en *Bayesian Analysis*, 1(1).

Blei, D., Lafferty, J. , 2006 b, "Dynamic topic models". En *International Conference on Machine Learning*, ACM , New York, NY.

Blei, D., Lafferty, J. , 2007, "A Correlated Topic Model of *Science*" en *The Annals of Applied Statistics*, Vol.1, No. 1, 17-35, Doi: 10.1214/07 AOAS 114.

Blei, D., 2011, "Variational Inference", disponible en: https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf.

Blei, David, 2012, "Probabilistic Topic Models" en *Communications of the ACM*, Abril, Vol.55, No. 4.

Blei, David, Kucukelbir, Alp y Jon McAuliffe, 2016, "Variational Inference: A Review for Statisticians"en arXiv:1601.00670v4 [stat.CO] 2 Nov.

Blei, David, 2017 a, "Probabilistic Topic Models and User Behavior". En el canal de *The School of Informatics at the University of Edinburgh*. Subido el 2 de febrero del 2017.

Blei, David, 2017 b, "Variational Inference": Foundations and Innovations", disponible en https://www.youtube.com/watch?v=Dv86zdWjJKQ&list=PLjc-tqOYn1C73S5-g1fE4EksnMrzsaIYb&index=9.

Chang, J. y Boyd-Graber, J., "Reading Tea Leaves: How Humans Interprets Topic Models", 2009, en *Neural Information Processing Systems,* University of Princeton,https://www.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf.

Cichocki, A., Cruces, S. y S. Amari, 2011, "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization en *Entropy*, 13, 134-170, doi: 10.3390/e 13010134.

Crain, S. y Ke Zhou, 2012, "Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond", en Aggarwal, Charu y ChengXiang Zhai (Eds.), *Mining Text Data*, Springer.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., 1990, "Indexing by latent semantic analysis". *J. Am. Soc. Inform. Sci. 41*, 6.

Dempster, A.P., Laird, N. M. y D. Rubin, 1977, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1, 1-38.

Domingos, Pedro, 2015, *The Master Algorithm*, Basic Books.

Ferguson, Thomas, 1972, "A Bayesian Analysis of Some Nonparametric Problems" en *Institute of Mathematical Statistics*, disponible en www.jstor.org.

Fernández Gallardo, P., 2004, "El secreto de Google y el álgebra lineal", Departamento de Matemáticas de la Universidad Autónoma de Madrid, disponible en: https://www.uam.es/personal_pdi/ciencias/gallardo/upm_google.pdf.

Févotte, Cedric y Jérome Idier, 2010, "Algorithms for nonnegative matrix factorizacion with the beta divergence"en arXiv.org, cs, 1010.1763, Cornell University, disponible en https://arxiv.org/abs/1010.1763.

Folzt, Brandon, 2012, https://www.youtube.com/watch?v=ConmIDAzRqI.

Foltz, B., 2013 a, "Simple Linear Regression.The Very Basics". Disponible en su canal de YouTube y en el enlace https://www.youtube.com/watch?v=ZkjP5RJLQF4&list=PLIeGtxpvyG-LoKUpV0fSY8BGKIMIdmfCi.

Folzt, B., 2013 b, https://www.youtube.com/watch?v=locZabK4Als.

Frigyik, Bela, Kapila, Amol y Maya Gupta, 2010, "Introduction to the Dirichlet Distribution and Related Processes" en UWEE Technical Report, Number UWEETR, 0006, Diciembre de 2010. Disponible en: https://www2.ee.washington.edu/techsite/papers/documents/UWEETR-2010-0006.pdf.

Gershman, S. y N. Goodman, 2014, "Amortized Inference in Probabilistic Reasoning", disponible: http://gershmanlab.webfactional.com/pubs/GershmanGoodman14.pdf.

Golub, G. y Ch. Van Loan, *Matrix Computations*, 1996, The John Hopkins University Press, Baltimore y Londres, disponible en: http://web.mit.edu/ehliu/Public/sclark/Golub%20G.H.,%20Van%20Loan%20C.F.-%20Matrix%20Computations.pdf.

