Ejemplo práctico: Text Mining con R

En este capítulo se va a desarrollar un caso práctico sobre minería de textos, con el fin de corroborar su utilidad y su importancia en el mundo actual. Para dicho análisis necesitaremos una fuente de información de donde extraer los textos a analizar, dicha fuente bien podría ser un artículo de opinión, discursos transcritos… Durante este capítulo, se procederá a realizar el proceso de minería de textos sobre publicaciones en redes sociales, en este caso Twitter.

Twitter es actualmente una dinámica e ingente fuente de contenidos que, dada su popularidad e impacto, se ha convertido en la principal fuente de información para estudios de Social Media Analytics. Multitud de empresas emplean este tipo de técnica para obtener información muy valiosa de individuos o de corporaciones. Análisis de reputación de empresas, productos o personalidades, estudios de impacto relacionados con marketing, extracción de opiniones y predicción de tendencias son sólo algunos ejemplos de aplicaciones.

Además de R existen otros lenguajes de programación como Python, MatLab u Octave. Si bien Python es el lenguaje que domina en este ámbito, este análisis se realizará a través de la programación en R, pues contiene tanto librerías, como paquetes que facilitan y extienden sus capacidades como herramienta de análisis de texto.

*Introducción.*

Tal y como ocurre en muchas redes sociales, Twitter otorga la posibilidad de compartir sus datos tanto con empresas como con desarrolladores y/o usuarios particulares. Aunque en la mayoría de casos se trata de web Services API, con frecuencia existen librerías que permiten interactuar con la API desde diversos lenguajes de programación.

La forma en la que Twitter permite acceder a su contenido es a través de lo que se conoce como Twitter App, al crear dicha Twitter App, se adquieren una serie de claves y tokens de identificación que permiten acceder a la aplicación y consultar la información necesaria.

Algo que se deberá tener en cuenta durante este análisis, es que Twitter tiene una normativa que regula la frecuencia máxima de peticiones, así como la cantidad máxima de tweets que se pueden extraer. Durante la fase de extracción de la información, se deberá tener en cuenta dichos limites con el objetivo de evitar ser sancionado por la organización.

*Acceso a la API*

Para acceder a la API de Twitter, como se indica en la documentación de la misma existen dos métodos de acceso Oauth2 y Oauth1a, los cuales dependerán del tipo de información que se desee extraer.

Como veremos a continuación el metodo de acceso será a través de Oauth1a puesto que se desea extraer información específica de algunos usuarios de la plataforma. Así pues, OAuth 1a para funcionar requiere de cuatro elementos que provienen de la aplicación de Twitter creada por el usuario desarrollador, clave del consumidor, clave del consumidor secreta, token de acceso y token de acceso secreto.

Todos estos recursos deberán ser proporcionados por la plataforma de Twitter parea desarrolladores una vez creada la aplicación con la cual se pretende acceder a la API.

*Extracción de datos*

Puesto que se pretende realizar una recopilación de aproximadamente 1000 tweets, se sigue la siguiente entrategia: