\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* COEN 242 Programming Assignment 3 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Group Members

---------------------------------------------------------

Shweta Kharat

Viashnavi

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Problem Statement

---------------------------------------------------------

TEXT MINING USING SPARK

Queries:

1. Determine the total number of minor revision available in the data set.

2. List out the page title and page id of all the pages that have at most five URL links

mentioned in their text field.

3. List out all the contributor with more than one contribution along with the revision-id. The

list of revision-id must be sorted in the descending order of timestamp.

Compile and Execute Instructions

---------------------------------------------------------

Below are the java files :

Solution 1:

Main.java - source code

Query1Sql.java

Query1Dataset.java

Solution 2:

Main.java - source code

Solution 3:

Query3Sql.java

Query3Dataset,java

Note:

---------------------------------------------------------

1. Change the path of the input file in java files

2. Compile and run java files to get the output

3. Output file will be saved at specified location

Output

---------------------------------------------------------

Problem 1:

Number of minor revisions: 3346414

Execution time: 0:3:21

Query3Sql.java Execution time: 25 minutes

Query1Dataset.java Execution time: 24 minutes

Problem 2:

Number of pages with at most 5 url links: 6861000

Execution time: 0:27:41