Big Data Processing: Assignment 1

Marks: 10

Deadline: 31.01.24, 11.55 pm IST

You are given a file with item ids and their fixed dimensional vectors. Each line represents one item. The first column of each line is the item id and remaining part is the vector for the item. Given a query vector (say q) of the same dimension, your goal is to find out the most similar k items to the query item q based on cosine similarity. For developing your code, you can take a few vectors from the data file as query vector and test on it.

For the above problem statement, you need to write a multi-threaded priority queue in python. The output of your program would be a space separated two column list printed on the terminal, where the first column is the item id and the second column is the similarity score in descending order of similarity.

Make sure you use python version 3.10 or newer.

Path to data file: The data file is attached in moodle.

We will evaluate your program on a linux system from command line with the arguments as follows:

```
python <your-code.py>  <data file>  <query item file>   <# threads>  <value of k>
```

The above format is very important for evaluation. Thus, your program arguments must follow the sequence.

Submission guidelines:

You need to submit the program as a single python file in moodle. The file name must follow the format: **assignment-1-roll.py** (where the roll denotes your roll number in capital letters that must match exactly with your IITKGP roll number). Please note that if you fail to follow the format, your program may not be evaluated at all.

Important notes:

1. No credit will be given if your program does not run and produces wrong output.
2. No credit will be given if your program in not multithreaded
3. No submission will be accepted after deadline.
4. It is your responsibility to check that the file has been submitted successfully.
5. Plagiarism from friend or from web will invite negative (**-10)** marks.