Twitter™

Una fuente de datos para la academia

Wilmer González / wilmer.a.gonzalez@ucv.ve

Dado el creciente volumen de datos existentes en las redes sociales, el aumento en la velocidad de generación y disponibilidad de uso de los mismos en ciertas plataformas, pueden representar una oportunidad en el análisis de patrones, ya que hacen factible la recopilación de estos datos, que bajo ciertas condiciones están dispuestos públicamente y cuyo estudio puede tener fines asociados a diferentes áreas del conocimiento como: Psicología, Sociología, Economía y aún más en Computación.

Fuente de datos

Particularmente, la Red Social Twitter™ permite este tipo de tareas al concebir el uso de herramientas de captura automatizada de datos que, naturalmente, integran su ***API***, lo cual se refiere al conjunto de subrutinas, funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro *software* como una capa de abstracción [1].

El levantamiento de información, es parte esencial de cualquier estudio, como tal, debe tratar de realizarse de la forma más eficiente y efectiva posible. Extraer datos mediante el ***API*** de twitter puede ayudarnos al cumplimiento de esta tarea, así como a la automatización de dicho proceso.

Condiciones de Uso del Twitter™

Inicialmente, la recopilación de estos datos (*Tweets*) es posible gracias al **Acuerdo de Privacidad** [2] que Twitter™ establece con sus usuarios, así como con los desarrolladores que harán uso del ***API***, mediante el **Contrato y Política para desarrolladores** [3].

Para poder realizar la captura de los datos se requiere definir el tipo de ***API*** que más se ajusta a las necesidades, que en este caso pueden ser los tipos ***REST*** y/o ***Streaming***, dado que ambos permiten la captura de los objetos definidos. Sin embargo, por simplicidad el tipo ***REST*** puede resultar conveniente. Así mismo se requiere decidir el lenguaje de programación mediante el cual se accederá a dicha ***API***, que provea las funcionalidades suficientes para las tareas que se realizarán.

Caso de Estudio

Una vez aclarado los puntos anteriores, se presenta el siguiente caso de estudio ejemplo referente al área de **Minería de Texto**, entendida como el análisis y procesamiento de la información para facilitar la toma de decisiones.

Es bueno acotar que la extracción de datos mediante el uso del ***API*** de twitter es importante porque representa una fuente variada y densa de contenido, que refleja la opinión, sentimientos, e incluso puede contener información de eventos pasados, actuales o que se llevarán a cabo.

Se usará el acceso de tipo ***REST*** al ***API*** de Twitter™ mediante el lenguaje de programación ***R***, dado que este posee una gran cantidad de paquetes útiles en la realización del análisis, que consistirá en estudiar patrones asociados a los *tweets* recopilados entre los días **'2015-15-11'** y **'2015-29-11'** que contengan la palabra clave ***'#6D'*** .

Las primeras etapas de este algoritmo deben dedicarse al pre-procesamiento de los datos, en este sentido, las actividades más comunes son:

- Definición de la codificación apropiada para el texto, para garantizar la compatibilidad con la mayor cantidad de plataformas.
- Tratamiento de signos de puntuación.

- Tratamiento de ***URLs.***
- Transformación a minúsculas de caracteres alfabéticos, para disminuir la redundancia de elementos léxicos.
- Tratamiento de palabras muy frecuentes pero poco informativas.

Una vez pre-procesados los *tweets*, se puede indagar en las palabras más frecuentes en el conjunto de todos los *tweets* (ver imagen tópicos), así como un gráfico similar puede indicar la cantidad de *tweets* generados por usuario.

Otras actividades posteriores al pre-procesamiento incluyen: Análisis de tópicos latentes (***LDA***), Análisis factoriales, Determinación de reglas de asociación (***A priori***), Análisis de series de tiempo con respecto a la frecuencia de los tópicos, Análisis semántico, Análisis pragmático, desambiguación de significados, Determinación de *tweets* ***SPAM***, Seguimiento de eventos en tiempo real, Determinación de características socioeconómicas de los usuarios derivadas de sus *tweets*, entre otros [4].

Aunque sencillo, el presente ejemplo tiene como fin ilustrar las oportunidades de captar y analizar datos que se pueden encontrar en datos abiertos como es el caso de twitter. ■

El código y datos asociados a los gráficos se encuentran disponibles en:

https://github.com/wilmeragsgh/vdtic-6d

Conclusión

En este artículo se presentaron una serie de pasos y conceptos relacionados con el levantamiento de datos desde el API de Twitter, pudiendo estos ser analizados posteriormente con fines académicos para diversas áreas del conocimiento. Entendiendo que la red social en cuestión posee características que conforman un fenómeno propio que no necesariamente es igual a la realidad, a pesar de ello, análisis estructurales, de comportamiento y otras propiedades pueden derivarse de los conceptos generales de redes sociales como una instancia de las mismas.■

Referencias:

- [1] https://es.wikipedia.org/wiki/Interfaz_de_programación_de_aplicaciones
- [2] https://dev.twitter.com/es/overview/terms/agreement-and-policy
- [3] https://twitter.com/privacy
- [4] Aggarwal and C.X. Zhai (eds.), Mining Text Data