CSE4/587 Data-intensive Computing Spring 2017

LAB2: DATA CLEANING AND MUNGING: 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. General pedagogical pattern for the labs is: one or two vignettes for learning the main concepts (or theme) of the lab, followed by 2 or 3 activities that apply the concepts.

In Lab1 we wrote a data client and very simple information server. In Lab2 we will focus on data cleaning and data munging. Data Cleaning refers to removal of unwanted data and adding additional data derived from raw data to allow easy and efficient processing. Data Munging is a process for converting raw data into a form that renders it convenient for consumption by downstream applications and processes.

GOALS:

Major goals of the lab2 are:

1. ***Convert*** data in various format into a form that is convenient for in-memory operations. Transform from external storage formats such as xml, sqllite into a common external format, comma separated value (.csv) convenient for exploratory data analysis using R. This allows for easy reading of data into the memory as data frames.
2. ***Extract*** (data munging) useful data from raw data collected by real survey instruments. You will use the actual survey document in interpreting the survey results.
3. ***Repurpose*** data from a popular domain (such as sports) for consumption by different genre of applications.
4. ***Transform*** data using operations such as grouping, categorization and binning to stage them for in-memory analysis. (R is optimized to work well with in-memory data.)
5. ***Document*** data cleaning steps using Markdown, a text-based HTML authoring format. This is an essential step in "reproducible research" and is offered within Jupyter platform.
6. ***Learn*** and understand the scientific data collection process, surveys, and nature of raw data and the need and motivation for cleaning and munging the data.

OBJECTIVES:

The lab goals will be accomplished through these specific objectives:

1. You will be working in R language environment: Jupyter or R Studio.
2. Experiment with and learn the tools and techniques needed for cleaning and munging in R. This is done implementing the steps detailed in an R package vignette.
3. Learn about some popular data sources and ***data environments*** such as Kaggle [5], Pew research [10] and Data.gov.

4. You will convert an XML file to csv, extract and repurpose data from sqllite database, transform raw data collected by Pew survey.
5. You will pre-process raw data from Pew research and perform exploratory data analysis (EDA) using the (i) survey responses, (ii) survey instruments and (iii) additional information about the survey. Survey based research is just one of many scientific approaches to data-based research methods [13].

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 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.

In Lab1 we acquired data from twitter using its REST API and processed it. In Lab2 we will prepare the data for (i) question answering (ii) change format to accommodate EDA (iii) understand data by plotting it, and then standardize and attempt to normalize data for support further analysis.

Also we will follow the pedagogical pattern

- Preparation before lab (pre-lab)
- One or more R vignette on a specific concept
- One or two simple activities related to core topic of the lab (in Lab2 it is Data Cleaning)
- Featured (major) activity (in Lab2 cleaning single-table survey data into standard and loosely normalized form).
- Documentation /narrative for the featured activity in Markdown web text format [7].

Preparation: Here are the preliminary requirements for the lab. Time needed: 1 or 2 hours (Day 1)

1. Work environment: You can work in our “Learning Environment” Jupyter [3] or on “Development Environment” in RStudio for R language [4]; (Production Environment will employ a robust programming language such as Java or C++).
2. Cleaning data also involves studying the “variables” (column headers in a table). Sometimes this requires (i) categorization, (ii) factoring and (iii) some basic plotting of variables. A fine example for understanding these concepts is the “titanic” vignette that analyses the survival of passengers

in the infamous titanic ship disaster. This data is analyzed by numerous people in so many different ways and is also featured as a data (and competition) on Kaggle [5]. We will use a particularly simple analysis that illustrates the use of qlpot and other useful plots for understanding data. Complete this vignette [6] on titanic data.

LAB 2: WHAT TO DO?

1. (5 points) R ***Vignette*** **(1 day: 2-3 hours)**
   R is essentially an in-memory processing language. Moreover in Lab1 you realized the importance of grouping data according to a certain rule (e.g, lat-long location categorization). Many devices such as mobile phones collect their data sqlite format. To address these needs we will work with a very useful R package in ***dplyr***. A vignette for dplyr is here [8]. Work through the commands discussed in the document and understand them. We will use it later. **Work on this before you work on other items in Lab2 and submit it as Jupyter notebook.**

2. (10 points) **Activity 1: Transform XML to data frame to CSV**. **(1 day: 2-3 hours)**
   XML has been a very popular approach to data markup for a long time. So we have large data collections in XML. We want to learn how to transform XML data to data frame for processing and persist the transformed data in to a csv file. R has a very nice package appropriately called *xml* that has functions to do this. Use the ***xml*** package to generate data from the xml file. We will use a simple data set [9] from Washington University to carry out this operation. Optionally you can also look at xml data files from data.gov. Complete this and submit it as Jupyter notebook. **Input**: WU data file in xml: **Output**: exactly equivalent file in CSV. **Process**: R and xml package.

