Angela Wu

CSCI 5930

05/02/2019

## Semester project: Data cleaning for OSMI Mental Health in Tech Survey 2016

OSMI Mental Health in Tech Survey 2016 was obtained from Kaggle. The researchers who conducted the survey were interested in estimating how mental health is viewed within the tech/IT workplace. The original purpose of this data was to drive our work in raising awareness and improving conditions for those with mental health disorders in the IT workplace. I would like to use this dataset as my training data to see if I am able to come up with a method that can predict whether or not a person currently has mention illness based on answers in the survey. To do so, I will have to first clean the data, transform it into a .tsv file, and this is the goal of this Unix tool semester project.

The following ideas are involved in the clean steps: Removing the columns that consist of the same values, normalizing the string values by turning them into categorical values, reduce the varieties of the answers, removing individuals who didn’t complete the survey, and removing the columns that are not needed for the analysis. All of these were done with unix tools, and mostly the sed, cut, past functions. To determine the types of the values, I used the combination of “sort | uniq -c”. The columns were reduced down to 30 from 63, and the number of individual when from 1434 to 702. The steps are documented in the UT project\_steps file.

Here are screenshots of partial before and after files viewed with excel.

Before :



After: