CSE 482 Exercise 1 (Date: September 08, 2020)

The purpose of this exercise is to help you get started using python libraries to load and process data. Follow the instructions below to complete the exercise. Use ipython notebook to write your code and to execute your program. Save the notebook as exercise1.ipnyb.

1. Install Anaconda python on your laptop (if you have not done so in the past). Make sure that pandas, json, and matplotlib library packages are installed.
   a. To check what packages are installed, type “conda list” on the command prompt
   b. To install a specific package, say, pandas, type: “conda install pandas”

   If you have not used iPython notebook, read the notes on “Python (Getting Started)” from the resources page on the class website.

2. Download the data file wiki_edit.txt from the class web page. The data file contains information about revisions that were made to Wikipedia articles in January 2005. Each line corresponds to a revision made by an editor to one of the articles. The format of the space-delimited columns are as follows:

   ```
   RevisionId  ArticleId  Timestamp  Editor
   ```

3. Use pandas read_table function to load the file into a data frame object named data.

   ```
   import pandas as p
   column_names = [‘RevisionId’, ‘ArticleId’, … ]
   data = p.read_table(filename, sep = ‘  ‘, header = None, names=column_names)
   ```

   a. Find the top-5 articles that have received the highest number of edits.
   b. Find the top-5 editors who have edited the most number of articles.

4. Download the JSON data file wh.json from the class web page. The data corresponds to tweet messages posted by the White House. Find the top 15 most frequent terms that appear in the tweet messages.

Deliverables: Submit the file exercise1.ipnyb OR exercise1.py. Make sure the notebook includes the results after executing your code.