MEMORIA PRÁCTICA 2

Javier Núñez Bon

En este documento se detalla el funcionamiento del código desarrollado para abordar una solución de MapReduce del dataset Multilingual Amazon Reviews Corpus. En primer lugar se detallará el software encargado de recoger la información y analizarla y posteriormente los entornos donde se ejecutó.

- ***Algoritmo desarrollado***: La clase MapReduceJob se desarrolló en Python, y se encuentra en el fichero **MRMyJob.py**. Este software está constituido por una función de Map y dos Reducers, para así poder recabar la información pedida en el enunciado de la práctica.

    - Mapper: Esta función únicamente recopila la información necesaria de cada entrada del dataset que se vaya a utilizar a posteriori. En este caso, se recoge el valor de los siguientes campos:

        - *product_category*
        - *stars*
        - *language*
        - *product_id*
        - número de carácteres del campo *review_body*

    - Reducer 1 (Product Category Reducer): El primer Reducer recoge la información disponible por el Mapper y calcula a partir de ellas las siguientes métricas:

        - Estrellas totales (necesario para las estadísticas globales)
        - Número de reviews de cinco estrellas o más
        - Contador de productos por categoría
        - Contador de productos por idioma

      Además, se calcula en este reducer cuál es la *review_body* menos extensa, se eliminan los productos duplicados para saber cuantos hay en el dataset por *product_category*, y cuál es el idioma que posee más entradas. Finalmente, se producen dos salidas, una para identificar las estadísticas de la categoría y otra para las estadísticas globales.

        - Salida de variables del Reducer 1 para las estadísticas globales:
            - Diccionario de contador de idiomas
            - Extensión de review menos extensa
            - Contador de reviews de 5 estrellas

- Salida de variables del Reducer 1 para las estadísticas por categoría:
    - Promedio de estrellas
    - Idioma más popular
    - Contador de productos sin duplicados
    - Reviews totales

- Reducer 2 (Final Reducer): Este Reducer es el encargado de producir la salida final acorde a lo pedido en la práctica. Analizando la salida del primer Reducer, recopila la información y la ordena para guardar en un fichero *output.json* los siguientes datos:

    - Por categoría:
        - Average number of stars for the product category
        - Language with the maximum number of reviews for this product category.
        - Total number of different products ("product_id") per product category.
        - Total number of reviews per category

    - Globales:
        - Total number of reviews per language.
        - Minimum Number of characters for all the reviews.
        - Total number of reviews with 5 stars.

- ***Prueba en local - Pseudo Distributed Operation***: Se instaló Hadoop en el equipo local y se ejecutó el algoritmo de MapReduce desarrollado para analizar los datasets mencionados, ubicados en el sistema de ficheros del propio Hadoop (HDFS).