Parallel and Distributed Computing

Assignment – 03 (ALL PDC SECTIONS)

Deadline: Thursday 10th June 2021 by: 11:59pm (not extendable)

Max points: 5 Absolutes

Instructions:

1. You MUST use JAVA.

2. Your programs must compile without warnings and execute correctly for full credit. Use good programming style, including the use of an appropriate amount of comments and suitable naming conventions.

3. In submitting this work, you are agreeing that it can be checked for plagiarism. Any plagiarized submission will be dealt with course plagiarism policy (zero marks for all current and future assignments) as announced and mentioned in the course outline.

4. You MUST submit the screenshots of code execution. All the outputs (screenshots) should include your name and roll number. All the screenshots must be included in a single zip file. The name of the zip file should be as follows (this carries weight): PDC_Spring2021_Assign3_[YOUR_ROLLNO]_ [YOUR_SECTION].zip

5. Late submission (even 1 minute) means No-Submission. To avoid any potential technical issue at the submission day, submit it as soon as possible.

Problem Statement

Stack Exchange is a network of question-and-answer websites on topics in diverse fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process. The reputation system allows the sites to be self-moderating.

The data files needed are attached.

You can find the metadata information here. You can also google if you have any confusions.

https://ia800107.us.archive.org/27/items/stackexchange/readme.txt

Create a **MapReduce** program for the questions below. You can run your code on standalone (local) Hadoop installation. The process is explained in the slides in file Hadoop Introduction. A lecture with standalone demo video is also uploaded.

1. Find the year in which most posts were created in Posts.xml file.

2. Find the average length of comments in Comments.xml file.

3. Given a keyword, identify the comments about the keyword to be either positive or negative collectively. This can be accomplished by first looking up each word in the comment (about the keyword) in the bag of positive and negative words i.e., positive.txt and negative.txt files respectively. The process is repeated for all the comments and collectively a positive or negative score can be used to classify the keyword to either used positively or negatively. Use Comments.xml file.

Example:

Comment 1: iBooks is free, Good Reader is inexpensive

Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 2: iBooks is a average alternative.

Comment Sentiment: Negative as most words are found in the bag of negative words.

Comment 3: I've tried downloading iBooks but it hangs a lot.

Comment Sentiment: Negative as most words are found in the bag of negative words.

Keyword: iBooks used negatively overall.