Homework 2: Semistructured Data

Due: 24 hours before the live session in Week 8

For this homework assignment, you may work on your own or you may work in a group of two people.

Semistructured Data Processing

The main outline of your assignment is to write a program that will read in JSON formatted data from a Mongo DB collection or from a file. This will be in a format that is structured with lines of data representing one type of unit, for example, one tweet for Twitter or one post from Facebook. Your program will contain the data as lists of JSON structures, which are just Python dictionaries and lists. Your program may also contain pandas dataframes for processed data.

The 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 definitely optional.

Questions:

Types of questions:

- Process one collection of data and summarize information from a number of fields. This is similar to the example programs for Twitter hashtags or Facebook counts but must access different and more fields than in those examples.

- Process one collection of data and separate it into different categories and give some 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.
- Process two or more collections of data and compare some summary data about the two collections. For example, collect Twitter user timelines from different political candidates and compare the number of retweets of their tweets.

You may use the programs twitter_lang.py as an example, but you must use different fields. You may also use twitter_hashtags.py or facebook_counts.py, but in these programs you must add a part to write a file. In all cases, you must change the comments to reflect your individual understanding of the program. If you only do one question, then it must be more complex than these simple examples; otherwise, you may choose additional questions.

Data:

You may collect data from Twitter, Facebook, or some other URL that returns JSON data. (If you want to use another format, such as XML, please ask.) If you collect your own, please collect at least several hundred data items, if possible.

You may request me to collect data for you, and if so, please make that request as soon as possible.

What to Submit:

1. The program that you write,

2. A report in a Word Document or PDF File. In it provide:

   • The data and its source, including any preprocessing

   • A clearly stated question that describes whether it is a summary or comparison question and what fields are being used in the data
   • A brief description of the program

   • A description of the output files

3. Output files (may be included in your jupyter notebook or your report)

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.

Submit your report, your program and your output file(s).

Group Work:

If you choose to work in a group (of two), you may write and submit one program, but you must process the data for (at least) two comparison questions. Each member of the group should write some part of the program, even if edited later together. Your report should describe the roles of the group members and who did what parts of the project, possibly including data, formulating questions, and debugging.