CSCE 42703-001: Assignment 1

Due 11:59pm Thursday, February 13, 2024

1 Letter Count

Write a Spark program which outputs the number of words that start with each letter. This means that for every letter we want to count the total number of (non-unique) words that start with that letter. In your implementation please ignore the letter case, i.e., consider all words as lower case. You can ignore all non-alphabetic characters.

Run your program over the input data "pg100.txt".

What to submit:

Submit the printout of the output file and the source code (.java or .py files).

2 Matrix Multiplication

Following the two-step method introduced in the class, write a Spark program to compute computing $P = M \cdot N$, where $M = \{m_{i,j}\}$ is a 1000×1000 matrix, $N = \{n_{j,k}\}$ is a 1000×1000 matrix, and $P = \{p_{i,k}\}$ is a 1000×1000 matrix such that

$$p_{i,k} = \sum_{j=0}^{999} m_{i,j} \cdot n_{j,k}.$$

The matrix M is stored in "M.txt" where each element is stored in the form of "$i, j, m_{i,j}$". Similarly matrix N is stored in "N.txt" where each element is stored in the form of "$j, k, n_{j,k}$". Only non-zero elements are stored in these files. Similarly, your program only needs to output non-zero elements in P.

Hint: You may use the **join**(*otherDataset*) transformation in Spark.

What to submit:

Submit the source code (.java or .py files) and the values of following elements in P: $p_{826,506}, p_{551,373}, p_{627,406}, p_{21,672}, p_{42,142}$ (note that zero elements are not stored).