Student ID # Z-15197421

Maciej Medyk

Assignment 04

DCAT Assignment in which we find related data across multiple data.gov data sites.

In this assignment the biggest challenge I found was to find the data from same time period that would have something in common with each other. I went to data.gov and downloaded two RDF files in which one deals with seat belt usage statistics per state from 2010 and second one deals with number of deaths in motor vehicle accidents per state from 2010. I wanted to see if there is correlation between seatbelt usage and safety of the occupants by simply comparing number of deaths percentage to seatbelt usage percentage. It is to be assumed that the higher the percentage of seatbelt use the lower the amount of deaths would be.

Another obstacle in this assignment is how to bring in RDF file into the program while maintaining individuality of records as one record may contain many lines. After analysis of the RDF files I noticed a pattern that I could exploit in code to bring RDF files in. With testing I noticed that I was able to bring any RDF file and it would be saved in array of arrays and each subarray would be individual record.

Last obstacle in RDF is that actual record values are preceded by a RDF meta prefix and followed by RDF meta surfix. In order to do calculations and comparison of data I had to strip the prefixes and suffixes out of each line I wanted to display or do calculations on. Again with there was a pattern that could have been exploided with python split function that allowed me to strip away and only save any value out of any line.

Afterwards all I did was to combine the two separate arrays that held data from two separate file into one array and combine was done by a common field “state”