UCM
4º GRADO EN MATEMÁTICAS C.COMPUTACIÓN

Práctica 4

Cristina Pino Bodas

Programación paralela
Pyspark

Resumen

Este documento recoge los resultados de los problemas planteados para llevar acabo un análisis de los datos de Bicimad del año 2019.
Dichos datos han sido obtenidos mediante la página web de bicimad:
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=d67921bb86e64610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default
En concreto hemos utilizado los datos de los usuarios referentes a los meses Enero, Febrero y Junio de 2019. Además de estos también hemos utilizado los datos de las estaciones correspondientes a Enero de 2019.

A continuación exponemos los problemas y resultados obtenidos mediante un análisis utilizando pyspark.

PROBLEMA 1

El primer problema que planteamos es conocer qué tipo de usuarios utilizan Bicimad. Estos se clasifican según el tipo de usuario de la siguiente forma:
0: No se ha podido determinar el tipo de usuario
1: Usuario anual (poseedor de un pase anual)
2: Usuario ocasional
3: Trabajador de la empresa

En primer lugar, observamos que en Enero de 2019 la mayoría de los usuarios son del tipo 1, es decir, usuarios anuales. A continuación, constatamos este resultado con los meses de Febrero y Junio en los cuales obtenemos el mismo resultado. Por tanto, se puede concluir que la mayoría de usuarios de Bicimad son usuarios anuales.

PROBLEMA 2

Ahora queremos conocer que grupo de edad es el que más usa bicimad. Los grupos de edades se representan de la siguiente forma:
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

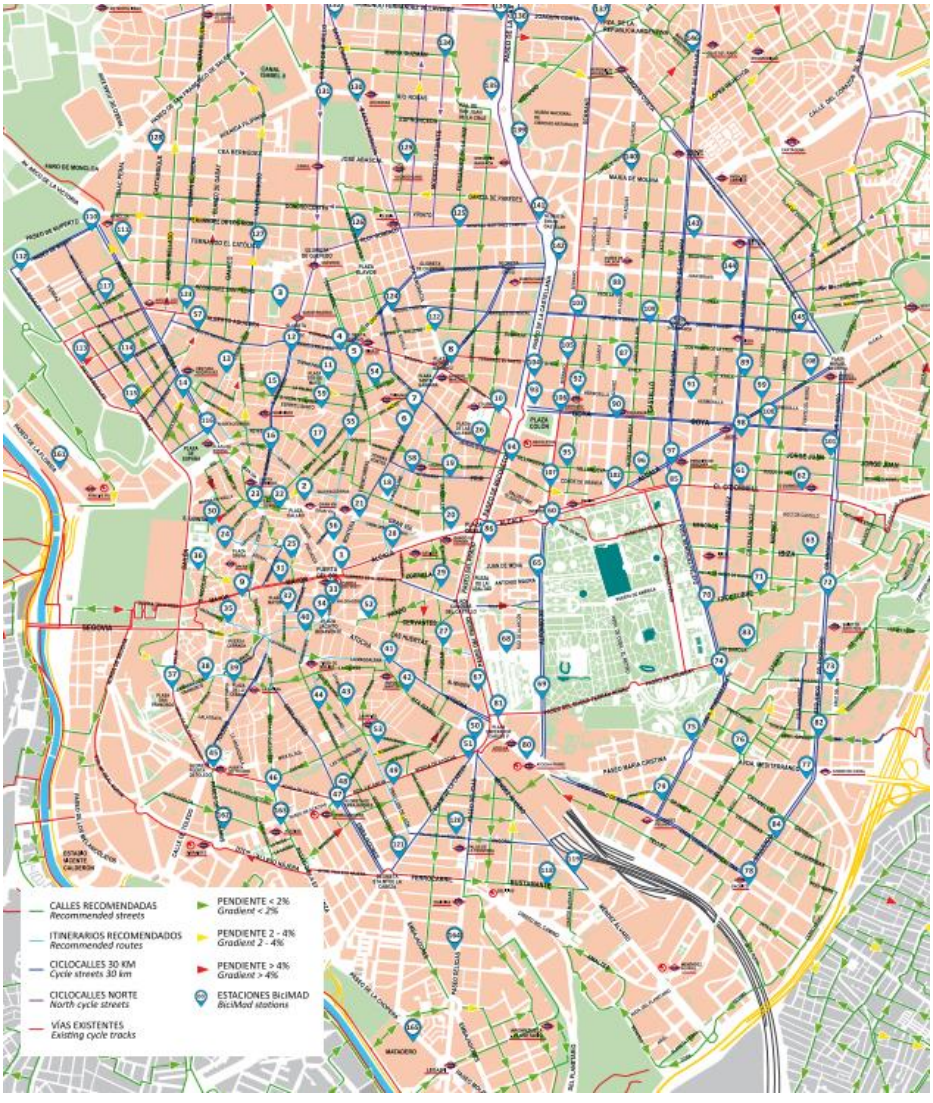
De la misma forma que en el problema 1 estudiamos y constatamos que grupos de edad son los que más representan a un usuario de bicimad y obtenemos el siguiente resultado:
Los usuarios que más utilizan bicimad son aquellos que tienen entre 27 y 40 años seguidos del grupo 5 (entre 41 y 65 años) y del grupo 3 (entre 19 y 26 años).

PROBLEMA 3

Para los siguientes problemas de usuarios, puesto que hemos concluido en el problema 1 que el tipo de usuarios en su mayoría corresponden con el tipo 1, nos quedamos con los datos relativos a estos usuarios en Enero de 2019.

Estudiamos los viajes que realiza cada usuario en función del origen y el destino.

1. Queremos saber los viajes realizados con origen en la estación 5 que corresponde con la calle Fuencarral. Obtenemos 1146 viajes.
   Con estación de origen 34 (plaza Jacinto Benavente) obtenemos 858.
2. Ahora obtenemos un listado con las estaciones y el número de viajes realizados con esa estación de origen y observamos que la estación 163 (Acacias) ha sido la que más se ha utilizado con este fin.
3. Ofrecemos el código necesario para conmocer cuantos viajes se han realizado entre una estación de origen y otra de destino. Por ejemplo:
   Viajes realizados con origen en la estación 5 y destino estación 2 0
   Viajes realizados con origen en la estación 82 y destino estación 83.....17



RECORRIDO POR LOS USUARIOS

De los usuarios que han realizado dos viajes en un día queremos saber cuántos de ellos han realizado el segundo viaje desde la estación de destino del primer viaje.
Obtenemos 30638.

De los usuarios que han realizado 4 viajes en un día queremos saber cuantos de ellos han realizado al menos un viaje de ida y vuelta.

Obtenemos 2671.

PROBLEMA 4

Queremos saber la probabilidad de encontrar bicis en una estación a una cierta hora. Para ello marcamos inicialmente el origen y la hora deseada.
En este caso hemos escogido como origen la estación 57 y hora 3:00:00. Obtenemos que a las 2:50 había 12 bicis ancladas en la estación deseada ('dock_bikes'). Por tanto, es muy probable que haya bicis disponibles para realizar nuestro viaje.

ANÁLISIS DEL TIEMPO DE LOS VIAJES

Para finalizar plantamos el código necesario para estudiar los viajes en función del tiempo. Por ejemplo, obtenemos que 13859 viajes que han durado más de media hora y 5038 más de una hora. Dentro de los que han durado más de una hora observamos el recorrido de algunos de ellos por ejemplo:

```
Origen: 49 Destino: 12 Tiempo (s): 6365
Origen: 136 Destino: 50 Tiempo (s): 9860
Origen: 164  Destino: 84 Tiempo (s): 8581
```