PRÁCTICA SPARK BICIMAD
GRUPO 10: Íñigo Iriondo, Álvaro Pleguezuelos y Alejandro Hernando

Definición del problema y motivación

El objetivo de esta práctica es analizar los datos del sistema de préstamo de bicicletas BICIMAD en Madrid para identificar patrones de uso. En particular, veremos qué estaciones son las más y menos populares durante diferentes franjas horarias del día y los códigos postales con mayor uso a lo largo del año (abril 2017 - marzo 2018).

Diseño e implementación de la solución

Hemos implementado un script que utiliza *PySpark* que consiste en:

1) Importación de *SparkContext* de *pyspark*, *json* y *matplotlib.pyplot*.
2) *map_line_to_dict*: función que convierte cada línea de los datos en un diccionario.
3) Creación de la clase *AnalisisBiciMad*: inicializa el contexto de *Spark* y lee los datos desde los archivos proporcionados y los analiza con las siguientes funciones:
    a) *group_by_key_count*: agrupa los datos y cuenta cuántos registros hay para cada clave.
    b) *analizar_datos*: es la función principal. Separa los datos en mañana, tarde y noche y utiliza *group_by_key_count* para encontrar las estaciones de bicicletas más y menos utilizadas en cada franja horaria. Después determina los cinco códigos postales más populares y, para cada estación del año, calcula y grafica los números de viajes en estos.
4) Inicialización de la clase que llama a *analizar_datos* con el input de los archivos (datos de abril 2017 a marzo 2018).

Evaluación de resultados y aplicación de la solución al conjunto de datos

El output que recibimos es:
La estación más usada por la mañana es la estación número 163 con 17283 viajes.
La estación más usada por la tarde es la estación número 43 con 26458 viajes.
La estación más usada por la noche es la estación número 57 con 13033 viajes.
La estación menos usada por la mañana es la estación número 2008 con 32 viajes.
La estación menos usada por la tarde es la estación número 2008 con 11 viajes.
La estación menos usada por la noche es la estación número 119 con 536 viajes.
Top 5 códigos postales con más viajes:
1. CP Desconocido con 1247466 viajes
2. CP 28005 con 253204 viajes
3. CP 28012 con 231182 viajes
4. CP 28007 con 205815 viajes

5. CP 28004 con 204194 viajes

De estos resultados podemos concluir que las estaciones 163, 43 y 57 deben estar preparadas para una alta demanda por la mañana, tarde y noche respectivamente, mientras que la 2008 tiene poco uso por la mañana y tarde que incluso podríamos prescindir de ella. También vemos que los códigos postales donde BICIMAD es más popular son el 28005, 28012, 28007 y 28004, especialmente en verano y otoño por lo que para esas fechas tendrían que estar todas las bicicletas y estaciones en perfectas condiciones. En cambio, en invierno baja ligeramente la demanda por lo que podría ser un buen momento para hacer reparaciones masivas o mantenimiento que podría inhabilitar el servicio.
Finalmente, es importante notar que obtenemos códigos postales vacíos lo cual sería algo a mejorar de la recogida de datos en el futuro.