Estudio conjunto de datos BICIMAD mediante SPARK

UNIVERSIDAD
COMPLUTENSE
MADRID

Programación paralela

Curso 2022-23

Teresa Ballester Navarro y David González Morala - Grupo 15

Madrid, Mayo de 2023.

1. Planteamiento de la práctica y motivación

El transporte sostenible y eficiente se ha convertido en una prioridad a la hora de elegir el medio de desplazamiento por parte de los ciudanos. En grande ciudades como Madrid, la implantación de un buen sistema de transporte público es necesaria, pues es una forma de moverse rápido por la ciudad. Una de las opciones más populares es el sistema de alquiler de bicicletas conocido como BICIMAD. Sin embargo, ha surgido una nueva empresa competidora que busca meterse en el mercado y ofrecer una alternativa a los usuarios de este servicio.

El objetivo de este análisis es estudiar la competencia que supone BICIMAD para esta nueva empresa. Para lograrlo, examinaremos diferentes aspectos que incluyen la cobertura geográfica, la experiencia del usuario, el rango de edad y la frecuencia de uso de este servicio.

Entender la competencia es esencial para tomar decisiones estratégicas a la hora de tener éxito y oportunidades de meterse en el mercado.

2. Conjunto de datos

En esta práctica de programación paralela diseñamos y emplementamos el análisis utilizando el entorno Spark. El dataset sobre el que trabajamos es el proporcionado por el ayuntamiento de Madrid del uso del sistema de bicicletas de préstamos de BICIMAD.

En concreto vamos a analizar 'Junio 2019', pues la empresa piensa comenzar a instaurarse en septiembre de 2019. Estudiaremos en profundiad los siguientes aspectos:

- Estudio de la frecuencia de uso del alquiler BICIMAD para realizar ciclos.
- Uso de ls bicicletas para dar un paseo. Es decir, la existencia de caminos y la frecuencia de estos.
- El rango de edad de los usuarios de BICIMAD y la frecuencia de uso del alquiler de bicicletas por rango de edad.
- Las estaciones más frecuentadas y menos.

Para ello haremos uso de los siguientes parámetros, que corresponden a las líneas correspondientes del fichero *.json*:

- _id : Identificador de la bicicleta
- user_day_code : Código de usuario diario
- idunplug_station : Número identificador de la estación donde se engancha la bicicleta.

- idplug_-station : Número identificador de la estación donde se engancha la bicicleta.
- travel_time : Tiempo total en segundos, entre el alquiler de una bicicleta y la devolución de esta en una estación.
- ageRange : Número que indica el rango de edades de los usuarios que han alquilado una bicicleta. Sus posibles valores son:
  - $0 --> 0$ a 11 años
  - $1 --> 12$ a 17 años
  - $2 --> 18$ a 24 años
  - $3 --> 25$ a 44 años
  - $4 --> 45$ a 64 años
  - $5 --> +65$ años

3. Análisis resultados

Para trabajar de forma paralela con los datos, utilizaremos RDD's (Resilient Distributed Datasets) y el entorno Spark. Los RDD's permiten el procesamiento distribuido de datos.

En primer lugar, transferimos los datos desde los archicos *.json* a un RDD mediante la biblioteca Spark, para trabajar paralelamente con ellos. Como punto de partida creamos un contecto de Spark, *sc*:

```
conf = SparkConf().setAppName('PracticaSpark')
sc = SparkContext(conf = conf)
```

Para filtrar los parámetros que necesitamos para el estudio, creamos la función *FiletoDic()* que dada una línea del archivo *.json*, la transforma en un diccionario (mediante la función *json.loads()* y nos devuelve la información deseada, explicada en 2.

Una vez definida esta función y obtenido los datos necesarios para nuestro análisis, aplicamos *filetoDic()* a cada uno de los elementos de nuestro RDD mediante *map*:

```
rdd1 = rdd.map(FiletoDic)
```

Los resultados obtenidos mediante el análisis realizado se muestra a continuación en formato de 4 tablas. La primeras dos tablas muestran el análisis hecho a los usuarios que realizan un ciclo , y de esos mismos usuarios analizamos los usuarios que realizan un camino. Las siguientes dos tablas muestran el análisis hecho a los usuarios en base a su edad.

Cuadro 1: Usuarios que realizan ciclos/caminos

	Tiempo medio	**número de usuarios**
Ciclos	27.27 min	16336 usuarios
Caminos	45.10 min	4780 usuarios

	Estación más frecuente/Veces usada	**Estación menos frecuente/Veces usada**
Ciclos	135/363 usos	2008/6 usos
Caminos	64/178 usos	28/1 uso

En las dos primeras tablas se observa que el tiempo medio aumenta si nos centramos en los usuarios que realizan un camino. Sin embargo si comparamos el número de usuarios se comprueba que la cantidad baja considerablemente en trayectos que corresponden con caminos. Por otro lado, nos fijamos en que la estación más frecuentada por usuarios que realizan un camino es la 64, por lo que una posibilidad sería colocar una mayor cantidad de bicicletas en esa estación. Análogamente también sería productivo una cantidad mayor en la estación 135.

Respecto a las otras dos tablas podemos observar que los adolescentes permanecen en promedio más tiempo usando las bicicletas que el resto de usuarios que pertenecen a otro rango de edad. Un dato llamativo es el número de usos por parte de los niños, pues resalta significativamente por encima del resto de resultados en ese apartado. Esto se puede deber a que junio es la entrada del verano y la subida de las temperaturas en esta época junto con el fin de las clases de los colegios e institutos hace que más niños usen las bicicletas con respecto al resto de personas de otras edades.

4. Instrucciones para la ejecución

Dado que toda nuestra implementación del análisis está en un archivo IPYNB, es decir, un documento de cuaderno usado por Jupyter Notebook. Aparece explicado paso a paso el análisis y es necesario indicar la ruta del archivo *.json* en esta parte del documento:

Cuadro 2: Usuarios según su rango de edad

	Tiempo medio	**Número de usos**
Niños	18.64 min	199830
Adolescentes	31.03 min	3628
Jóvenes	14.10 min	1928
Jóvenes/Adultos	15.08 min	23712
Adultos	15.10 min	12421
Mayores	17.14 min	95046

	Estación más frecuente/Veces usada	**Estación menos frecuente/Veces usada**
Niños	135/2949 veces	28/162 veces
Adolescentes	42/127 veces	43/1 vez
Jóvenes	82/81 veces	98/1 vez
Jóvenes/Adultos	83/434 veces	2008/5 veces
Adultos	129/2172 veces	207/100 veces
Mayores	90/1391 veces	207/54 veces

```
with open·$(r"C:/Users/j4gon/Downloads/junio2019.json", encoding='latin1')$ as f:\\
data = f.readlines()\\
rdd = sc.textFile$(r"C:/Users/j4gon/Downloads/junio2019.json")$
```

Jupyter Notebook es un entorno computacional interactivo diseñado para ayudar a los científicos a trabajar con el lenguaje Python y sus datos.

5. Conclusiones

Tras analizar lo datos correspondientes a BICIMAD en junio de 2019 concluimos que en la época entrante a verano es conveniente poner más bicicletas de tamaño más pequeño para que así más niños puedan utilzarlas.Concluimos que hoy en día las bicicletas siguen suponiendo una gran alternativa como medio de transporte, por lo que nuestra compañía tratará de ofrecer los mejores servicios para el uso de bicicletas.