Griffiths, Th., y Mark Steyvers, 2004, "Finding Scientific Topics", en*PNAS*, 101(1): 5228-5235, disponible en:

http://psiexp.ss.uci.edu/research/papers/sciencetopics.pdf.

Grimmer, J., 2010, "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases." En *Polit. Anal.18*, 1.

Hoffman, M., Blei, D. y F. Bach, 2010, "Online Learning for Latent Dirichlet Allocation", disponible en https://endymecy.gitbooks.io/spark-ml-source-analysis/content/%E8%81%9A%E7%B1%BB/LDA/docs/Online%20Learning%20for%20Latent%20Dirichlet%20Allocation.pdf.

Hoffman, M. et al., 2013, "Stochastic Variational Inference", en *Journal of Machine Learning Research,* 14, 1303-1347.Disponible en: http://www.columbia.edu/~jwp2128/Papers/HoffmanBleiWangPaisley2013.pdf.

Hong, Liangjie y Brian Davidson, 2010, "Empirical Study of Topic Modeling in Twitter" en las *Actas del 1st Workshop on Social Media Analytics (SOMA'10)*, Washington DC.

Johnson, P y M. Beverlin, 2013, "Beta distribution", disponible en: http://pj.freefaculty.org/guides/stat/Distributions/DistributionWriteups/Beta/Beta.pdf.

Kalchbrenner, N., Grefenstette, E. y Ph. Blunsom, 2014, „A Convolutional Neural Network for Modeling Sentences", en arXiv: 1404.2188v1 [cs.CL].

Kinsley, Harrison: Canal de You Tube sobre uso de Python para tareas de minado de texto en Twitter. Sentdex: https://www.youtube.com/user/sentdex/featured.

Kohlberg, Lawrence, 1981, *The Philosophy of Moral Development*, Harper & Raw.

Kolman, Bernard y David R. Hill, 2006, *Álgebra lineal*, Pearson Education.

LeCun, Yann, Bengio Yoshua y Geoffrey Hinton, 2015, "Deep Learning", 436, Vol. 521, *Nature*, Macmillan Publishers.

Leskovec, Rajaraman y Ullman, 2016, *Serie de la Universidad de Stanford sobre Reducción de la Dimensionalidad*. Disponibles en los siguientes enlaces de You Tube, publicados en: https://www.youtube.com/watch?v=yLdOS6xyM_Q, https://www.youtube.com/watch?v=UyAfmAZU_WI, https://www.youtube.com/watch?v=P5mlg91as1c&t=45s, https://www.youtube.com/watch?v=K38wVcdNuFc.

Leskovec, Jure, Anand Rajaraman y Jeffrey Ullman, 2010, *Mining of Masive Data Sets*. Universidad de Stanford, disponible en:

http://infolab.stanford.edu/~ullman/mmds/book.pdf.

Luce, R. Duncan y Howard Raiffa, 1957, *Games and decisions*, John Wiley & Sons, New York.

Martin, D. y Michael Berry, 2006, "Mathematical Foundations Behind Latent Semantic Analysis", disponible en:
http://mall.psy.ohio-state.edu/LexicalSemantics/MartinBerry2006.pdf.

Merton, Robert, 1968, "The Matthew Effect in Science", *Science* 159(3810):56-63, January 5, disponible en

http://www.garfield.library.upenn.edu/merton/matthew1.pdf.

Mimno, D., McCallum, A., 2008, "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression".En *Uncertainty in Artificial Intelligence*.

Mimno, D., et al., 2012, "Sparse Stochastic Variational Inference" en *Proceedings of the 29th International Conference on Machine Learning*, Escocia.

Minka, Thomas y John Lafferty, 2002, "Expectation-Propagation for the Generative Aspect Model", disponible en https://tminka.github.io/papers/aspect/minka-aspect.pdf.

Mosterín, Jesus, 1998, ¡*Vivan los animales!*, Editorial Debate, Madrid.

Mosterín, Jesús, 2000, *Conceptos y teorías en la ciencia*, Alianza Editorial, Madrid.

