Big Data Processing: Assignment 1

Marks: 10

Deadline: 10.02.23, 11.55 pm IST

In this assignment you have to write a multi-threaded python program for the following problem. ***Make sure you use python version 3.10 or newer.***

You are given a text document collection (in plain text format) along with the class labels. The documents in a folder corresponds to a particular class. The goal is to produce top k (k in an integer) unique word n-gram from the collection based on their class salience score. A word n-gram is a consecutive sequence of n words that appear in a document. The class salience score of a n-gram is defined as $\frac{count\ of\ the\ n-gram\ in\ a\ class}{\#\ documents\ in\ the\ class}$**.** Thus, if there are 20 classes, and a particular n-gram appears in all the classes, then the n-gram will have 20 scores (one for each class). The top k will be strictly based on descending order of score of the n-grams.

Tokenization rule (breaking documents into words): You must generate words from a document by breaking it on any non-alphanumeric character. You must also lowercase all the words.

Link to data: https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/20_newsgroups.tar.gz

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

python <your-code.py> <path to data directory> <# threads> <value of n for n-gram> <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 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.