Non-Graded Lab Assignment Set 1: Programming in R

This document contains several small programming assignments for you to practise programming in R.

For doing this lab assignment, you may want to obtain a textbook on R from the library. A good example of an e-book on R (available in the public domain) is as follows: https://www.cs.upc.edu/~robert/teaching/estadistica/rprogramming.pdf

At the end of this document, please refer to the resources for learning R.

During the lab classes, your TA and I will address your doubts in case you encounter any problems with these programming assignments in R. If you have any doubts, please DO NOT hesitate to ask your TA or me either during lab classes or during consultation hours.

Please note that this is a non-graded assignment; the purpose is for you to learn. Your learning from this non-graded assignment will be assessed from the next assignment, which will be a graded assignment that will serve as a capstone.

1. Create a table with 1000 rows and 3 attributes: <StudentID>, <DepartmentID>, <CGPA>

A sample table is shown below.

StudentID	DepartmentID	<CGPA>
11041	14	8.5
11042	13	7.8

To populate the values of the attributes in this table, you can generate random numbers using any programming language of your choice.

After you have completed populating the table with the attribute values, process the following sample queries: (For reading the values from the table, use read.table, read.csv etc. Learn to use arguments such as the colClasses argument. You can also practise reading using some of the packages in R).

(a) What percentage of the students have a CGPA above 7.5?

(b) What percentage of the students have a CGPA between 7.5 and 8.5?

(c) What is the mean CGPA across all the students in a given Department? (Here, your query should have any one DepartmentID, say d)

(d) Now randomly delete some of the CGPA values for the students in your selected Department (d). And then compute the mean CGPA across all the students in a given Department?

(e) Now increase the number of your table to 100,000 rows with the same 3 attributes as above. Repeat (a) to (d), but this time you need to handle the memory consumption issue as well. Here, 100,000 rows is just an example. If your system does not scale up to 100,000 rows, just use a lower number of rows. The objective of this question is to facilitate you towards understanding scalability issues w.r.t R. In general, when you wish to use R with larger datasets, it is preferable for you to have good knowledge of your system, including (but not limited to) memory requirements.

(f) Create a visualization for the distribution of student CGPAs in your data. There is no ONE correct answer to this kind of visualization, hence try out some ways of visualizing the data and then select any ONE of the ways for data visualization with appropriate justification.

2. (a) Practise using data frames in R for storing tabular data. In particular, use the dplyr and other packages, and functions such as data.frame(),data.matrix() , read.table( ), read.csv(). etc. Moreover, go through the readr package in detail.

(b) Practise using functions such as lapply( ), sapply( ), tapply( ), mapply( ), split( ) etc.

3. Use any text file as your input e.g., it could be a text file of an e-book, any PDF file, or any webpage (Don't download any copyrighted material). Practise using functions such as readLines on this input text file.

(a) Select any random word from your input file and print its number of occurrences.

(b) Which words in your input file occur more than 10 times? Here, 10 is just an example of a threshold parameter; you can use any other appropriate value for the threshold parameter depending upon your input text file.

(c) Now repeat (a)-(b) using a much larger-sized text file. Remember that you need to handle memory requirements and other scalability aspects as your input text file size keeps increasing.

(d) Which keywords represent frequent itemsets in your input file? Here, you need to define support thresholds etc., and also decide upon the granularity issues (line, paragraph, page etc. of the document). In other words, in the terminology of association rule mining, is every line regarded as a separate transaction, or is every paragraph regarded as a separate transaction and so on?

(e) Plot a distribution of the keywords occurring in your input text file in a neat visualization format.

(f) Create an effective visualization for the frequent itemsets of keywords in your input text file.

Deliverable: You need to upload your completed codes on Blackboard by April 4, 2017 at 3 pm. It is NOT mandatory to show any demos for this lab assignment since it is a non-graded assignment. However, if you have any doubts, you are HIGHLY encouraged to ask.

Please note the following points:

1. This lab assignment is non-graded; it is for your learning only.

2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program files. All programs should be submitted on Blackboard on or before the stated deadline shown below.

3. The **deadline** for assignment submission is **April 4, 2017 at 3 pm**. This is a hard deadline and will NOT be extended.

Resources for learning R from a data science perspective

- https://www.datacamp.com/ (click on "Start learning R")
- https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf (Excellent FREE textbook; 157 pages, but worth reading if you have the time. Alternatively, you can use it as a reference book.)
- http://adv-r.had.co.nz/
- https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/

- https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/learning-path-r-data-science/
- https://www.kaggle.com/wiki/Tutorials (This site provides a very comprehensive list of tutorials in R as well as other aspects of data science.)