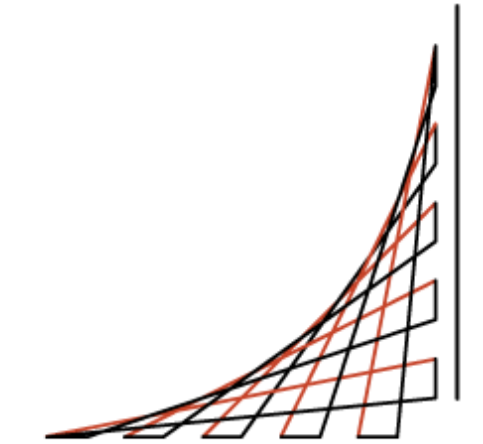
Big Data: Tweets in real time

Juan Manuel Villate Isaza
Juan David Navarro Jiménez

Big Data Visualización y Análisis Interactivo bajo Entornos Escalables y Multiplataforma



ESCUELA COLOMBIANA DE INGENIERÍA
JULIO GARAVITO

Índice

1. Introducción

En este proyecto se busca usar algunas de las técnicas y metodologías aprendidas durante el curso, para esto vamos a tomar un grupo de datos y analizaremos la diferente información que obtuvimos de los datos recolectados.
Entonces para la obtención de los datos vamos a consumir un API que nos permite obtener tweets relacionados a un tema, hemos decido elegir el tema del virus Covid-19 ya que es un virus que está afectando a todo el mundo y creemos que con este tema tendremos muchos más datos para poder analizarlos.

2. Descripción de los Datos

Este documento describe brevemente uno de los conjuntos de datos a ocupar para la asignación práctica del curso 'Big Data: Visualización y análisis interactivo en entornos escalables y multiplataforma'. También describe las principales preguntas a responder con la ayuda de visualización de datos.

2.1. Datos

Los datos 'Tweets about covid in real time' se almacenan como tabla, donde cada fila es un tweet, que menciona la palabra "Covid", el cual pudo ser publicado por cualquier usuario, ubicado en cualquier ciudad del mundo, las columnas indican varios atributos almacenados por tweet (por ejemplo, fecha de creación, usuario que lo tuiteó, ubicación en el mundo, etc.). estos datos fueron sacados haciendo peticiones a la API de Twitter (creamos cuentas Twitter de desarrollador para poder consumir la API de Twitter)

Las funciones para consultar en tiempo real los tweets está en el archivo analysisTweetsRealTime.ipynb en la carpeta del proyecto /slides

La documentación de la API de twiiter que usamos se puede encontrar en el siguiente link https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

2.2. Campos

- **Date (date time):** Registra el momento cuando se publica el tweet.
- **Hashtags (list):** Registra los hashtags que se mencionan en el tweet.
- **Text (text):** Registra el contenido del tweet.
- **User (text):** Registrar el username del usuario que creó el tweet.
- **Verified User (boolean):** Registra si el usuario está verificado en twitter.
- **Registered On (date time):** Registra la fecha en que el usuario se registró en twitter.
- **Followers User (number):** Registra cuántos seguidores tiene el usuario.
- **Following User (number):** Registra cuántos usuarios están siguiendo el usuario.
- **Tweets user (number):** Registra cuántos tweets tiene el usuario.
- **Favorites tweets User (number):** Registra cuántos tweets favoritos tiene el usuario.
- **Retweet (number):** Registra cuántos retweets tiene el tweet.
- **Quote (number):** Registra cuántas citas tiene el tweet.
- **Reply (number):** Registra cuántas respuestas tiene el tweet.
- **likes (number):** Registra cuantos likes tiene el tweet.
- **IsQuote (boolean):** Registra si trata de un Tweet citado.

- **Original Date (date time):** Registra el momento cuando se publicó el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original text (text):** Registra el contenido del tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original User (text):** Registra el username del usuario que creo el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original User Verified(boolean):** Registra si el usuario del tweet original está verificado en twitter (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original Location (text/location)):** Registra la ciudad y país de donde es el usuario que creo el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original Quote Count(number):** Registra cuántas citas tiene el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original Reply Count(number):** Registra cuántas Respuestas tiene el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original Retweet Count(number):** Registra cuántas Retweets tiene el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Original Favorite Count(number):** Registra cuántos Likes tiene el tweet original (Este campo sólo aparece cuando el Tweet es una cita de otro Tweet).
- **Language (text):** Registra el idioma con el cual se publicó el tweet.
- **Device (text):** Registra el dispositivo que se usó para publicar el tweet.
- **Location (text/location):** Registra el país y ciudad del tweet.
- **Latitude (location):** Registra la latitud de la ubicación donde se generó el tweet.
- **Longitude (location):** Registra la latitud de la ubicación donde se generó el tweet.

3. Preguntas

- ¿Cuál es la distribución de la ubicación de los tweets?
- ¿Cuál es la distribución y diferentes dispositivos usados para realizar los tweets?
- ¿Cuál es la distribución de idiomas en los tweets realizados?
- ¿Cuál es la variación entre todos los datos usuarios follows, followers y favoritos?
- ¿Cuál es la variación entre cuentas verificadas y no verificadas?
- ¿Cómo es la dispersión de los tweets en Colombia?
- ¿Cuál es el idioma más usado en tweets realizados?
- ¿Cuál es el tiempo de respuesta a tweet original?
- ¿Cuál es la diferencia de tiempo en la creación de las cuentas?

