CSE4/587 Data-intensive Computing Spring 2017

FINAL PROJECT: COMMUNICATING THE RESULTS OF DATA ANALYTICS: B. RAMAMURTHY

OVERVIEW:

The hands-on practical learning components of the course comprises two types of activities: labs covering one or two knowledge units (skills, competencies) of data-intensive computing and a single term project serving as a capstone covering the entire data pipeline. In the first half of the course we learned data analytics and quantitative reasoning using R language. In the second half we focused on big data approaches for data analytics with Hadoop MapReduce and Apache Spark. The finale or final project is a data analytics showcase that provides an ***interactive dashboard*** for communicating the results of a data problem that you already analyzed or a new data problem that you will analyze for this project.

In Lab1 we wrote a data client and very simple information server. In Lab2 we worked on data cleaning and data munging. In lab (Lab 3) we applied machine learning algorithms and statistical models to data with the ultimate goal of being able to predict the outcome for a certain query or classify data according to certain criteria. More specifically, we explored algorithms discussed in Chapter 3 of Doing Data Science textbook [1]: linear regression, k-nearest neighbors, k-means. In lab4, we are exploring approaches that deal with big data, especially text data, using the Google's MapReduce algorithm. In Lab 5 we will explore processing graph data using Spark. This final project will capture all these aspects in a dashboard, plus interaction for assessing some "what if" scenarios. It will build the "data product" step in the data science process described in [1].

GOALS:

Major goals of the lab4 are to:

1. **Identify** a problem with multi-variate data and different forms analysis resulting in charts.
2. **Discover patterns** by analyzing the data**.**
3. **Design a dashboard** with multiple charts that offer insights into the data**.**
4. **Build a data product by adding interaction** to the charts to allow a user to understand the behavior of the data by changing values of parameters.
5. **Write and publish** a report that describes the dashboard that you created.

OBJECTIVES:

The lab goals will be accomplished through these specific objectives based on this Tableau book [2, 3].

1. Knowing the goals of analysis, the target audience
2. Using the right data
3. Selecting the suitable visualizations
4. Designing for interaction

5. Deciding the medium (RShiny [4, 5, 6], d3,js [7, 8] etc) and channel (web, app, mobile etc.)

LAB DESCRIPTION:

Introduction: In this age of analytics, data science process plays a critical role for many organizations. Several organizations including the federal government (data.gov) have their data available to the public for various purposes. Social network applications such as Twitter and Facebook collect enormous amount of data contributed by their numerous and prolific user. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make subset of their data available for use by registered users for free. Some of them as downloadable data files (.csv, .xlsx) as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as a web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrap the data from these web pages and extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products. Very often data collected is not in the format required for the downstream processes such as EDA, analysis, prediction and visualization. The data needs to be cleaned, curated and munged before modeling and algorithmic processing.

Also we will follow the pedagogical pattern

- Preparation before lab (pre-lab)
- Learn from working on some of the solved problems on Tableau. (Somewhat like R Vignette).
- Apply the knowledge to meet the goals of this project to create a data analysis showcase.

PREPARATION: Here are the preliminary requirements for the lab. **Time needed: 2 to 3 hours (Day 1)**

1. Work environment: We will be downloading and exploring the Tableau environment [ref] to get some ideas about communicating data. If you want to work with other environment download and prepare that environment.
2. Download the data sets for Tableau learning from UBbox [9]: These are specified as Excel files. (yes, Tableau works well with MS Excel!) License will be sent to you.
3. Review the variables (rows and columns) of the Excel data files form NHLTop100 players and World Population and Income data files.
4. Also review the Introduction to Tableau provided in the lecture notes based on the Tableau reference book [2]. Also see the tutorial available at [3].

LAB 3: WHAT TO DO?

1. **(10 points) Data dashboard example. (Day 2, Time Needed: 3-4 hours)**
   An activity with Tableau is posted in UBbox [9]. This activity is similar to an R-vignette. In this case you learn about the features of Tableau software. Go through the instruction given and the work on them to complete the exercises. Submit the final product (dashboard, story book, worksheet etc.) on timberlake.

2. **(15 points) Story telling with your data (Day 3, Time needed: 4-5 hours)**
   Now repeat the steps with the data set you have selected to work with. Make sure the data as well as the charts are non-trivial such as pie charts. You have to include interaction to vary the parameters. The data and analysis should solve a real-world problem and not a synthetic problem. Clearly define the data schema and the scope of your analysis (what are doing?).
3. **(20 points) Implement a data product (Day 4,5: Time required: 4-5 hours each day)**
   This is the last module of the data science process described in Doing Data science text [1]. It could be a simple prediction problem, classification problem and should be dependent on the earlier data analysis and exploration in step 2. This could be in Tableau, RShiny, d3.js or any high level language you prefer.
4. **(5 points) Document the product developed (Day 6, 4-5 hours)**
   Choose a "channel" to make your data product available (through the web or mobile channels). Document the steps in using the product including data input format and expected output.

DUE DATE: 5/7/2017 BY 11.59PM. ONLINE SUBMISSION.

REFERENCES:

[1] C. ONeil and R. Schutt, Doing Data Science, ISBN:978-1-4493-5865-5. Oreilly Media, Doing Data Science, http://shop.oreilly.com/product/0636920028529.do, 2013.

[2] B. Jones. Communicating Data with Tableau: Designing, Developing, and Delivering Data Visualizations. O'reilly, first edition, 2014.

[3] Tableau Tutorial. https://www.tableau.com/learn/training, last viewed 2017.

[4] RShiny. RShiny bt RStudio. https://shiny.rstudio.com/, last viewed 2017.

[5] RShiny Gallery. https://shiny.rstudio.com/gallery/, last viewed 2017.

[6] RShiny Tutorial. https://shiny.rstudio.com/tutorial/, last viewed 2017.

[7] D3.js, https://d3js.org/, last viewed 2017.

[8] D3.js gallery. https://github.com/d3/d3/wiki/Gallery, last viewed 2017.

[9] UBBox link for Term Project data and instructions.
https://buffalo.box.com/s/jk2m74jloejv5t3f92r3no1quweq7we7, last viewed 2017.