

CSE4/587 DATA-INTENSIVE COMPUTING SPRING 2018

LAB1: DATA COLLECTION AND EXPLORATORY DATA ANALYSIS: B. RAMAMURTHY **DUE: 3/8/2019**

OVERVIEW:

The hands-on practical learning component of the course comprises two types of activities: learning from existing data explorations and labs covering one or two knowledge units (skills, competencies) of data-intensive computing. This document describes Lab1: Data Collection and exploratory data analysis (EDA) that replicates professional data analysis on a topic of current interest, and extending the analysis to include another public data source.

LEARNING OUTCOMES

- ✓ Work on data analysis related to socially relevant current topic. (What better than “flu”?)
- ✓ Learn by analysis of existing data analysis examples or vignettes. (Twitter vignette)
- ✓ Learn from data analysis and reporting examples available in public sources. (Flu.gov)
- ✓ Apply methods for collecting data from publicly available data sources: flu.gov, fluvew.
- ✓ Install a work environment for carrying out various activities of the data science process: Jupyter, R, RStudio.
- ✓ Extract data using APIs and OAuth keys (for collecting tweets).
- ✓ Process the data collected for simple data analysis and charting (reproducible research).

OBJECTIVES:

The lab goals will be accomplished through these specific objectives:

1. **Educate** ourselves about influenza or flu that is rampant this season in the USA. In general, be cognizant of what is going on in world around you.
2. **Install** Jupyter [1] notebook environment and within it R language [2] kernel, and RStudio [3]. See the instructions here [4].
3. **Familiarize** yourself with R language through Jupyter environment by running all the examples discussed in this handout [5]. Do not miss this step.
4. **Explore** the real flu data, and replicate and learn from the analysis performed by experts in Center for Disease Control (CDC) [11] and related organizations.
5. **Collect** data by querying Twitter REST API [6]. You will have to get a developer account on twitter and also get the credentials for your application (the twitter client) that you will be writing. Good query word related to “flu” gets you good data.
6. **Process** data using twitterR [7] library package of R.
7. **Visualize** geo spatial information extracted from the tweets using geo-map libraries of R: ggplot2, ggmap, maps, and maptools [8]. Maps and geo codes are supported by Google map API.
8. **Compare** CDC flu map with your own home-brewed flu map of the USA derived from the twitter data you obtained.

9. **(optional) Create** your own optional documentation updating the data-twitter-R-analysis by providing a cool vignette for some other domain than flu.

LAB DESCRIPTION:

Introduction: An important and critical phase of the data-science process is data collection. 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 curated, packages and sold as products.

What is an API? Why is this so important? A standard, secure and programmatic access to data is provided through an Application Programming Interface (API). An API offers a method for one or two way communication among software (as well as hardware) components as long as they carry the right credentials. These credentials for authentication for programmatic access is defined by another standard OAuth (Open Authentication) delegation protocol [6]. (There may be other approaches. We will use this approach.)

LAB 1: WHAT TO DO?

PAIR PROGRAMMING: We are going to allow pair programming for this lab. You will work in groups of one or two. **(No groups >2)**. You will get an F for the course if your group plagiarizes or copies somebody else's work or some other group's work. You can discuss anything ONLY with your pair team member. **Members in the pair have to work on the entire problem and submit your own notebooks, and submission. You cannot seek help from outside source or out-source your lab work to any online agents or personal in other organizations or universities or freelancers.**

Preparation: Here are the preliminary requirements for the lab.

1. **Work environment:** You will working on Jupyter with R kernel. Install Jupyter and R kernel as instructed by the handout [4]. This will be our "Learning Environment". Later on we will explore "Development Environment" in RStudio; Hear is this application development process depicted in figure 1. You can see why we use many tools: Jupyter, and R studio.

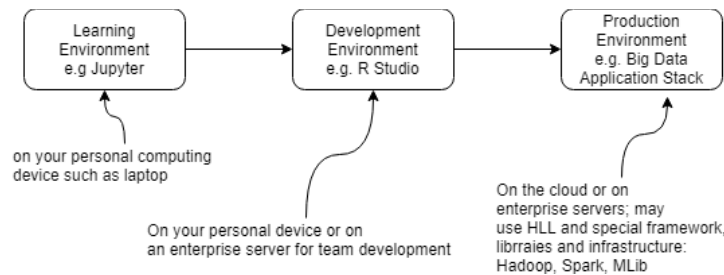


Figure 1 Development Process

2. Create an account on twitter as a user as well as a developer. In the developer site, the tab Apps is of particular interest. (After you create a twitter developer account,) you will click on Apps to create a new app called Lab1. Fill in the required fields **as per the instructions given there**. Once you submit you should be able to get the OAuth credentials [9] that had four parts: Customer API key, Customer API secret, Access Token Key and Access Token Secret. All are needed for programmatically working with Twitter. (Yes, you can auto-tweet, if you know what I mean ;-)
3. R community has created a package for working with Twitter data called “twitterR”. Read the vignette by Jeff Gentry [12] about the package he contributed. Work on the vignette.