4. Desarrollo

4.1. Consideraciones Iniciales

4.1.1. Descargar el repositorio

Se debe descargar el repositorio de GitHub que se encuentra en el siguiente link https://github.com/villate13/BDproject el cual tiene el contenido desarrollado en el proyecto.

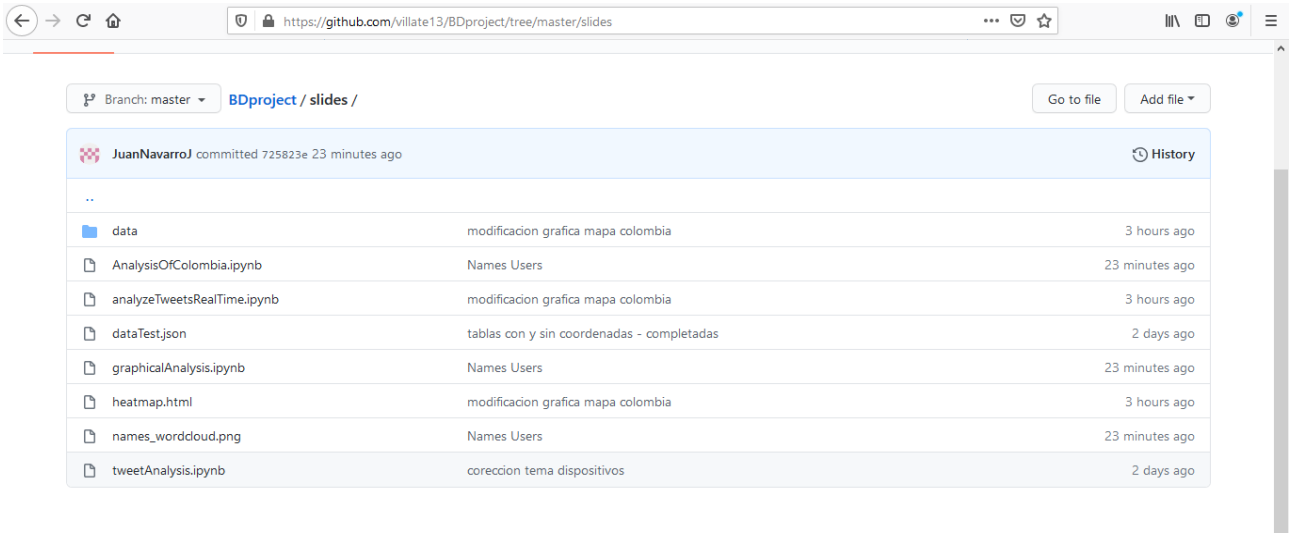


Figura 1: Repositorio del proyecto.

- **AnalysisOfColombia.ipynb :** En este Jupyter Notebook se realizo un mapa con la dispersión de los tweets únicamente en Colombia.
- **analyzeTweetsRealTime.ipynb :** En este Jupyter Notebook se realizó la implementación para la recolección de los tweets en tiempo real por medio de la Api de Twitter que utilizamos.
- **dataTest.json :** Este JSON es un archivo de muestra de como la API nos da el resultado de los tweets consultados.

- **graphicalAnalysis.ipynb :** En este Jupyter Notebook se realizaron las gráficas y análisis de el total de tweets obtenidos.
- **heatmap.html :** Este archivo Html es un mapa generado al final del Jupyter Notebook "analyzeTweetsRealTime" en el cual utilizando las coordenadas de los tweets generamos un tipo de mapa de calor gracias a GoogleMapPlotter.
- **names-wordcloud.png :** Esta es una imagen generada para visualizar algunos de las arrobas o usernames de los usuarios los cuales tomamos para hacer el análisis.
- **tweetAnalysis.ipynb :** Este Jupyter Notebook tiene la misma implementación del "analyzeTweetsRealTime .ipynb" pero este nos permitía consultar los tweets en otras fechas por medio de la Api de Twitter que utilizamos.

