**Mini Project 2**

**IST652 - Scripting for Data Analysis**

**Due: Sunday, April 19 by Midnight**

For this mini project, you must work individually.

**Semi-structured Data Processing:**

The goal of your assignment is to write a program that will read in JSON (or another semi-structured data format), from a Mongo DB collection, from a file, or from the web.

You’ll then convert to a format that is structured with lines of data representing one type of unit (*for example, one tweet from Twitter*). Your program should contain the data as lists of JSON structures (*which, as you’ll recall, are just python dictionaries and lists*). Your program may also contain pandas dataframes for processed data.

Finally, your program will do some processing to collect data from some of the fields that will answer one or more questions, as described below, and write a file with the data suitable for answering each question. Remember that some fields may be optional or have null values, so you may need to test for those conditions.  ***Graphing is optional***.

**Questions:**

1. Process one collection of data and summarize information from ***several*** fields.
   * This is like the example program for analyzing twitter hashtags, but must access different and more fields than in those examples.
2. Process one collection of data and separate it into different categories and provided summary statistics on those categories.
   * For example, bin the tweets by day or by hour and report on the number of tweets per day or hour.
3. Process two or more collections of data and compare some summary data about the two collections together.
   * For example, collect Twitter user timelines from different political candidates and compare the number of retweets of their tweets.

**Data:**

You may collect data from Twitter, or some other URL that returns JSON data (You may also choose another semi-structured data format: XML, HTML, YAML, etc.). You should collect at least ***several hundred data items***, if possible.

**Deliverable [total: 15 points]:**

For this mini project, you must submit your data set, a program**\***, a report**\*\***, and output files. Your program must be submitted as a jupyter notebook file (**.ipynb**). You may submit the above either as separate files or as a single compressed file (*.zip/.rar*).

* **\*** A program (.ipynb) which does the following **[subtotal: 10 points]**:
  + Reads in data from MongoDB, a file, or the web [1 points]
  + Cleans and formats the data [2 points]
  + Analyzes/Summarizes the data in ***three*** different ways [6 points (2 x 3)]
  + Outputs the summaries in new files (3) with column headers [1 point]
* **\*\*** A report which describes the following **[subtotal: 5 points]**:
* The data and its source [1 point]
* A description of your data exploration and data cleaning steps [1 point]
* ***Three*** clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed. [1 point]
* A description of the program [1 point]
* A description of the output files [1 point]

For your program, you may use any of the code developed in class as a template, but it is **absolutely essential** that you use appropriate variable names and that you write original comments for what your program does. Recall that good comments demonstrate your understanding of the code that you write and the problem that you are trying to solve.