Anteproyecto de Trabajo Fin de Grado

David Márquez Mínguez

26 de noviembre de 2020

Título: Análisis y comparacion de paquetes para el desarrollo de web scraping

Departamento: Departamento de Ciencias de la Computación

Autor: David Márquez Mínguez

Tutor: Juan José Cuadrado Gallego

1 Introducción

El análisis de datos en particular y la ciencia de los datos de forma general, son disciplinas con múltiples facetas y enfoques, no solo en el campo informático sino en multitud de ambitos ya que permite obtener información a partir de una serie de datos.

En este trabajo la atención se centrará en la extracción de información a partir de fragmentos de texto. A diferencia de lo que comúnmente se conoce como minería de datos, en la minería de texto la información no se obtiene directamente a partir de los datos, sino que esta información se obtiene a partir de grandes cantidades de texto. Una vez extraida, la información no suelen estar ni estructurada ni limpia, por ello se deben realizar acciones de pre-procesamiento con el objetivo de poder realizar un análisis correcto.[**?**].

La minería de texto tiene multitud de aplicaciones y puede ser empleada en diferentes campos, no solo científicos. Algunas de las aplicaciones que se destacan son: búsqueda de información, reconocimiento de texto, clustering, clasificación, análisis de sentimientos... Estas aplicaciones se emplean a diario en grandes corporaciones como Amazon o Google para conocer nuestra opinión sin tener que ni siquiera que hacer uso de cuestionarios.

2 Objetivos y campo de aplicación

El objetivo fundamental del trabajo es la comprensión y el aprendizaje de las diferentes herramientas que permiten realizar un exhaustivo análisis sobre cualquier texto, ya sean cartas, artículos periodísticos, discursos transcritos... En este caso se analizarán publicaciones de diferentes usuarios en redes sociales con el objetivo de determinar si la minería de textos aplicada en dicho contexto funciona como método de análisis.

Como objetivos adicionales se pretende:

1. Describir y explicar las técnicas utilizadas tanto para la extracción de datos en cualquier texto, como para el análisis en cuestión, asi como técnicas adicionales existentes para el mismo proceso.

2. Realizar un caso de estudio en particular aplicando Text Mining y obtener la mayor cantidad de información posible de un texto cualquiera. Dicho caso de estudio contendrá varias etapas de desarrollo, en las que se deberán realizar técnicas como análisis exploratorios, limpieza de datos, análisis de sentimientos...

3. Por último se pretende realizar un estudio que muestre y compare los datos obtenidos durante el proceso analítico.

3 Descripcion del trabajo

Como se ha determinado en el apartado anterior, el objetivo fundamental del trabajo es el análisis de cualquier tipo de texto, asi como la exposición exhaustiva de los diferentes algoritmos empleados en el proceso y su funcionamiento correspondiente.

Como herramienta de prueba se ha decidido emplear publicaciones de diferentes usuarios en una determina red social, en concreto Twitter. Esta red social, al ser una plataforma que permite a multitud de usuarios de todo el mundo compartir opiniones o sentimientos sobre ciertos temas, parece un buen lugar donde poder realizar el análisis. Las publicaciones a analizar en concreto se denominan tweets y permiten a los usuarios expresarse con un máximo de 280 caracteres.

Como herramienta de análisis se pretende emplear el lenguaje de programación R, con el objetivo de disponer de una mayor flexibilidad en el análisis comparado con una aplicación software independiente. Si bien existen otras herramientas de programación como Python que dominan a la perfección este ámbito, R contiene librerías que facilitan y extienden capacidades como herramienta de análisis de texto.

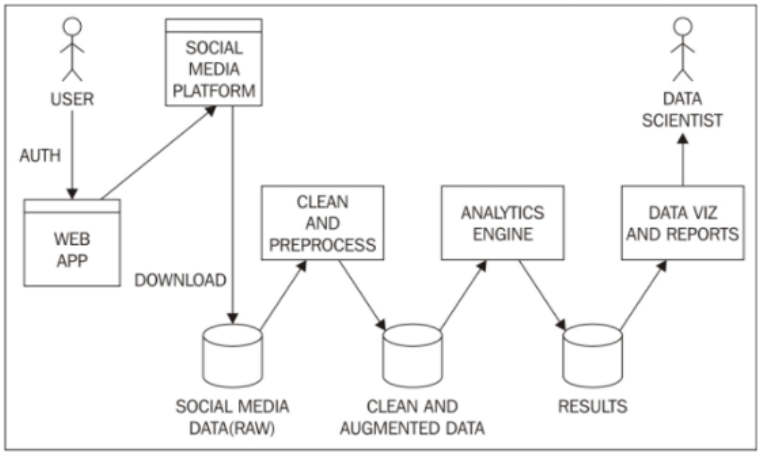

Figura 2: Minería de textos en redes sociales, esquema genérico.

Como se puede observar en la Figura 2, se muestra un esquema sobre el proceso general que se seguirá en cuanto al desarrollo del caso de estudio. Se determinan las etapas y procesos que se realizaran hasta conseguir los datos deseados.[**?**]

Como ocurre en muchas redes sociales, Twitter pone a disposición de los usuarios una API [**?**] que permite extraer información de la propia aplicación. A diferencia del resto de redes sociales Twitter no solo provee a los usuarios de una web services API, sino que permite la posibilidad de emplear librerías como rweet o twitteR que son capaces de comunicarse con dicha API.