Part 1: Complete the R-handout [5] discussed in lecture on Jupyter and submit the notebook generated. (10%) Due: 2/14

Part 2: Learning by repeating topical data analysis (50%) Due: 2/28

1. Navigate to the CDC site of flu data and analysis, flu.gov [11] and fluview [10]. Take a few minutes to review the contents and understand the context.
2. The fluview site discusses many variables or features that affect the reality that is playing out in front of our eyes.
3. Browse through the page, identify and review the charts and data (plain tables, .csv) under each chart; here are the ones with data.
 - 1) Influenza national summary (green and yellow chart)
 - 2) Positive tested
 - 3) Influenza sub-type pie-charts
 - 4) Mortality
 - 5) Pediatric deaths
 - 6) Influenza-like illness
 - 7) Flu heat map of USA (Required)

You can down all these data and more from this link [13].
4. Now repeat at least five of the seven flu charts discussed in the flu report for the week of Jan 28th, 2019. Beware, by the time we review it, it might have moved on to the next week and the data and the charts may be different. The last one, heat map is required for part 3 of your lab.
5. (5/50%) The chart 1) and 2) are provided only for the partial year. Aren’t we curious about the flu pattern for the entire year or 52 weeks? You can get this data appropriately selecting data from [13] and repeating the charting process using R (on Jupyter or R studio).

6. (5/50%) Now repeat this chart in 5) for just New York state. You can get this data also from the same source. Just explore around.

Part 3: Twitter Application Development (40%) Due: 3/8

We will develop applications that are “data clients” for twitter data. Twitter supports many APIs, we will use Search API that is a part of the REST API.

1. Learning Jupyter, R and twitterR: All these can be achieved by one activity: working with twitterR package library vignette. Type in the R language instructions for each example discussed: try it with different names and twitter hash tags. For our context we need query words to be associated with influenza or flu. Apply your domain knowledge or intuition to query for an effective search word to get most closely tracking tweets for “flu”.
2. We are NOT interested in sentiment analysis. We are interested in sheer number of tweets on a topic that is associated with “flu” or a related term that you uniquely determine that will be important influencer. You have to choose a good topic. Understand the Search API that we are using for can give you only limited number of tweets per day and also only a sampling of the all the tweets. You will collect at least 20000 tweets (How could we categorize them?). Group them by geo-location as in Google maps API (one more API) and plot them on the map of USA. Map the geolocations to states, and color the states according to the number of tweets or mentions per state. Also understand that geo-locations are approximate. Person tweeting may not be in the state at which the actual incident of the flu is happening. We don’t have to worry about this. We assume the location of the tweet and the occurrence of the tweet are in the same place.
3. Input: Search word or hash tag related to flu. Data client processing: Obtain and group tweets by location into categories mentioned in “fluvew”. Output: plot them on the states/ color the states by the strength.
Issue 1: Of course, there is an issue with location meta-data. This is not available (N/A) if the user does hide his/her location. Only 1% of the tweets have geolocations. Then how can we get “set of locations”?
Here is a verified approach using function of twitterR
 - a) Convert search result tweets into dataframe
 - b) Lookup screen names from this dataframe
 - c) Convert these screen names into another dataframe
 - d) Keep only users (user names) with location info
 - e) Get the geocode of the locations from this dataframe
 - f) Hints on TwitterR functions you may need: twListToDF, lookupUsers, geocode; Look up these functions in the twitterR manual.
4. Compare: Now compare your map and with map from CDC for the same date ranges. You can do that side by side in a Jupyter notebook by running the respective R commands.
5. Iterate: Try the comparison with different query words related to flu. Keep your words secret.
6. Bundle all the work in a Jupyter notebook for submission, even if you work on RStudio. Make sure your document your application development using markdowns.
7. (5/40%) Run R server and make your analysis available for the world to see and interact. Standby for more details.