3. (15 points) **Activity 2: Extract and Repurpose Data. (2 days: 2-4 hours each day)**
   Kaggle (kaggle.com) is an excellent environment for learning about modern data sets, data sources and data-science related activities and competitions. We will use one of the data sets posted European Soccer Data for several years. Data is stored in normalized sqlite tables, and the data is provided for several years. It is quite "clean". If you look at the interest of Kaggle users for this particular data set, it looks like most are interested in "predicting" the winners. The data contains mostly lot of technical details about the players and the teams. We want to extract one or two simple csv files with observations that are of interest to common public. One could be about players (player.csv) and the other could be about the teams (team.csv). Think about what you want to "interrogate", "learn" or "query" about European Soccer. These should drive your meta-data (observations) of data filtered out from the many tables of the original data. Think of this problem as "staging" the data for Q&A (Question and Answer) system such Amazon Alexa [14]. This activity is like creating views from a relational database only that we will be performing that using R and R packages.
   Input: Kagge Sqlite European Soccer Database **Output**: R data frames persisted in csv files to facilitate easy Q&A, **Process**: sqllite to R data frames to csv.
   Use the set of questions given below as guideline for developing the variables (meta data) for the data frames.

4. (20 points) **Featured Activity 3: Convert and Transform Raw Data (2 -3 days: 3-4 hours a day)**
Pew Research Center [10] has been collecting data for a long time about social issues using survey approach. Study their home page here and the types of questions they try to answer using the data sets collected [11] and analyzed using scientific methods. We are especially interested in look at the data PRC collected about Gaming, Jobs and Broadband. Clean and munge this data using the methods described during lecture, methods described in ***dplyr***, and your own intuition. Rename observation, rename data categories, remove unwanted observations, null rows, etc. Your guiding principle in these operations is the set of questions you are trying to answer using this data. What do you want to learn from this data? You will have document every change/transformation you make to the data for provenance purposes. Also be mindful that any modification to the raw data may result in "loss" of information and in some cases even may be unethical or illegal – please be aware of this. In other words, your documentation of what you did and justification are very critical in this process.

I. Here is some information about the data:
JUNE 10-JULY 12, 2015 – GAMING, JOBS AND BROADBAND

"This dataset contains questions about video games and gaming; job seeking and the internet; workforce automation; online dating; and home broadband, cable and smartphone use among Americans." The dataset has been downloaded by me and is available as a resource on UBox at [12]. You don't have download it from Pew Research Center. **Please understand this data is for use only in this course. Do not republish it anywhere else as yours.**

II. Download and unzip the file from UBox link given. Here is the list of files and contents.

File Name	Contents	Our Reference
June 10-July 12, 2015 – Gaming, Jobs and Broadband - csv	CSV raw data file	File1
June 10-July 12, 2015 – Gaming, Jobs and Broadband - Questionnaire	Survey questions	File2
June 10-July 12, 2015 – Gaming, Jobs and Broadband - Topline	Results per question	File3
June 10-July 12, 2015 – Gaming, Jobs and Broadband - Crosstab	Tabulated results	File4
June 10-July 12, 2015 – Gaming, Jobs and Broadband - sav	SPSS file	We will not use this file

III. Study the csv file (File 1). Both RStudio and Jupyter load the File1 well and you can see the 2001 observations of 140 variables. Study the variables. Understand the ***independent*** variables: very interesting ones for categorization and factorization: Gender, Income, Age, Race.

IV. Use File3 and File4 to select three or more topics in the data that interest you. For example, smart phone ownership or Internet usage or job seeking using smart phones. Show your creativity. DO NOT copy me or fellow students. We will refer to this list of your choice of topics as ***MyTopics***. **Document every step using Markdown web text format.**

V. Develop questions and hypothesis. How does smart phone ownership vary with Age? Income? I feel lower age group has higher ownership of smart phones and lower income has higher ownership of smart phones.

VI. For new data frames with only relevant variables. Group (dplyr), categorize, write functions to bulk process, plot (qplot, hist) to evaluate your hypothesis. Study the outcomes. We want to see plots to justify your cleaning and development of new data frames.

VII. Data cleaning is an iterative process. Based on the outcomes and observations above you will have to manipulate your newly created data frames further. Save the data frames as csv files.

5. **You have to work on your own. This is an individual lab. You will get an F for the course if you plagiarize or copy somebody else's work or share your work with somebody.**

DUE DATE: 3/5/2017 BY 11.59PM. ONLINE SUBMISSION.

REFERENCES:

[1] https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html

[2] https://www.tutorialspoint.com/r/r_xml_files.htm

[3] Jupyter. http://jupyter.org/, last viewed 2017.

[4] The R Language. https://cran.r-project.org/, last viewed 2017.

[5] Kaggle Data science Platform, https://www.kaggle.com/, last viewed 2017.

[6] Titanic Survivors vignette. http://rstudio-pubs-static.s3.amazonaws.com/25402_590f6acfb2544355883f8dcf1c441dc6.html, last viewed 2017.

[7] J. Gruber. Markdown http://daringfireball.net/projects/markdown/ last viewed 2017.

[8] R Vignette for dplyr: https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html, last viewed 2017.

[9] XML data from Washington University, http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/data/courses/reed.xml , last viewed 2017.

[10] Pew Research Center. http://www.pewresearch.org/, last viewed 2017.

[11] Pew Research Center Datasets, http://www.pewinternet.org/datasets/, last viewed 2017.

[12] Pew Research Data set for Lab2: https://buffalo.box.com/s/tjf7suaux5batnvo03pbrjom6josu0h7

[13] Hammond, Flora, Handbook for clinical research : design, statistics, and implementation Publisher: Demos Medical Publishing, LLC., ISBN:1-936287-54-4, 978-1-936287-54-3. 01/01/2014

[14] Amazon Alexa, https://developer.amazon.com/alexa-skills-kit, last viewed 2017.