Graded Lab Assignment 4: Programming in R with text input datasets

This graded lab assignment is designed to be a capstone assignment for you to use your learning from the non-graded assignment on R towards reading, manipulating and performing visualizations on **real-world unstructured/text datasets** using R. Remember that more than 80% of the data in today's world is actually unstructured data; hence it is important for you to be capable of mining text datasets using languages such as R.

Required:

Download and use any non-copyrighted and free text/PDF file from the public domain as your input text dataset. For example, it could be a text file of an e-book, any PDF file, or any webpage (Don't download any copyrighted material).

For your input dataset, you can think of all the words in a given paragraph, page, section, chapter etc. as constituting a transaction. You can decide the granularity of a transaction in this case.

Determine which keywords represent frequent itemsets in your input file and plot a neat visualization to indicate the same.

Remember that you need to handle memory requirements and other scalability aspects as your input text file size keeps increasing.

Deliverables:

- You need to upload your completed codes on Blackboard by **April 28, 2017 at 3 pm**. This is a hard deadline and will NOT be extended.
- You are also required to provide a demo of your program.
- You need to write a document (12 point Times New Roman font, 1.5 line spacing, maximum of 3 pages) to explain the results of your visualizations in a concise manner. ***In essence, you should use this document to summarize your interpretation of the results.*** You should also submit this document (in PDF format only) via Blackboard.
- Please note that this assignment is a subset of your non-graded lab assignment. Hence, if you have already completed the non-graded lab assignment adequately, you can use the same codes as those of your non-graded lab assignment with modifications for any new visualization that you plan to do for this assignment.

Please note the following points:

1. This lab assignment will contribute to **10% of your grades** for the course.

2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program file. All programs should be submitted on Blackboard on or before the stated deadline. Please note that your programs may be subjected to plagiarism checks.

3. The **deadline** for assignment submission is **April 28, 2017 at 3 pm**.

4. This is a **HARD deadline** and no points will be awarded for the assignment if you submit after the deadline, unless there are extenuating circumstances.

5. The grading criteria for this assignment will be based on effort, adherence to learning points from your previous non-graded assignment, code quality, visualization, results and scalability.

6. Any act of plagiarism will result in a zero for the entire assignment. Hence, please avoid any form of plagiarism.

7. This is a group assignment, hence please do NOT collaborate with your fellow students in other groups towards the completion of this assignment. However, you are obviously required to collaborate effectively with members of your own group to ensure that you are able to function as a team player.

8. If you encounter any problems with downloading the datasets, please contact your TA (Saurabh Mishra).

This assignment carries 10 points.