CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO ENTREGABLE I

Preprocesamiento de datos, integración y gestión de datos mediante una DB NOSQL

INTRODUCCIÓN

En este primer Trabajo Práctico entregable del curso, se integrarán los conocimientos relacionados con el preprocesamiento de datos, integración, construcción de variables y gestión de datos mediante una Base de Datos NO-SQL orientada a documentos.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R y la Base de Datos MongoDB con el objetivo de ejercitar los conceptos abordados en las clases teóricas.

Los datos fueron generados con un script R[1] que *scrapea tweets* desde Twitter a partir de la API REST de la red social.

OBJETIVO GENERAL

El objetivo general de este trabajo es realizar un *análisis exploratorio* del dataset[2] y el posterior *preprocesamiento*, de acuerdo a las técnicas vistas en clase, a efectos de generar un modelo que clasifique los tweets de acuerdo a las características que hacen que sea *popular o no (posea retweets y favs)*[3].

Por otro lado, se busca extraer conocimiento del texto de los tweets, estudiando cuales son los términos más importantes.

Asimismo, se deberán realizar consultas sobre el dataset utilizando la sintaxis de consultas para Bases de Datos NoSQL.

Si bien el trabajo tiene un carácter netamente exploratorio y se definen las consignas de manera abierta, se evaluará la aplicación de las técnicas vistas en clase así como también el carácter innovador de la solución propuesta.

[1] El script *scraping-tweets.R* está publicado en el sitio de la materia.

[2] El dataset consta de dos colecciones: *tweets_mongo* (con características de tweets) y *users_mongo* (con las características de los usuarios que realizaron los tweets).

[3] Un tweet es considerado popular si posee al menos 1 marca (tweets o favs) por parte de otros usuarios.

CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

CONSIGNAS PROPUESTAS[4]

1. **ANÁLISIS EXPLORATORIO DE DATOS**
    a. A partir del dataset[5] elegido, explore sus características y detalle que *features* posee.
    b. De acuerdo a las técnicas abordadas en las prácticas de laboratorio, realice un análisis exploratorio para identificar cual es la distribución de sus variables y si existe relación entre las mismas[6].

2. **CONSULTAS NOSQL**: Defina las consultas que respondan a las siguientes preguntas:
    a. ¿Cuál es la proporción de tweets que contienen al menos una marca (RT o fav)?
    b. ¿Cuál es el usuario con mayor cantidad de tweets?
    c. ¿Cuántos usuarios twittean desde la aplicación web de Twitter?
    d. Seguramente notó que la colección contiene sólo tweets en español ¿Existe alguna cuenta que su lenguaje no sea español?
    e. Realice un perfil del usuario que posee mayor cantidad de seguidores. ¿Existe relación con la cantidad de marcas de sus tweets?

3. **PREPROCESAMIENTO:** A partir de las técnicas de preprocesamiento vistas en clase, y las features del dataset (tweets & users):
    a. Escoja y -en los casos que corresponda- transforme al menos 12 atributos del dataset que ayuden a predecir cuándo un tweet es popular. Deberá documentar paso a paso las operaciones realizadas y fundamentar su elección.
    b. Redefina el concepto de *popular* propuesto por el equipo docente y analice los cambios en la distribución de la clase con respecto a las instancias en el dataset. ¿Cuál sería el umbral óptimo a su criterio para que un tweet sea considerado popular? Justifique sus respuestas y aporte evidencias.

---

[4] En todas las consignas -principalmente en las relacionadas con análisis exploratorio y preprocesamiento- deberá documentar las actividades realizadas y argumentar sus decisiones en función de las técnicas (analíticas y gráficas) abordadas en clase.

[5] Puede importar el dataset a Mongo a partir de los archivos *dump_tweets.json* y *dump_users.json* utilizando el script *load-db.R* que está en el sitio de la materia o utilizando el comando *mongoimport* de **MongoDB**.

[6] Deberán buscarse relaciones entre las variables numéricas y también entre variables categóricas.

4. CONSTRUCCIÓN DEL MODELO

a. Performe un modelo a partir de un método de clasificación, por ejemplo con un árbol J48, en función del objetivo del trabajo. El proceso de construcción del modelo deberá ser iterativa en términos que deberá buscar la configuración del dataset que posea mayor precisión.
b. Documente estas iteraciones y haga referencia a cuales son las transformaciones con las que obtuvo una mejor configuración del modelo.

5. PROCESAMIENTO DE TEXTO:

a. Aplique las técnicas de procesamiento de texto abordadas en clase sobre el texto de los tweets.
b. Utilice una técnica gráfica que permita determinar cuáles son los términos que más aparecen en la colección de datos.