Universitat Rovira i Virgili
Escola Tècnica Superior d'Enginyeria

SISTEMAS DISTRIBUIDOS

PRÁCTICA 2

AUTORES:
ALEX MAILLO JUNCAL
DAVID GASENI HIDALGO

DOCENTE:
Pedro Antonio García López

01/06/2021
2020 - 2021

ÍNDICE

BASES DEL PROYECTO

Esta práctica académica se ha centrado en la extracción de resultados de un seguimiento en la evolución sentimental de la sociedad, en la red social Twitter, durante el periodo de pandemia de COVID-19.

Se plantea la recopilación de mensajes desde inicios de 2020 hasta el mes de mayo de 2021. A partir de los diferentes mensajes se prevé cuantificar el sentimiento del mensaje en cuantías proporcionales en base a la alegría, neutralidad o enfado detectados en el texto.

DESARROLLO

Las herramientas que se utilizan en el proyecto son:

- Python
    - Vader sentiment
    - Pandas
    - Matplotlib
    - Numpy
- Twitter API
- IBM Cloud
- Lithops, *framework* para Python en *clouds*
- Docker

Obtención del *dataset:*

El *dataset* de *tweets* se ha recopilado a partir de las ID de éstos que proporcionan un equipo de recolección masiva de información de *Twitter*. Este trabajo está abierto al público y **proporciona solo las ID de los *tweet***. Se cita el *dataset* masivo en el apartado de referencias. Hemos partido a partir de este *dataset* de IDs, ya que la API de twitter sólo permite recolectar tweets de hasta una semana de antigüedad, cosa que si hacemos un *crawling* con el propio ID del tweet no ocurre.

A partir de las ID de los *tweets* se ha usado la API de Twitter para conseguir todos los demás datos: fecha, localización, si es *retweet* o no y el texto completo.
Esta obtención de los mensajes con información más completa, se ha ejecutado en el Cloud de IBM, siendo una función programada con *triggers*.
Los *triggers* permiten establecer eventos que pueden desencadenar la ejecución de la función, esto ha sido necesario porque la ejecución de IBM Cloud permite un máximo de 10 minutos a cada función. Además, como la versión utilizada de la API de Twitter solo permite un máximo de 900 peticiones cada 15 minutos se ha tenido que asegurar que cada ~10 minutos se ejecutase y que los datos se guardasen con iteraciones atómicas, para evitar perder *tweets* en el caso que haya un *timeout* tanto por el Cloud como por la API de Twitter.

Para permitir la ejecución remota en el Cloud de librerías extra de Python se debe proporcionar un *runtime* que las contenga, por esto se ha usado Docker y se ha subido un contenedor en DockerHub con Python y todas la librerías requeridas.

Tratamiento de los datos:

El primer paso para procesar la información ha sido la obtención del sentimiento del texto de los mensajes. Para realizar esta tarea de forma más rápida hemos usado el paralelismo que la librería lithops nos proporciona sobre el cloud, recogiendo así conjuntos de *tweets* en una lista para permitir la aplicación de la función de análisis de sentimiento con un *map*, de esta manera cada ‘iteración’ del *map* trata múltiples datos.

Muestra de resultados:

Para mostrar los resultados se han usado las librerías Pandas, Matplotlib y Numpy, que permiten la visualización de gráficas a partir de código Python y un tratamiento de datos más cómodo.

Las características principales que se tienen en cuenta en la evaluación de la información son el tiempo, el volumen de *tweets,* localización y la diferenciación entre *tweet* base y *retweet*. Los resultados han sido dirigidos hacia la observación de variaciones, tanto en volumen como en sentimiento, de los *tweets* en función de su origen, ya sea geográfico o si es un *retweet*.

CONCLUSIONES

El procesamiento en paralelo en el Cloud es una herramienta eficiente para el tratamiento de información masiva en paralelo, teniendo las herramientas necesarias.

En el caso de esta práctica, la obtención de información se ha realizado mediante una API con limitaciones de peticiones por tiempo, por lo tanto, en esta fase no se ha notado la paralelización del Cloud.

En la siguiente fase, de procesamiento de la información, se ha podido aplicar un mismo tratamiento a miles de elementos simultáneamente.

Los resultados se pueden ver en el repositorio GitHub: https://github.com/Buzzerage/SD_BigDataChallenge

Se han podido observar variaciones entre la pronunciación del sentimiento y la cantidad de *retweets* existentes. Además, esta posible relación afectaría de manera diferente según el origen geográfico de los datos.

REFERENCIAS

Dataset usado:
Banda, Juan M., Tekumalla, Ramya, Wang, Guanyu, Yu, Jingyuan, Liu, Tuo, Ding, Yuning, … Chowell, Gerardo. (2021). A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration (Version 62) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4767764