4 Metodología y plan de trabajo

En consecuencia de los objetivos y campo de aplicación determinados en la sección 2, se detallan las fases de desarrollo del trabajo. En la Figura 3 se pueden apreciar las distintas tareas que compondrán el trabajo, así como las distintas subtareas de los mismos.

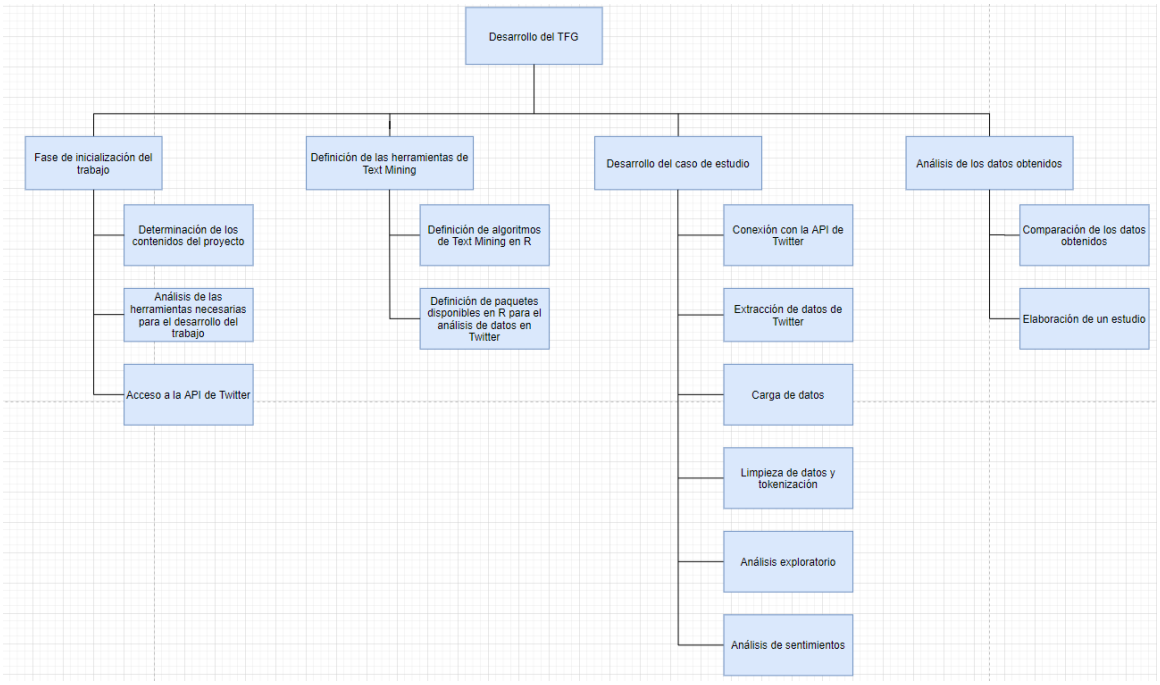

Figura 3: Estructura de descomposición del trabajo.

A continuación, se determina la dedicación aproximada a cada una de las fases que componen el proyecto.

1. Fase de inicialización del trabajo: (0,5 meses):
   - Determinación del contenido del proyecto.
   - Análisis de las herramientas necesarias para el desarrollo del trabajo.
   - Acceso a la API de Twitter.
2. Definición de las herramientas de Text Mining (2 meses):
   - Definición de algortimos de Text Minign en R.
   - Definición de paquetes disponibles en R para el análisis de datos en Twitter.
3. Desarrollo del caso de estudio (3 meses):

- Conexión con la API de Twitter.
- Extracción de datos de Twitter.
- Carga de datos.
- Limpieza de datos y tokenización.
- Análisis exploratorio.
- Análisis de sentimientos.

4. Análisis de los datos obtenidos (1 mes):

   - Comparación y muestra de los datos obtenidos.
   - Elaboración de un estudio.

En cada una de estas acciones no solo se deberá tener en cuenta la fase de procesamiento, sino que también se deberá tener en cuenta la correspondiente documentación del proceso realizado asi como las diferentes consultas bibliográficas que se crean oportunas.

5 Medios

Se determinan a continuación los medios necesarios para el desarrollo del trabajo, así como las herramientas que lo complementan.

En primer lugar, se empleará el lenguaje de programación R para el desarrollo del TFG así como la aplicación de paquetes y algoritmos dentro del mismo para realizar un correcto minado de los textos.[**?**]. Se pretende emplear RGui, aunque es posible emplear herramientas de código abierto como RStudio que permiten escribir y ejecutar programas en R.[**?**].

Se empleará Twitter como una herramienta de prueba, por ello se debe tener acceso a la API. Para que Twitter conceda acceso a su API es necesario crearse una cuenta Twitter Apps y justificar el uso que se hará con dicha herramienta.[**?**]

El trabajo se escribirá en LaTex, por ello será necesario la instalación de dicho compilador de textos, concretamente en su versión para Windows.[**?**]. Además, como se pretende incluir fragmentos de código R en el documento, se deberá utilizar el componente Sweave. [**?**].

Referencias