Big Data Processing: Assignment 3
Marks: 10
Deadline: 15.04.23, 11.55 pm IST

Word Association using Spark:

You will be given a file with text documents, where each line corresponds to one document. For a given word (say W), the goal is to find:

1) Find top k positively associated word to W.
2) Find top k negatively associated word to W.

The association is computed based on word co-occurrence in documents using pointwise mutual information (PMI) scores. A word must not contain anything other than English letters. While computing co-occurrence, you must lowercase all the words and you must also remove the stopwords available here: https://github.com/terrier-org/terrier-desktop/blob/master/share/stopword-list.txt

$$\text{PMI(w1, w2)} = \log_2 \frac{P(w1,w2)}{P(w1)*P(w2)}$$

where P(w1,w2) = co/N, P(w) = m/N
co -> # documents where two words appear
m -> # documents where w present
N -> # documents

You goal is to write spark program for the above problem. You can use either scala or pyspark. Your code must have the main function.

Output format: Output needs to be printed on screen. First the list of positively associated words along with the PMI score. Then the list of negatively associated words along with the PMI scores.

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

```
spark-submit <your-code>  <path to file> <query-word> <k>
```

Where "query-word" is the given word, k is the top positively associated and negatively associated words to "query-word". This format is very important for evaluation. Thus, your program arguments must follow the sequence. Your program must have a main function.

Submission guidelines:

You need to submit the program as a single python/scala file in moodle. The file name must follow the format: **assignment-3-roll.py/assignment-3-roll.scala** (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.

Important notes:

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