Munzert, Simon, Rubba, Christian, Meissner, Peter y Dominic Nyhuis, 2015,*Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*, John Wiley & Sons, U.K., 2015.

Neal, Radford M., 1995, *Bayesian Learning for Neural Networks*, Tesis Doctoral, Toronto, disponible en http://www.csri.utoronto.ca/~radford/ftp/thesis.pdf.

Newman, D. et al., 2009, "Distributed Algorithms for Topic Models" en *Journal of Machine Learning Research*, 10, 1801-1828.

Oneata, Dan, 2016, "Probabilistic Latent Semantic Analysis", disponible en http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV1011/oneata.pdf.

Ongaro, A y S. Migliorati, "A generalization of the Dirichlet distribution" en Journal of Multivariate Analysis 114 (2013), 412-426. Disponible en: https://ac.els-cdn.com/S0047259X12001753/1-s2.0-S0047259X12001753-main.pdf?_tid=8fc6f460-0b8c-11e8-ac83-00000aacb35d&acdnat=1517956077_8352a04b6f6247a09db8e958cb0dd6b6

Puschmann, Cornelius y Tatjana Scheffler, "Topic modeling for media and communication research: a short primer", en *HIIG Discussion Paper Series*, Alexander von Humboldt Institut für Internet und Gesellschaft, 5.

Pedregosa et al., 2011,"Scikit-learn: Machine Learning in Python", en JMLR 12, pp. 2825-2830.

Pritam, Gundecha y Huan Liu, 2010, "Mining Social Media: A Brief Introduction" en *Tutorials in Operations Research*, Informs, http://dx.doi.org/10.1287/educ.1120.0105.

Raganath, R., Tang, L., Charlin, L. y D. Blei, 2015, "Deep Exponential Families" en *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, San Diego, JMLR: W&CP, Vol. 38.

Rogers, Simon y Mark Girolami, 2012, *A First Course in Machine Learning*, CRC Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smith, P., 2004, "The author-topic model for authors and documents". En *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence,* AUAI Press.

Rumerlhart, D.E., McClelland, J.L and the PDP Research Group, 1987, *Parallel Distributed Processing*, A Bradford Book.

Schafer, Corey: Canal de You Tube con tutoriales sobre Python, https://www.youtube.com/channel/UCCezIgC97PvUuR4_gbFUs5g.

Scikit-learn, documentación,http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#sklearn.decomposition.LatentDirichletAllocation.

Sklar, Max, 2014, "The Dirichlet Distribution", disponible en https://www.hakkalabs.co/articles/the-dirichlet-distribution.

Spiegelhalter, D., Best, N., Carlin, B., Van der Linde, 2002, „Bayesian Measures of Model Complexity and Fit" en J.R. Statis. Soc. B, 64, Part 4, 583-639.

Steyvers, M., Griffiths, T., 2006, "Probabilistic topic models". En T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (eds): *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum.

Strawson, P.F., 1992, *Analysis and Metaphysics*, Oxford University Press.

Sullivan, Scott, 2017, "LDA Algorithm Description", disponible en https://www.youtube.com/watch?v=DWJYZq_fQ2A.

Tabachnick, B. y L. Fidell, 2013, *Using Multivariate Statistic*, Pearson.

Tallis, Raymond, *Aping Mankind.Neuromania, Darwinitis and the Misrepresentation of Humanity*, Routledge, 2014.

Teh, Y., et al, 2006, "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation", disponible en https://papers.nips.cc/paper/3113-a-collapsed-variational-bayesian-inference-algorithm-for-latent-dirichlet-allocation.pdf.

Waal, Alta, Jacobus Venter y Etienne Barnard, 2008, "Applying Topic Modeling to Forensic Data", en *Advances in Digital Forensics IV*.

Wallach, H., 2006, "Topic modeling: Beyond bag of words" en *Proceedings of the 23rd International Conference on Machine Learning*.

Zhao, Yanchang, 2015, *R and Data Mining: Examples and Case Studies*, Elsiever.