CPSC 4330 Big Data Analytics
Winter 2025

Homework 2

Due: 10:55 am, Friday, Feb. 7

In this assignment, you will compute TFIDF for the terms(words) in a document corpus. The document corpus is a set of 30 text files that I have downloaded from https://www.gutenberg.org/. They are in the zipped file (ebooks.zip) on Canvas. You can download the zip file from Canvas, unzip it, and use all the 30 text files as the input of your program.

Write a Java MapReduce program to compute TFIDF for every (term, document) pair in the document corpus. As discussed in class, the TFIDF computation consists of multiple MapReduce jobs. You can review the slides on TFIDF to understand the algorithm of computing TFIDF. Please note that in class discussion, there are four jobs. You can also merge job 3 and job 4 into one job to complete the computation using only three jobs.

Your program should meet the following requirements.

(1) Terms are case-insensitive. For example, "Above" and "above" are the same term.

(2) Remove all the terms that are empty (i.e. "").

(3) A term is a string that contains only alphabets. For example, "f2f" is not a valid term and should be removed in the computation.

(4) For the precision of TFIDF, leave 8 digits after the decimal point (e.g. 0.00000609).

(5) Natural logarithm (base e) is used in computing TF-IDF.

Also, for this problem, you don't need to run your program on AWS.

See below a few lines of sample output (the file names in your output may be in a different order). Please note that I only put the TFIDF of the term "abandon" below to help you understand the format of the output.

You may find that in your output the TFIDF values of some words (e.g. "a" "about" "they" "have", etc.) are 0. Those words are stop words that appear very commonly across documents.

Submission

Use Canvas to submit all your source code files (.java).

```
....... //skip some data (this line is not part of the sample output)

abandon@408-0.txt    0.00003425

abandon@1232-0.txt   0.00003157

abandon@98-0.txt     0.00000591

abandon@pg64317.txt  0.00001575

abandon@84-0.txt     0.00002133

abandon@2554-0.txt   0.00001185

abandon@2701-0.txt   0.00001132

abandon@pg174.txt    0.00001002

abandon@4300-0.txt   0.00000309

abandon@345-0.txt    0.00000503

abandon@2600-0.txt   0.00003565

abandon@1260-0.txt   0.00001303

abandon@135-0.txt    0.00001739

....... //skip some data (this line is not part of the sample output)
```