LAB2. K-MEANS PARALLELIZATION in PYTHON

Creation of the datasets

In the lab material you would find a file named “proteins-generator.py”. You have to use it to generate computers datasets for the lab. To generate a data set, execute the command:

$>python computers-generator.py numrows

Being “numrows” a parameter specifying the number of protein chains in the dataset.

For development:
$>python computers-generator.py 5000

To test performance of the solution to deliver:

$>python computers-generator.py 500000

The file "computers.csv", created include a data set about a list of computers, including the following information per computer:

```
.write(str(price))
  f.write(",")
  f.write(str(speed))
  f.write(",")
  f.write(str(hd))
  f.write(",")
  f.write(str(ram))
  f.write(",")
  f.write(str(screen))
  f.write(",")
  f.write(str(cores))
  f.write(",")
  f.write(cd)
  f.write(",")
  f.write(laptop)
  f.write(",")
  f.write(str(trend))
```

```
  f.write("\n")
```

"id",”price”,"speed","hd","ram","screen",”cores”,"cd","laptop", “trend"

IMPORTANT: Do not modify, touch the file or create transformed fields. For the lab delivery extra files will not be accepted. We will use the same command to generate the dataset.

Notice: In the data you have 1 field that are numerical.

cd

As iy has only two values, you can substitute them with 0 (no) and 1 (yes) to normalize de data.

Laboratory Description

You are asked to extract useful information the computer data set implementing a program using the k-means algorithm in Python.

Use the path “computers.csv”. for the file. Do not include the full path in your computer.

Part one – Python serial

1.- Construct the elbow graph and find the optimal clusters number (k).

OPTION A

2.- Implement the k-means algorithm

3.- Cluster the data using the optimum value using k-means.

OPTION B

3.- Cluster the data using the optimum value using k-means with an existing function..

4.-Measure time

5.- Plot the results of the elbow graph.

6.- Plot the first 2 dimensions of the clusters

7.- Find the cluster with the highest average price and print it.

6.- Print a heat map using the values of the clusters centroids.

Part two – Python parallel, multiprocessing

1.- Write a parallel version of you program using multiprocessing

2.- Measure the time and optimize the program to get the fastest version you can.

3.- Plot the first 2 dimensions of the clusters

4- Find the cluster with the highest average price and print it.

5.- Print a heat map using the values of the clusters centroids.

Part three – Python parallel, threading

1.- Write a parallel version of you program using threads

2.- Measure the time and optimize the program to get the fastest version you can.
.- Plot the first 2 dimensions of the clusters

4- Find the cluster with the highest average price and print it.

5.- Print a heat map using the values of the clusters centroids.

Part four

10.- Write a memory explaining your results (maximum 12 pages)

Laboratory Delivery

uc3m | Universidad Carlos III de Madrid

Master in Big Data

TECHNOLOGICAL FUNDAMENTALS IN THE BIG DATA WORLD

Maximum group: 3 people.

Do not use Jupyter notebook

You have to deliver a compressed file named: "yournia_computers_2022.zip" including:

- Report with the memory (include author names)
- Three Python programs with serial and parallel versions of the program with multiprocessing. Names:
  - Computer-serial.py
  - Computer-mp.py
  - Computer-th.py

Delivery date: October 23rd[rd] 2022. 23:30 hours.

Scoring

OPTION A - Maximum score. 10
OPTION B - Maximum score. 7