Fundamental of BIG DATA ANALYTICS

ASSIGNMENT 3

SPRING 2023

Due Date: 12th May 2023 (Submit Code file online on google classroom)

Instructions:

- *The name of the file should be your rollnumber-Question number*
- *Do not copy the work of your peers. In case cheating is detected, then your case will be referred to DC.*

Question 1: (10 marks)

Write a PySPARK code that inputs a text file and performs the following tasks

a) Count and print the number of **three long consecutive words** in a sentence that starts with the same English alphabet. We say that a word is long if it is greater than four alphabets.
b) Print the number of three long consecutive words starting with each alphabet.
c) Print the number of times each combination of three long consecutive words occur in the file.

Input :

Horrid Henry's hound hunts in the massive Murree mountains. While silly stupid Samuel's dark dreadful dragon likes to hunt in skies. Horrid Henry's hound and Samuel's dreadful dragon Dany are fast friends and like to hunt and play together. They call themselves fantastic fanciful foursome.

Output:

Total count 7
H => 3
M=> 1
S => 1
D => 1
F => 1

Horrid Henry's hound =>2
Henry's hound hunts => 1
massive Murree mountains =>1
silly stupid Samuel's =>1
dreadful dragon Dany=>1
fantastic fanciful foursome =>1

Question 2: (10 marks)

Write a PySpark program to input two files, Input.txt and Reference.txt. The Input.txt is enormous and contains results of diagnosis of various medical tests of different patients. The results are in binary, where 1 indicates yes and 0 indicates No.

Input.txt **Format: PatientID followed by T1, T2, T3 … test results in binary** **1:1 0 1 0 1 1 1 1 0 1 0** **2:1 0 1 0 1 1 1 1 0 0 0** **3:1 0 0 0 0 1 1 1 0 1 0**

4:1 1 1 1 1 1 1 1 0 1 1

The Reference file is a small file and consists of the test results of a few people. These people serve as a reference point.

Reference.txt
Format: ReferenceID followed by T1, T2, T3 … test results in binary

R1:1 0 0 0 1 1 1 1 0 0 0
R2:1 1 1 1 1 1 1 1 1 1 1

For each patient in the Input file, we wish to find the closest reference in the Reference file. To determine the distance between two binary sets of variables, we can use Simple Matching Coefficient or Jaccard Formula. Search for details about these formulas online.

Output

1 => R1
2=> R1
3=> R1
4=> R2

Question 3: List of Co-authors between any two co-authors (10 marks)

Write a Pyspark code that gets a list of co-authors of each author in the given input file and finds common co-authors between **authors**. Also, print the count of the common co-authors and the list of co-authors.

Input (Author -> List of Co-authors)

Y. Lu -> B. Cao, C. Rego, F. Glover, K. Kiani
B. Cao -> C. Rego, F. Glover, K. Kiani, Y. Lu
C. Rego -> B. Cao, Y. Lu
F. Glover -> B. Cao, Y. Lu
B. Hosseini -> K. Kiani
K. Kiani -> B. Hosseini, B. Cao, Y. Lu

Output pair of author -> (count) common co-authors

B. Cao, Y. Lu -> (3) C. Rego, F. Glover, K. Kiani
C. Rego, Y. Lu -> (1) B. Cao
F. Glover, Y. Lu -> (1) B. Cao
B. Cao, C. Rego -> (1) Y. Lu
B. Cao, F. Glover -> (1) Y. Lu
C. Rego, F. Glover -> (2) B. Cao, Y. Lu