Facultad de Ciencias Matemáticas

Programación Paralela

curso 2021-2022

Práctica 4: Análisis de datos de Bicimad con Spark

Esther Arribas García, Lucía Bragado Pérez, Álvaro Seco Silva

Mayo, 2022

Índice

1. Introducción

En esta práctica se trabaja en el planteamiento, diseño e implementación de una solución a un problema de análisis de datos reales utilizando Spark. El conjunto de datos de estudio es el proporcionado por el ayuntamiento de Madrid sobre el uso del sistema de bicicletas de préstamo Bicimad.

Realizaremos una comparativa del uso de este servicio pre y post pandemia, principalemente el tiempo de uso según el grupo de edad, la cantidad de salidas y llegadas según el barrio, y el número de usuarios totales al mes.

Para visualizar los resultados obtenidos, mostraremos histogramas y tablas.

2. Motivación y definición del problema a resolver

Con vistas a realizar un análisis de datos que pueda ofrecer resultados interesantes desde el punto de vista social, consideramos que una buena idea sería hacer una comparativa del uso de Bicimad antes y después de la pandemia por COVID-19.

De esta manera, seleccionamos los datos correspondientes a los meses de Mayo a Octubre (ambos inclusive) de los años 2019 y 2020.

Analizamos los siguientes puntos de dicho conjunto de datos:

- Tiempo medio de uso de la bicicleta en función del grupo de edad (6 meses pre-confinamiento VS 6 meses post-confinamiento).
- Cantidad de salidas y llegadas por barrio (pre-confinamiento y post-confinamiento).
- Cantidad de usuarios (sin repeticiones) de Bicimad cada mes (pre-confinamiento VS post-confinamiento).

3. Datos empleados

Los datos que hemos empleado en el estudio están disponibles el siguiente enlace Datos Bicimad proporcionados por el Ayuntamiento de Madrid.

De este conjunto de datos extraemos los correspondientes a los meses que nos interesan. Los datos de cada mes vienen dados por separado en un archivo con extensión *.json*.

Cada línea de estos ficheros es un objeto de JavaScript, con los siguientes atributos:

- **_id**: identificador del movimiento
- **user_day_code**: código del usuario (para una misma fecha todos los movimientos de un mismo usuario tienen el mismo código).
- **idplug_base**: número de la base en la que se engancha la bicicleta.
- **user_type**: número que indica el tipo de usuario que ha realizado el movimiento. Como los posibles valores cambian del año 2019 al 2020, no tendremos en cuenta este atributo en el análisis comparativo pre-post pandemia.
- **idunplug_base**: número de la base de la que se desengancha la bicicleta.
- **travel_time**: tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.
- **idunplug_station**: número de la estación de la que se desengancha la bicicleta.

- **ageRange**: número que indica el rango de edad del usuario que ha realizado el movimiento. Sus posibles valores son:

  0: No se ha podido determinar el rango de edad del usuario

  1: El usuario tiene entre 0 y 16 años

  2: El usuario tiene entre 17 y 18 años

  3: El usuario tiene entre 19 y 26 años

  4: El usuario tiene entre 27 y 40 años

  5: El usuario tiene entre 41 y 65 años

  6: El usuario tiene 66 años o más

- **idplug_station**: número de la estación en la que se engancha la bicicleta.
- **unplug_hourTime**: Franja horaria en la que se realiza el desenganche de la bicicleta (se facilita la hora de inicio del movimiento, sin la información de minutos y segundos)
- **zip_code**: número de la estación de la que se desengancha la bicicleta.

FALTA LA PARTE DE DONDE HEMOS COGIDO LO DE LOS BARRIOS.

4. Diseño e implementación de la solución

Para trabajar de forma paralela con los datos, usaremos *rdd*.

En primer lugar, para el análisis del tiempo medio en función de la edad y las salidas y llegadas por barrio creamos dos rdd, uno para cada año (*rdd19* y *rdd20*), uniendo los archivos descargados de cada mes.

En segundo lugar, para el apartado de cantidad de usuarios por mes pre y post confinamiento, consideramos dos listas, una para cada año, *lrdd19* y *lrdd20* cuyos elementos son los rdd independientes correspondientes a cada mes(sin unir, a diferencia de *rdd19* y *rdd20*).

Para el análisis del tiempo medio en función del grupo de edad en cada año, de cada línea del fichero rdd (*rdd19* o *rdd20*) obtenemos la tupla (grupoEdad, tiempoViaje) a través de un map (*mapper_edad*). Después agrupamos estos datos por grupo de edad con un groupByKey obteniendo tuplas (grupoEdad, listaTiempos-Viaje). Finalmente, para cada grupo de edad hacemos la media de los elementos de la listaTiempos-Viaje para obtener la media por grupos de edad. Representamos el resultado en un histograma doble.

FALTA EL ANÁLISIS DE LAS SALIDAS Y LLEGADAS SEGÚN BARRIO.

Para el análisis de la cantidad de usuarios por mes pre y post confinamiento, para cada lista *lrdd19* y *lrdd20* y cada elemento de dichas listas (que corresponde a un mes) transformamos cada línea del fichero rdd a una única variable códigoUsuario a través de un map (*mapper_usuario_unico*). Despúes, eliminamos códigos repetidos (pues queremos contar usuarios diferentes) con un groupByKey. Una vez eliminadas las repeticiones, obtenemos la cantidad de usuarios como la longitud de la lista de códigos después de la agrupación.

