DATA2001: Data Science: Big Data and Data Diversity

Tutorial 2

Welcome to the Tutorial 2 of DATA2001. In this tutorial you will learn how to process and analyse delimited data in text files. Download the survey datasets from Canvas and refer to the lecture 2 slides to perform the following analysis.

Reading and accessing Student Surveys

Read the student survey data from the csv file using Dictionary data structure (key-value) pairs. Define some constants for each column name of the csv file. (E.g., 'submitted'=TIMESTAMP, 'Tutorial class'=CLASS_ID. Visualise the 9th, 15th, 31th and the 2nd last record of the dictionary (hint: follow the python codes in the lecture slides).

Data Cleaning

(a) Define a python method to clean the 'Tutorial class' column by analysing the texts in cells as follows:

SIT Lab 115 (2-4pm) = T14A

SIT Lab 116 (2-4pm) = T14B

SIT Lab 117 (12-2pm) = T12A

SIT Lab 114 (2-4pm) = T14C

All strings need to be cleaned and converted in such a way that this column contains only these four (T14A, T14B, T12A, T14C) codes.

(b) Define a python method that considers the following levels of Python experience and convert each level to its corresponding integer number as follows:

‘None’ = 0

‘Basic Understanding’ = 1

‘Written some simple Python programs’ = 2

‘Competent Python programmer (familiar with eg. functions and classes)’ = 4

‘Have written complex Python programs already’ = 5

(c) Define the cardinality for each respondent by considering responses in regards to their experience described in 'other programming languages competency'. For example, a response like 'C#, Java, Javascript/ECMAScript, Matlab' = 4, 'Matlab,R' = 2, ... likewise. (Hint: consider the number of comma separated values in each cell).

Define a new list with a name ('number of language competency') that will contain the converted cardinality values calculated above.

(d) Do the same thing for the 'Relational Database Competency' column (e.g., ‘Microsoft Access, Microsoft SQL Server, Oracle, PostgreSQL’ = 4, ‘MySQL, PostgreSQL’ = 2).

Define a new list as well and name it as 'number of SQL competency'.

(e) Similar to question (b) define integer levels for the column 'SQL Competency' as follows:

‘Never heard of SQL’ = 0

‘Heard of it, but never used it’ = 1

‘Can interpret some SQL statements’ = 2

‘Written some SQL queries already’ = 3

‘Can already create tables and write complex SQL queries’ = 4

Analysing Date/Time

Display the time component of the 'Submitted' Column by ignoring the date part for the first and last 10 records (rows).

E.g., 2018-03-13 23:41:57

Hour: 23

Minute: 41

Second: 57

Visualizations

(a) Visualize a frequency polygon of the degree 'year' of respondents. In this polygon, the x-axis will contain the degree years (1, 2, 3, 4) and y-axis will contain how many respondents are at the corresponding year (e.g., how many students are in the first year of their degree)

(b) Visualize box plots of two lists you defined above ('number of languages competency' and 'number of SQL competency')

(c) Optional for advanced programmers: visualize the number students per degree (e.g., how many students are enrolled in computer science degree)