Proyecto Integrador de saberes

Este proyecto busca que los estudiantes de las asignaturas de Fundamentos de Base de Datos y Programación Funcional y Reactiva demuestren lo que han aprendido durante todo el ciclo académico Octubre 2019 - Febrero 2020.

El proyecto se plantea los siguientes objetivos.

Objetivo general

- Realizar un análisis exploratorio de datos a un data set

Objetivos específicos

- Construir un modelo relacional de base datos que represente a las entidades que se encuentran presentes en el dataset
- Construir un conjunto de consultas SQL para extraer datos almacenados en la base de datos con el fin de proporcionar información relevante.
- Utilizar conceptos de programación funcional para realizar análisis exploratorio de datos.
- Construir un conjunto de visualizaciones que permitan ver el resultado de los análisis realizados.

Descripción

El análisis exploratorio de datos (EDA por sus siglas en inglés) es un enfoque para analizar conjuntos de datos (datasets) que permite resumir sus características principales, a menudo utilizando métodos de visualización de datos.

El EDA se debe realizar utilizando estadística descriptiva e inferencial, así como también utilizando correlación entre variables. Los resultados del EDA se deben visualizar utilizando la herramienta Vega-Lite (https://vega.github.io/vega-lite/). Para facilidad se sugiere que por cada análisis se cree una visualización independiente y no se trate de crear un sitio Web con todas las visualizaciones.

El dataset que se debe analizar está formado por información que proviene de la red social Twitter, es decir se trata de un conjunto de Tweets (el detalle del dataset lo encontrarán más adelante). Esta información se extrajo utilizando el API de búsqueda de Twitter, a través de la herramienta TAGS (https://tags.hawksey.info).

El dataset es un archivo separado por comas que posee 18 columnas. La descripción de cada columna es la siguiente:

Columna	Tipo	Descripción	Ejemplo
id_str	String	Representa al identificador único de un Tweet	1038336680929964032
from_user	String	El nombre del usuario, tal como lo ha definido	MarielaAyerve

text	String	El texto, en UTF-8, de la publicación.	RT @Salud_CZ7: #Loja | Cristian Abendaño, psicólogo clinico, aborda sobre la prevención del suicido. Enfatiza como factores de protección la.comunicación, resiliencia, habilidades sociales, tiempo de calidad @lahoraecuador https://t.co/CCAw6qB6SD
created_at	String	Fecha UTC cuando se creó este Tweet	Sat Sep 08 08:02:10 +0000 2018
time	DateTime	Fecha	08/09/2018 08:02:10
geo_coordinates	String	Puede ser nulo. Representa la ubicación geográfica de este Tweet según lo informado por el usuario o la aplicación del cliente.	loc: 18.501,-69.917
user_lang	String	Idioma con el que el usuario configuró su cuenta.	es
in_reply_to_user_id_str	String	Puede ser nulo. Si el Tweet es una respuesta, este campo contendrá la el ID del autor del Tweet original. Esto no siempre será necesariamente el usuario mencionado directamente en el Tweet.	6253282
in_reply_to_screen_name	String	Puede ser nulo. Si el Tweet es una respuesta, este campo contendrá el nombre de pantalla del autor original del Tweet	twitterapi
from_user_id_str	String	La cadena representa al identificador único para este usuario	915406661157818368

in_reply_to_status_id_str	String	Puede ser nulo. Si el Tweet es una respuesta, este campo contendrá la representación de cadena del ID del Tweet original.	1051222721923756032
source	String	Aplicación utilizada para publicar el Tweet. (Formato HTML)	Twitter for Android
profile_image_url	String	URL que apunta a la imagen de perfil del usuario	http://pbs.twimg.com/profile_images/915452554905038848/1rVz0aNX_normal.jpg
user_followers_count	Int	El número de seguidores que esta cuenta tiene actualmente. Bajo ciertas condiciones de coacción, este campo indicará temporalmente "0"	41
user_friends_count	Int	El número de usuarios que esta cuenta está siguiendo (también conocido como sus "seguidores"). Bajo ciertas condiciones de coacción, este campo indicará temporalmente "0"	83
user_location	String	Puede ser nulo. La ubicación definida por el usuario en el perfil de su cuenta.	San Francisco, CA
status_url	String	URL que apunta al tweet original	http://twitter.com/MarielaAyerve/statuses/1038336680929964032

entities_str	String	Entidades que se han analizado del texto del Tweet. (Formato JSON). Mayor información la pueden encontrar aquí: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/entities-object	{"hashtags":[{"text":"ODS","indices":[92,96]}],"symbols":[],"user_mentions":[{"screen_name":"aqua_vall","name":"AquaVall💧","id":875691904196907000,"id_str":"875691904196907008","indices":[3,13]},{"screen_name":"aqua_vall","name":"AquaVall💧","id":875691904196907000,"id_str":"875691904196907008","indices":[52,62]}],"urls":[{"url":"https://t.co/LMUKvAPXhx","expanded_url":"https://twitter.com/VisionResponsab/status/1108394166533394437","display_url":"twitter.com/VisionResponsa…","indices":[97,120]}]}

El dataset con el que deben trabajar contiene información acerca de la iniciativa de la Naciones Unidas denominada Objetivos de Desarrollo Sostenible ODS (https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/). La información en su mayoría se encuentra en español y corresponde a una recolección hecha en el mes de marzo de 2019.

Resultados esperados

Como resultado se espera que se genere como mínimo los siguientes análisis con sus respectivas visualizaciones:

- Número de Tweets y ReTweets por día.
- Número de Tweets y ReTweets por hora.
- Aplicaciones más utilizadas para publicar Tweets.
- Distribución de Hashtags.
- Distribución de menciones.
- Distribución de URLs
- Distribución de media.
- ¿Existe una correlación entre el número de amigos y la cantidad de seguidores?
- El comportamiento de los usuarios. Por cada usuario se debe presentar: la cantidad de seguidores y de amigos, también el número de Tweets y re-tweets.
- Cuántas veces se ha mencionado a un usuario.

Metodología de trabajo

Los estudiantes forman grupos de dos personas. A cada grupo se le asignará un dataset que deberá analizar y cumplir con los resultados esperados.

El control de versiones y documentación se lo llevará a través de Github y su wiki respectivamente. Para ello, él docente creará una tarea grupal a través de GitClassroom (http://classroom.github.com). Un miembro, de los grupos que se crearon previamente en clases, agregará a su compañero y a los docentes de Fundamentos de base de datos y Programación funcional y reactiva.

Cronograma tentativo

	Entregable 1	Entregable 2	Entregable 3
Semana 5	• Modelo funcional del dataset entregado • Población de la base de datos		
Semana 6		• Obtención de información de la base de datos a través de consultas SQL • Conjunto de archivos en formato CSV, obtenidos a través de procedimientos asociados al paradigma funcional para cumplir con los resultados esperados.	
Semana 8			• Visualizaciones de cada uno de los resultados esperados en páginas HTML independientes