5. Instrucciones para ejecutar el código

Para poder obtener los resultados se necesita ejecutar el código *bicimad.py* desde la misma carpeta donde estén guardados los archivos *json* correspondientes a los meses de Mayo a Octubre de 2019 y 2020, con el nombre que reciben al ser descargados desde Datos Bicimad, esto es, los siguientes 12 archivos:

201905_Usage_Bicimad.json,

201906_Usage_Bicimad.json,

201907_movements.json,

201908_movements.json,

201908_movements.json,

201909_movements.json,

201910_movements.json,

202005_movements.json,

202006_movements.json,

202007_movements.json,

202008_movements.json,

202009_movements.json,

202010_movements.json.

6. Evaluación de resultados

La comparativa pre-post confinamiento del tiempo medio de uso de la bicicleta en función del grupo de edad se muestra en el siguiente diagrama de barras:

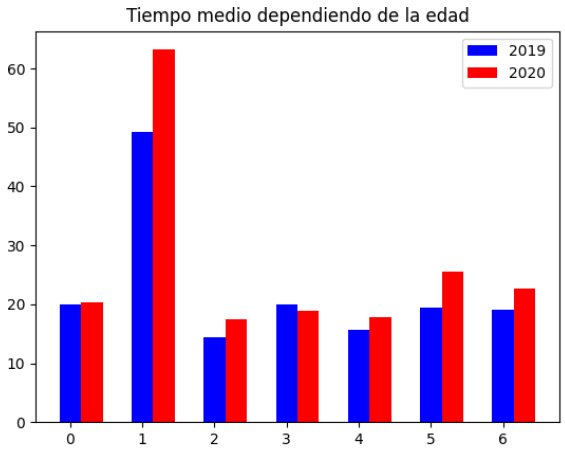

Figura 1: Tiempo medio (en minutos) de uso según grupo de edad pre-post pandemia

Notamos que hay el grupo de edad 1, correspondiente a niños y adolescentes (de 0 a 16 años) es el grupo que usa el servicio de Bicimad por mayor tiempo tanto en pre-confinamiento (unos 50 min de media) como en post-confinamiento (una hora de media). El resto de grupos de edad obtienen tiempos medios muy similares entre sí en ambos años (en el rango de los 15 min a los 25min).

En cuanto a la comparativa pre-post pandemia, el tiempo medio de uso después de la pandemia suele ser ligeramente superior al de antes de la pandemia, salvo para el grupo de edad 3 (de 19 a 26 años) en la que el tiempo medio es mayor antes de la pandemia. Estos resultados resultan coherentes, pues los ciudadanos que decidían usar el servicio de Bicimad posiblemente repetían los mismos trayectos que hacían antes de la pandemia, obteniendo tiempos medios similares.

FALTA EL APARTADO 2

El siguiente histograma doble muestra la comparativa pre y post pandemia de la cantidad de usuarios únicos

del servicio de Bicimad:

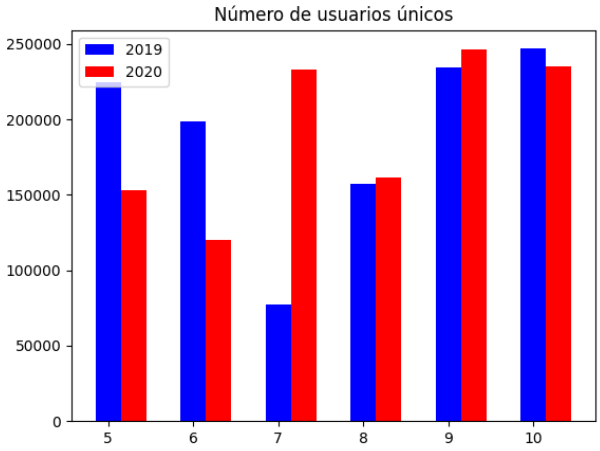

Figura 2: Cantidad de usuarios únicos de Bicimad en cada mes pre y post pandemia

Esta gráfica nos proporciona mucha información desde el punto de vista social. En primer lugar, en el mes de Mayo, correspondiente al momento en que se volvió a las calles después del confinamiento, había cierta reticencia a usar Bicimad, siendo la diferencia respecto al año anterior mayor que 50000 usuarios (CALCULAMOS EL NÚMERO EXACTO?no es complicao). Esta diferencia se mantuvo durante el mes de Junio.

Sin embargo, en el mes de Julio observamos un gran cambio: el uso en 2019 cae drásticamente (posiblemente debido a la gente abandonando la capital por vacaciones) mientras que en 2020 aumentan los usos en unos 100000 usuarios (NUM EXACTO?), probablemente por el miedo de la ciudadanía a salir de vacaciones y exponerse a un mayor contagio, usar la bicicleta fue un gran refugio para muchas personas.

Tras el mes vacacional de Julio, en los sucesivos meses de Agosto, Septiembre y Octubre tras la pandemia se consiguió equilibrar la cantidad de usuarios que había el año anterior, ratificando la vuelta a la normalidad.

7. Conclusión

En definitiva, en esta práctica se ha trabajado con la librería *pyspark* de Python para hacer un análisis comparativo del servicio Bicimad antes y después de la pandemia.

Para ello hemos trabajado con RDDs (Resilient Distributed Datasets), que sirven para operar de forma paralela con un conjunto inmutable y particionado de datos.

Los resultados obtenidos nos permiten obtener conclusiones significativos de un servicio real en nuestra ciudad.