Spark programming

Author: Alassane Watt

Common friends in a social network

Exercice 1

Execute the following command:
`spark-submit --master spark://sar01:7077 sn_tiny.csv`

Exercice 2

Execute the following command:
`spark-submit --master spark://sar01:7077 filename` for each `filename` in the problem statement.

Results

Dataset	Nb nodes	Nb links	Nb pairs	Execution time (sec)
sn_tiny.csv			10	5.4
sn_10k_100k.csv	10^4	10^5	756	11.5
sn_100k_100k.csv	10^5	10^5	50	5.8
sn_1k_100k.csv	10^3	10^5	29826	185.1
sn_1m_1m.csv	10^6	10^6	60	13.1

We notice that the execution time grows with the ratio N_{links}/N_{nodes}. It also grows as N_{links} grows.

Exercice 3

1) The script is written in **friends_stats.py**.

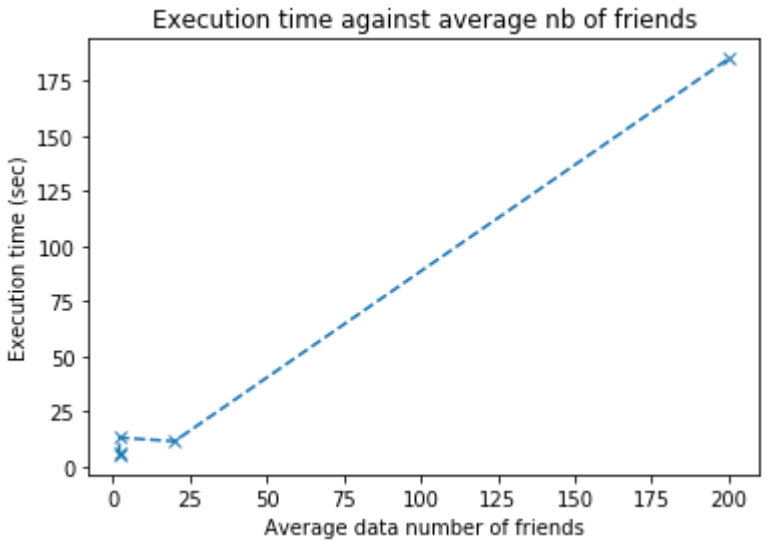
2) Execute `spark-submit --master spark://sar01:7077 friends_stats.py filename` for each `filename` in the problem statement.

3) **Results**

Dataset	Min	Max	Average
sn_tiny.csv	2	5	2
sn_10k_100k.csv	6	42	20
sn_100k_100k.csv	0	11	2
sn_1k_100k.csv	158	248	200
sn_1m_1m.csv	0	12	2

In [210]:

```
import matplotlib.pyplot as plt
data = [(2, 5.4), (20, 11.5), (2, 5.8), (200, 185.1), (2, 13.1)]
data = sorted(data, key=lambda x:x[0])
plt.plot([x[0] for x in data], [y[1] for y in data], "--x")
plt.xlabel("Average data number of friends")
plt.ylabel("Execution time (sec)")
plt.title("Execution time against average nb of friends")
plt.show()
```



We see that globally as the average nb of friends grows the execution time grows. It is not very highlighted due to the lack of points in the plot. This confirms our previous observation in exercise 2.

Creating an inverted index

Exercice 4

Execute the following command:

```
spark-submit --master spark://sar01:7077 inverted_index.py
```

The ouput shows that there are many words appearing only in one article.