

Apache Hadoop Análisis de Sentimiento

Introducción

Este proyecto recopila diferentes tweets sobre el coronavirus y hace un análisis de sentimiento de los tweets de cada usuario. Finalmente, crea un top 10 de los usuarios más pesimistas que tengan un total de tweets mayor que 10. Por ello, este es un top 10 de usuarios activos más pesimistas.

Datos

Se han recopilado 2 GB de datos utilizando la API de Twitter. Estos tweets eran tweets en inglés con los hashtags #coronavirus, #covid19 y #covid. Para un manejo más cómodo y rápido de los datos, se adjuntará un fichero .txt llamado sample_tweets el cual solo contendrá una muestra de dichos 2GB de tweets.

Funciones

Se han creado 4 funciones:

- mapper.py
 - Elimina los símbolos y letras de los tweets → # @ * RT
 - Devuelve el **nombre de usuario ||| texto del tweet ||| positividad del tweet en un rango de -1 a 1 ||| subjetividad del tweet en un rango de 0 a 1**
 - reducer.py
 - devuelve **nombre de usuario ||| lista de tweets ||| lista de polaridades ||| lista de subjetividades**
 - active_users_avg_sentiment.py
 - Teniendo en cuenta los usuarios que tienen más de 10 tweets, devuelve: **nombre de usuario ||| lista de tweets ||| lista de polaridades ||| lista de subjetividades ||| media de las polaridades ||| media de las subjetividades**
 - top_10_negative_users.py
 - Ordena los usuarios por negatividad y selecciona el top 10
 - Devuelve:
 - **THESE ARE THE TOP 10 MOST NEGATIVE USERS ACCORDING TO #COVID, #COVID19 AND #CORONAVIRUS IN TWITTER**
- USER #1**

USERNAME: username

Tweets:

Todos los tweets con sus polaridades y subjetividades

AVERAGE POLARITY OF THE USER IS: mediaPolaridad

AVERAGE SUBJECTIVITY OF THE USER IS: mediaSubjetividad

=====
=====

.....

=====
=====

USER #10

USERNAME: username

Tweets:

Todos los tweets con sus polaridades y subjetividades

AVERAGE POLARITY OF THE USER IS: mediaPolaridad

AVERAGE SUBJECTIVITY OF THE USER IS: mediaSubjetividad

Tiempos de Ejecución con Archivo de 2GB

- `hs ./mapper.py ./reducer.py covid_english_tweets.txt outputFirst`
 - 4m22,369s
- `hs ./active_users_avg_sentiment.py ./top_10_negative_users.py outputFirst/part-00000 outputFinal`
 - 0m7,450s

Ejecutar Demo Funciones con Hadoop

`hs ./mapper.py ./reducer.py sample_tweets.txt outputFirst`

`hs ./active_users_avg_sentiment.py ./top_10_negative_users.py outputFirst/part-00000 outputFinal`