CSC 177-01 Data Warehousing and Data Mining (Spring 2019)

Mini-Project 1: Data Cleaning and Analytics

Due at 2:00 pm, Friday, February 22, 2019

Demo Session: class time, Friday, February 22, 2019

In this project you will practice with data cleaning and analytics.

You will use the file "clinton_trump_tweets.txt". The file contains the tweets on Twitter collected during 2016 US presidential election. The file contains tab-separated entries with 14 columns that correspond to the following fields:

Name, ScreenName, UserID, FollowersCount, FriendsCount, Location, Description, CreatedAt, StatusID, Language, Place, RetweetCount, FavoriteCount, Text

Note: the file contains ISO-8859-1 encoded data. If you use *read_table* in pandas to read the file in, set parameter "encoding" = "ISO-8859-1".

1. Data Cleaning (40 pts)

Task 1.1: First, you need to clean up the data. Each line of the file is a tweet. **Throw away all tweets that are retweets** (the text starts with RT), and **from the text keep only the hashtags** (words that start with #) **and the handles** (words that start with @).

Task 1.2: Remove the hashtags/handles that have been used less than 20 times. Then remove the users that have used less than 20 tweets.

By doing such cleaning, we only focus on most active users and popular hashtags/handles.

2. Data Analysis by Visualization (30 pts)

Do the following on the cleaned data:

Task 2.1: Plot the number of the tweets against top-30 locations with the most tweets.
Task 2.2: Show the respective tweet word clouds of the top-3 locations with the most tweets**.**

Task 2.3: Plot the number of the tweets against top-50 users with the most tweets.
Task 2.4: Show the respective tweet word clouds of the top-3 users with the most tweets.

Task 2.5: Plot the occurrences of the top-100 most frequent hashtags/handles in the cleaned data
Task 2.6: Show the tweet word cloud of all the hashtags/handles in the cleaned data

Examine the results and include your conclusions in your report.

3. Grading Breakdown

Your grade for this project will be based on the soundness of your design, the novelty of your work, and the effort you put into the project.

Implementation	**70 pts**
Your report	**20 pts**
In-class demo	**10 pts**

4. Deliverables:

(1) All your source code/figure in Python Jupyter notebook.

(2) Your report in PDF format, with the name and id of each team member, course title, assignment id, and due date on the first page. As for length, I would expect a **report with more than one page**. Your report should include the following sections (but not limited to):

1. **Problem Statement**
2. **Methodology**
3. **Experimental Results and Analysis**
4. **Task Division and Project Reflection**

In the section "**Task Division and Project Reflection**", describe the following:

- who is responsible for which part,
- challenges your group encountered and how you solved them
- and what you have learned from the project as a team.

10 pts will be deducted for missing the section of task division and project reflection.

Notice: please don't send me the original dataset. ☺

All the files must be submitted **by team leader** on Canvas before

2:00 pm, Friday, February 22, 2019

NO late submissions will be accepted.

5. Demo Session:

Each team member must demo your work during the scheduled demo session. Each team have **three minutes** to demo your work in class. The following is how you should allocate your time:

- Findings/results (1 minute)
- Task division (1 minute)
- Challenges encountered and what you have learned from the project (1 minutes)

Failure to show up in defense session will result in **zero** point for the project.

6. Datasets:

Please use the following link on Google drive:

https://drive.google.com/open?id=1SXio43zM6JUiCtiCthwJVnE-CueRLrz8

Note: Use the file "clinton_trump_tweets.txt" for Mini-project 1.

We will use the file "clinton_trump_user_classes.txt" for Mini-project 2 and Mini-project 3.

7. Teaming:

Students must work in teams of 2 or 3 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserve the right to assign different grades to team members depending on their contributions. So you should choose partners carefully!

Hints:

- Use comparison operators for high-performance boolean selection in Pandas dataframes.
- Define custom Python functions and apply them to any Pandas DataFrame column by using *apply()*.
- To remove hashtags/handles used less than 20 times, you may concatenate all the hashtags/handles and break the resulting big string into a list, from which you create a Pandas Series, let's say, *hashtag_series*. Then use the following code to obtain the hashtags/handles appearing at least 20 times:

  top_hashtag = hashtag_series.value_counts()
  top_hash_list = top_hash[top_hash>=20].index.tolist()

- To remove users with less than 20 tweets, you may want to use ***groupby(), count()*** and ***sort_values().***
- For high-performance data filtering, you may use function ***isin()*** defined on Pandas Series. See examples here:

  https://www.geeksforgeeks.org/python-pandas-dataframe-isin/

- If you want to merge two numpy arrays, check out Numpy function ***np.concatenate()***

  https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.concatenate.html

- Pandas supports high-performance SQL join operations. Use Pandas function ***pd.merge()*** to merge (or to say, join) two dataframes based on values in one particular column. See an example here:

  https://chrisalbon.com/python/data_wrangling/pandas_join_merge_dataframe/

- Convert a Pandas Dataframe to its corresponding Numpy array representation, use the *DataFrame.values* attribute.

  http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.values.html#pandas.DataFrame.values