CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO ENTREGABLE I

Preprocesamiento de datos

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.

Los datos fueron generados con un script R 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 y el posterior *preprocesamiento*, de acuerdo a las técnicas vistas en clase para entender las relaciones existentes entre algunas variables del dataset.

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.

ACERCA DE LOS DATOS

Se trata de un conjunto de datos obtenidos desde la API REST de Twitter en formato JSON y que fueron almacenados en una base de datos NoSQL MongoDB conforme fueron descargados. A su vez se han generado dos colecciones una para el almacenamiento de los **tweets** y otra para almacenar los **usuarios**.

La estructura de datos de un tweet se muestra en la Figura 1, allí pueden observarse algunos de los componentes de la estructura jerárquica del documento. Para conocer más acerca de la estructura de datos de los tweets consulte [1] y para la descripción de los atributos del usuario [2].


```
{
        "created_at": "Thu May 10 17:41:57 +0000 2018",
        "id_str": "994633657141813248",
        "text": "Just another Extended Tweet with more than 140 characters, generated as a documentation example, showing that [\"tru... https://t.co/U7Se4NM7Eu",
        "display_text_range": [0, 140],
        "truncated": true,
        "user": {
                "id_str": "944480690",
                "screen_name": "FloodSocial"
        },
        "extended_tweet": {
                "full_text": "Just another Extended Tweet with more than 140 characters, generated as a documentation example, showing that [\"truncated\": true] and the presence of an \"extended_tweet\" object with complete text and \"entities\" #documentation #parsingJSON #GeoTagged https://t.co/e9yhQTJSIA",
                "display_text_range": [0, 249],
                "entities": {
                        "hashtags": [{
                                "text": "documentation",
                                "indices": [211, 225]
                        }, {
                                "text": "parsingJSON",
                                "indices": [226, 238]
                        }, {
                                "text": "GeoTagged",
                                "indices": [239, 249]
                        }]
                }
        },
        "entities": {
                "hashtags": []
        }
}
```


Figura 1: Estructura básica de un tweet

Los datos deben ser descargados del siguiente enlace:

https://drive.google.com/open?id=1JJboF06KyYgJBI7NtU2MwI1EjqLbqvn6

Los mismos fueron exportados con mongoexport y deben ser incorporados a la base con el comando **mongorestore** desde una consola CMD de Windows o una terminal Linux.

Archivo de usuarios[1]:

```
mongorestore -h localhost -d DMUBA -c users_mongo_covid19
--archive=./users_covid_curso2020.gz --gzip
```

Archivo de tweets:

```
mongorestore -h localhost -d DMUBA -c tweets_mongo_covid19
--archive=./tweets_covid_curso2020.gz --gzip
```

Importante: El parámetro **--archive** debe tener la ruta completa al archivo para que el comando mongorestore encuentre el archivo y pueda restaurarlo.

ALCANCE DEL TRABAJO

Se espera que puedan desarrollar los temas que fueron vistos durante las clases teóricas y prácticas, indagando en nuevos datos con complejidades propias de un problema del mundo real.

Los problemas principales a resolver son:

- Formular algunas preguntas que guíen el trabajo de descubrimiento de conocimiento.
- Análisis exploratorio: Será necesario que indaguen en los datos "crudos" para poder tomar decisiones de diseño de su trabajo. Por ejemplo, decidir

[1] De acuerdo a la versión de MongoDB, podría ser necesario utilizar el siguiente comando:
mongorestore -h localhost --nsInclude DMUBA.users_mongo_covid19 --archive=/users_covid_curso2020.gz --gzip

sobre aspectos de la integración, tareas de limpieza, identificador de presencia de ruido, etc.

- Integración de datos: La estructura jerárquica de los tweets va a requerir hacer algunas transformaciones para convertir las colecciones de tweets y usuarios en una matriz de datos que sea funcional.

  Además de integrar información de tweets y usuarios se podrán incorporar otras fuentes de información (ver Anexo) que puedan ser combinadas con la información de Twitter para enriquecer el análisis. (No es obligatorio realizar esta tarea, invierta el tiempo justo y necesario dado que puede llevar demasiado tiempo integrar datos de fuentes heterogéneas)

- Reducción de datos y dimensiones: Deberá tomar decisiones sobre qué variables duplican información, qué variables pueden dejarse afuera debido a su escaso aporte información para el objetivo de KDD.
- Limpieza de datos: Será necesario mejorar la calidad de los datos a partir de corregir inconsistencias, normalizar nombres, corregir tipos de datos, etc. Esta etapa se vuelve más compleja al incorporar análisis de los textos de los tweets donde la limpieza es clave a la hora de identificar a los objetos de estudio.
- Análisis de valores atípicos: Tratar de buscar no solo valores extremos sino también algunas observaciones que sean outliers de algún conjunto específico.
- Transformaciones: Varios de los análisis var a requerir re-escalado de variables o transformaciones logarítmicas. Tengan presente estas cuestiones durante los análisis.
- Generación de variables: Es posible combinar algunas variables para generar otras, por ejemplo si el usuario corresponde o no a una persona física o a una organización. La organización puede ser pública o privada, etc.
- Visualizaciones: Se espera que puedan hacer un análisis en el tiempo sobre algunas variables del dataset o analizar relaciones entre variables utilizando herramientas gráficas. Es importante que tengan presente que los gráficos

son una herramienta que facilita entender el problema, avalar o refutar alguna hipótesis de trabajo, es decir, es parte de la historia que quieren contar a partir de los datos. Por lo tanto, deben ser comprensibles por quien los vaya a leer. Todos los gráficos que se incorporen deben tener su correspondiente leyenda, nombres en los ejes, unidades de medidas, título, etc.

INFORME

El informe deberá contener las siguientes secciones:

- Título, nombre y apellido de los integrantes del grupo.
- Un resumen de hasta 200 palabras.
- Enumeración de las preguntas que se abordan en el análisis.
- Descripción de cada una de las etapas del proceso de KDD que se lleva a cabo para el trabajo (no incluir conceptos teóricos sólo sus aportes).
- La extensión máxima será de 15 páginas incluyendo texto, gráficos y anexos, evitando incluir código en el informe entregable.
- El informe deberá ser entregado en formato *pdf* (Portable Document Format) por correo electrónico al equipo docente.

CONFORMACIÓN DE GRUPOS

El trabajo deberá ser realizado en grupos, los cuales serán definidos por el equipo docente.

FECHA DE ENTREGA

El trabajo deberá ser entregado el día 9 de Junio 2020 hasta las 23:59 hs.

REFERENCIAS

[1] Introduction to Tweet JSON [link]
[2] User Object (Metadata) [link]

ANEXO

- Para tweets de Argentina datos de INDEC. (i.e radios censales) [link]
- European Centre for Disease Prevention and Control. Download today's data on the geographic distribution of COVID-19 cases worldwide [link]
- Datos.gob.ar https://datos.gob.ar/