CSE487 Data-intensive Computing Spring 2014

Due date: 2/28/2014 by 11.59PM, online submission

Title: Problem solving and exploratory data analysis using R.

Introduction/Context: Data Science or data-intensive computing process involves a phase where preliminary understanding of the characteristics of the data is explored. This phase is commonly known as exploratory data analysis (EDA) [1, 2]. This is a very useful phase before launching onto a full scale big-data analytics that often requires extensive infrastructure support and complex or newer algorithms such as Hadoop-Mapreduce [3] on a large cluster. The algorithms and methods we will use in EDA are also highly applicable and useful to perform (run) statistical analysis on the outputs of big-data processing. In my experience with data, EDA is perfect complement to big-data analytics. EDA lets you gain intuition about the data in the early phases and it also offers you a systematic approach to choosing among the many possible results of big-data analytics.

Problem Statement: You will work on a fairly large (semi)structured data set (minimum size 1GB). (We will leave the unstructured to the next 2 projects). The domain of the data needs to be national (USA only) or global (entire world). Examples are USA consumer sentiment data [4] and World Health Organization (WHO) data sets [5]. The data you choose should have been collected for a period of time. The example in the book for NY Times is for a month. You will run statistics on this data to perform EDA using the process/steps listed in Chapter 2 of your text book. You are required to prepare a report/document based on the EDA. Also you should provide the provenance for the results/outcomes (graphs etc.) you have listed/documented in your report. Creativity: 10%, Report: 10%, Degn&Imp : 80%

Data Characteristics: By structured and semi-structured data we mean data stored in tables (files), databases (RDBMS) and spreadsheets (like Excel CSV files). Type of the data can be numerical and more. Minimally it has to be numerical. The entire data could be stored in several files/tables, and spread over multiple directories (depending on the data). (We will add more details based on questions from you.)

What to do?

1. Form your team one or two people. Make sure you tell us about the members of your team.
2. Read chapter 2 and understand EDA and the statistical approaches discussed there in.
3. Download R statistical analysis language development environment and learn its capabilities through online tutorials. RStudio is also a good tool to install.
4. Complete the example discussed in your text book in pages 36-44. Complete the questions in p.38 and solve them using R code. You will submit the R code that you developed for this example and at least two interesting screen shots of plots.
5. Also complete the RealDirect case study with a different data set (given) than the one discussed.
6. In a typical project assignment we give a data set and also the set of questions to work on. In a slight twist, for this project we are going to turn the tables on you. You will research, seek out, and select a suitable data set. If it turns out that data set does not yield any patterns on analysis or anything interesting, you may have to start from the beginning and choose another. Use the data set

analyzed in Chapter 2 as a model. (Multiple file and tables, each row representing a user, collected over one month, and so on.) As mentioned before an important criterion is that the data should have global and/or national implication. (While I don't mind people choosing data sets from regional and local, I am not very familiar with those. I heard that there great data sets available from the Great Lakes research in the CSEE dept.) The data set and statistics have to be unique for each team.

7. Make sure that the data is interesting for you to work on and will provide some equally interesting insights.
8. Minimally the data should have dates or some derivative of data, geo-locations or some derivative of geo-location. You can also add markers to your data sets to support certain types of analysis.
9. You can also source your data from multiple sources while this is NOT a requirement. You can assume that the data is clean and does not have to be cleaned. Choose your data set appropriately.
10. Now for the most critical aspect: What statistics will you run on the data collected? The statistics discussed in Chapter 2 is quite basic and you should be able to run summaries, plots and pdfs as discussed in Chapter 2. Perform these and make sure you record the outcome in your report.
11. Perform the analysis for a single data set (file). Extend it to multiple files/data sets. Visualize some metrics and distribution over time. Describe the outcome in your report.
12. More specific analysis we want to perform are: K-means clustering (we will discuss this in class) with various k values depending on your data set. K-means is an unsupervised learning algorithm that is really very useful in many situations. (In one of the previous offerings of the course we used MapReduce to do this [6]. You can try this MR version later and compare.) I expect the clustering to be done on "state in USA" or "Countries in the world" or some criteria so that it can satisfy the next requirement. K-means in R is discussed in your Data Science text.
13. The second requirement is in the reporting of the results from k-means. We want the clustering to be mapped on to a USA map or a world map or a map (even a virtual map: see your text book for samples) that pertains to your data set. You will have to use several "maps" packages of R.
14. You are required to prepare a detailed **report and documentation** that will allow us to repeat the experiment/analysis you have carried out and also provide the provenance for the results you have generated. Make sure your report and the files have the names of the team members.
15. Use **elegant directory structure** and naming conventions for directories and files to capture all the work for project 1 and then zip them into one zip file. Include a README (for repeatability)
16. You will submit the report, data, and the R code:
    submit_cse487 project1.zip or submit_cse587 project1.zip

References:

1. C. ONeil and R. Schutt. Doing Data Science. Orielly, 2013.
2. J. Tukey. Exploratory Data Analysis. Pearson, 1977.
3. Hadoop-mapreduce. http://hadoop.apache.org/, last viewed Jan 2014.
4. University of Michigan Consumer Sentiment Data. http://www.sca.isr.umich.edu/, last viewed Jan 2014.
5. World Health Organization (WHO), http://www.who.int/research/en/, Last viewed Jan 2014.
6. L.S. Gordon,MapReduce and K-means clustering, http://blog.data-miners.com/2008/02/mapreduce-and-k-means-clustering.html,last viewed Jan 2104.