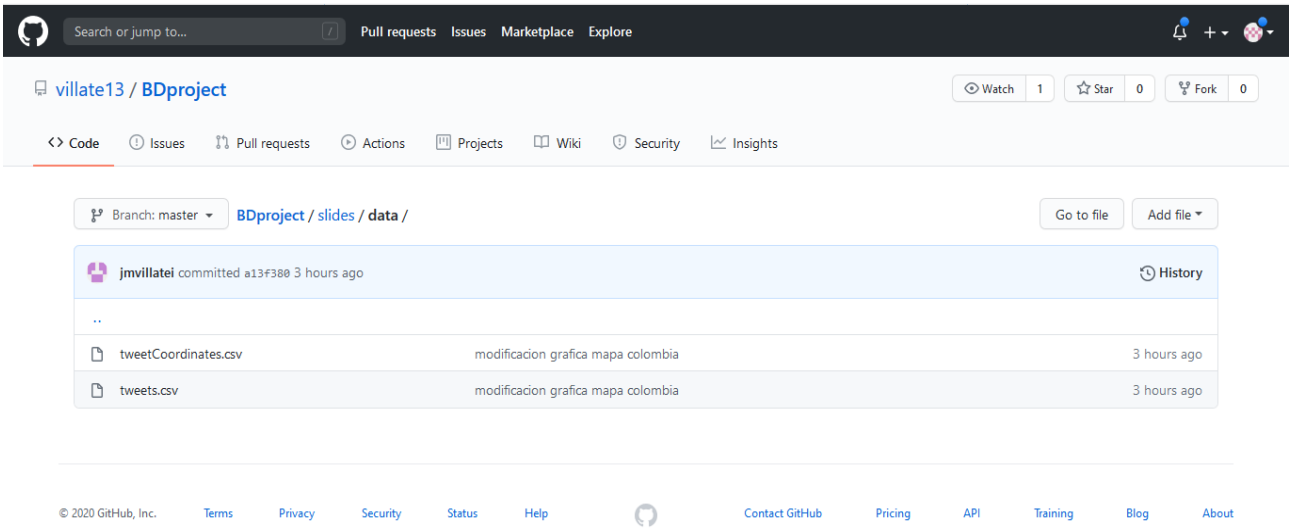


Figura 2: Repositorio del proyecto.

- **tweetCoordinates.csv :** Este archivo contiene todos los tweets de los usuarios que tienen registrada una ubicación de donde se realizo el tweet.
- **tweets.csv :** Este archivo contiene el total de tweets obtenidos durante la ejecución del código para consumir los tweets en tiempo real que se encuentra en Jupyter Notebook "analyzeTweetsRealTime.ipynb".

4.1.2. Imagen de docker

Para poder ejecutar el contenido de los jupyter notebooks es necesario correr la ultima imagen creada en el curso llamada "jupyter/my-datascience-vaex".

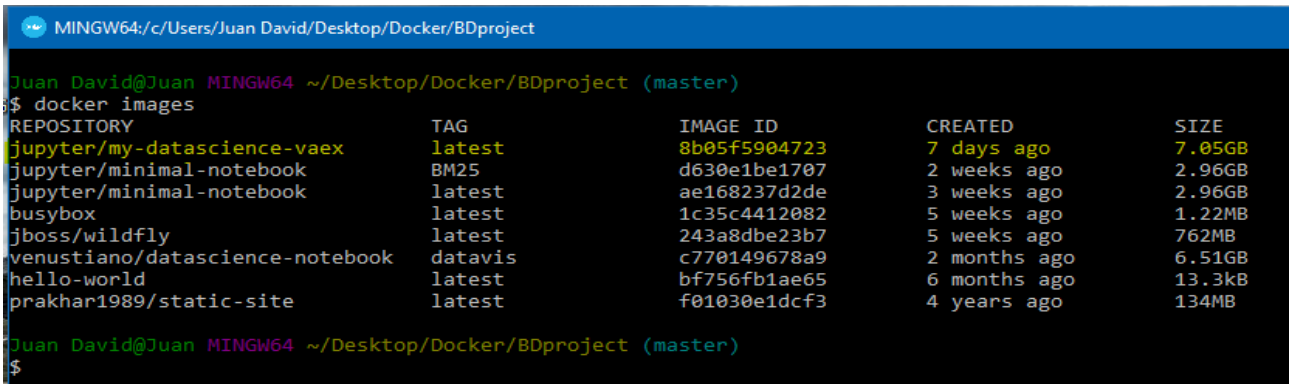


Figura 3: Imágenes de Docker.

4.1.3. Crear un contenedor

Para crear un contenedor basados en la imagen "jupyter/my-datascience-vaex" debemos ejecutar el siguiente comando: **docker run –rm -p 8888:8888 -v "$PWD":/home/jovyan/work jupyter/my-datascience-vaex**

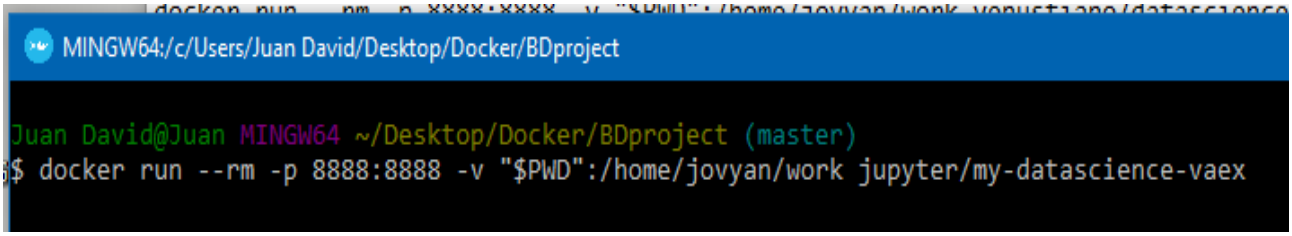


Figura 4: Ejecutar el contenedor.