CIND-110
Data Organization for Data Analysts

Lab Manual Module 10
Data Mining

Instructor: Dr. Tamer ABDOU

Contents

1. Objectives

In this lab we will focus on Data Preparation, Data Exploration and Data Mining of the data-sets drawn from the projects documented at:
https://github.com/ansymo/msr2013-bug_dataset/tree/master/data/v02

2. Installation

In this lab we will use Rstudio with latest version of R. Please follow the procedure given in Lab 9 to install Rstudio in your environment.

Please install jsonlite and curl packages if not installed already.

Install jsonlite and curl packages in R

```
install.packages("jsonlite")
install.packages("curl")
```

3. Load the Bug Status JSON file into R

- Download the dataMining.R file from the course shell to your environment.
- Open the file in Rstudio.
- Execute the steps in part-A of dataMining.R to load the bug status data to be mined.

4. Explore the data

- The data has been loaded into two lists, each consisting of thousands of data frames, for the bug tickets in the Eclipse project and the Mozilla project.
- Execute the steps in part-B of dataMining.R to view the status of a single ticket from a single project as an R data-frame.

Each data frame has the columns 'what', 'when', and 'who' for the bug ticket operation performed, when it happened as a Unix timestamp, and the userid of the person performing the operation, respectively.

5. Summarize the data

Follow some of the suggestions in part-C of dataMining.R to summarize the data and report descriptive statistics on the data.

- Find the total number of tickets in a bug list.
- Tabulate the number of operations performed in a bug list.
- Find the total number of operations performed across all tickets in a bug list.
- Find the number of unique users performing operations in a bug list.
- Draw a histogram of when new tickets in a bug list were created by year.
- Find the number of times tickets were re-opened
- Draw histogram of ticket time to resolution for tickets resolved once only.

Try your own approaches to mine statistical insights from the data.Be sure to run the queries on both bug lists and even draw some comparisons between the Mozilla project and the Eclipse project.