INTRODUCCION

El análisis de datos y la ciencia de los datos de forma general, es una disciplina con múltiples facetas y enfoques, no solo en el campo informático sino en multitud de disciplinas que permiten 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 través de los datos, sino que esta información se obtiene a partir de grandes cantidades de texto. La información obtenida a partir de fragmentos de texto no suele ser estructurada ni limpia, por ello se deben realizar acciones de post-procesamiento con el objetivo de poder realizar un análisis correcto con los datos obtenidos.

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.

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 redes sociales funciona como método de análisis.

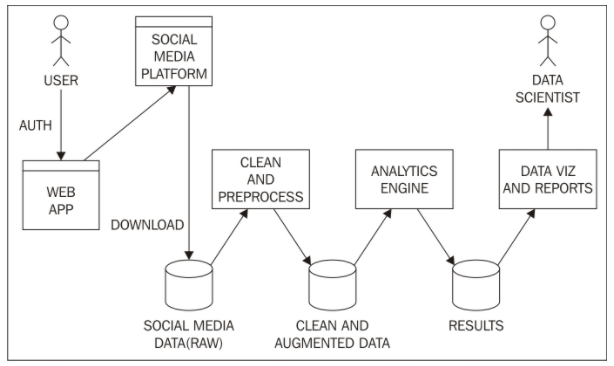
Como objetivos adicionales se pretende:

1. Describir y explicar las técnicas utilizadas tanto para la extracción de datos de la red social, como para el análisis en cuestión, asi como técnicas adicionales existentes para el mismo proceso.
2. Realizar un caso de estudio en particular analizando ciertas publicaciones de usuarios y poder obtener la mayor cantidad de información posible. 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 de análisis.

DESCRIPCIÓN 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 análisis se ha decidido emplear publicaciones de diferentes usuarios en una determina red social, por ejemplo 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 aprendizaje, se pretende emplear el lenguaje de programación R para realizar dicho análisis, 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.

Minería de textos en redes sociales, esquema genérico

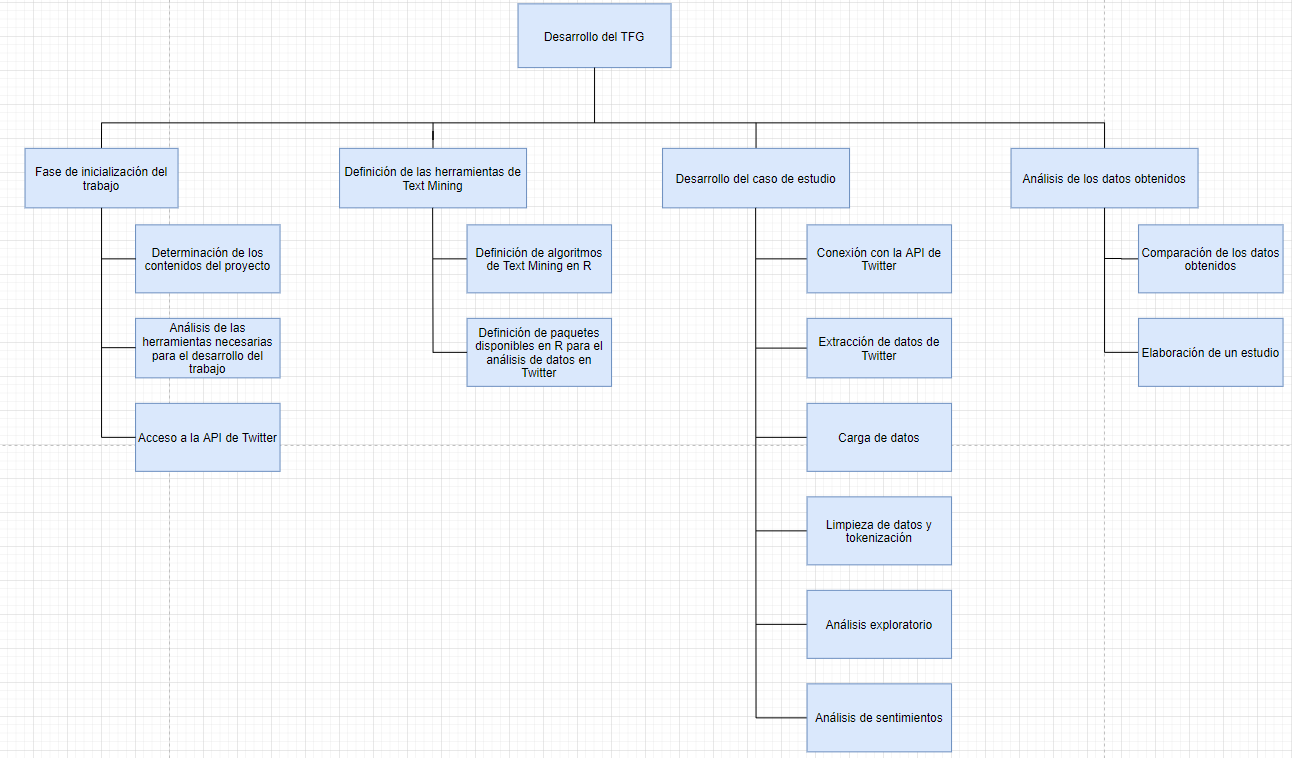
Como se puede observar en la figura 2, se muestra un esquema sobre el proceso general que se seguirá en cuanto al desarrollo del caso de estudio. Se determinan las etapas y procesos que se realizaran hasta conseguir los datos deseados.

Como ocurre en muchas redes sociales, Twitter pone a disposición de los usuarios una API que permite extraer información. A diferencia del resto de redes sociales Twitter no solo provee a los usuarios de una web cervices API, sino que es posible emplear librerías como rweet o twitteR que son capaces de comunicarse con la API.

La minería de datos y en concreto la minería de textos, es una herramienta muy útil no solo en el ámbito informático, sino en multitud de disciplinas que pretenden ampliar y mejorar el funcionamiento de sus aplicaciones. En este caso, se pretende combinar dicha herramienta con tecnologías que actualmente están en auge como son las redes sociales, y que además están en continuo crecimiento.

METODOLOGIA Y PLAN DE TRABAJO

Como consecuencia de los objetivos y campo de aplicación determinados en la sección 2, se detallan las fases de desarrollo del trabajo. En la figura se pueden apreciar las distintas tareas que compondrán el trabajo, donde cada una de las tareas se descompone en subtareas.

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

* .
* .
* .
* .

En estas acciones no solo se debe 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.

MEDIOS

Se determinan a continuación los medios necesarios para el desarrollo del trabajo, asi 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.\cite{r}. 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 análisis, 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 MikText. Además, como se pretende incluir fragmentos de código R en el documento, se deberá utilizar el componente Sweave.