Potential Capstone Projects

June 12, 2016

These are some of the datasets that I find interesting and can make a good capstone project out of. These data sets are obtained from the Data is plural newsletter page.

1 Are we alone ?

- Potentially habitable planets recorded by the NASA Kepler spacecraft can be found here and here. The data available is from the year 2009 onwards.
- The data sets include confirmed and candidate planets which we can use as train and test sets respectively to test some hypothesis which describes conditions for habitable planets in alien star systems.
- The datasets can be downloaded in various formats including the .csv format. We can download either the entire set or just a part of it, for examples only the variables that would be interesting to us.

2 Who favors Trump ?

- Here we try to figure out which part of the American population is favoring Trump by analyzing the rise of the Ku Klux Klan in these datasets and compare this with the states that Trump won during his nomination for the 2016 Presidential election.
- We then also use this dataset describing county typology codes in all the states of USA to determine which kind of population prefers Trump. This dataset also shows amount of education and poverty level in the various counties of each state. The data is in .xls format.
- It would be interesting to find if there is some data showing which states support the NRA and to see if there is a correlation between this and them favoring Trump.

3 Is cricket a boring game ?

- Cricket is a game played by very few nations, mostly erstwhile English colonies, and is generally considered a rather boring game by football fans, which is a much more widely played sport. One reason for this is that cricket with its various formats and rules can be a very long and tedious sport. While this statement is perhaps true to a great extent in the past it can no longer be true now with shorter formats of the form appearing increasing the appeal of the sport to a wider audience. We want to debunk the myth of cricket being boring by studying this dataset which has data of all formats of the game, thus helping us compare the shorter to the longer formats of the game. We then introduce some measure to quantify how "exciting" a game can be by looking at runs scored and the pace at which they are scored etc.
- Another way to measure how exciting a game can be is by looking at how frequently during a game "something interesting" happens. In football this amounts to how often a goal is scored in a game. In cricket this amounts to finding how often a 4 or 6 is scored or an important wicket falls. This can be easily analyzed using the dataset from this page. This is the data is plural url.

4 Is a gloomy day also a wet day ?

- We want to analyze the correlation between the amount of rainfall a place receives and annual cloud coverage of that place. The global cloud coverage dataset can be found here and the global rainfall dataset, here.
- It would be interesting to see if there are places on earth where most days are cloudy but the place receives very little rainfall !

5 How isolated are languages ?

- In this problem we want to analyze if there are similarities between languages that are spoken by geographically well separated groups. This can be done by studying similarities in the structures of these languages using datasets here.
- This can also have implications for migratory patterns of humans across the ages.