**User Generated Content Analytics (Fall 2018)**

**Assignment #2 (Individual work), New due date: 10/11/18 by 11:59 on Canvas**

Note: While this assignment is about Texas politics and the upcoming senate election, upon doing this assignment you will obviously see how the overall approach could be applied to a business setting where we analyze what attributes people associate with two competing products or brands, how they feel about such attributes (sentiments), and how the sentiments vary by geography.

1. Collect 4k tweets on the 2018 Texas Senate race using **getTweets.py**. It is preferable to use general hashtags and search keywords (like Texas Senate race 2018) rather than “Beto” or “Cruz” because of the probability of Cruz or Beto (or O’Rourke) being close to 1 in the latter cases. Let the occurrences of Beto O’Rourke or Ted Cruz come naturally in a subset of the tweets you collect.
2. Find **FOUR** key issues mentioned by the public in the tweets – e.g., character, personality, healthcare, crime, jobs, etc. (these are examples only, use wordfrequency.py to find out). Replace words if necessary (Using **find\_and\_replace.py** script).
3. Perform ***lift*** and ***sentiment analysis*** on each candidate and issue:

|  |  |  |
| --- | --- | --- |
|  | **Lift** | **Sentiment Score (if lift > 1)** |
| Issue 1 & Beto O’Rourke | ? | ? |
| Issue 2 & Beto O’Rourke | ? | ? |
| Issue 3 & Beto O’Rourke | ? | ? |
| Issue 4 & Beto O’Rourke | ? | ? |
| Issue 1 & Ted Cruz | ? | ? |
| Issue 2 & Ted Cruz | ? | ? |
| Issue 3 & Ted Cruz | ? | ? |
| Issue 4 & Ted Cruz | ? | ? |

1. Show the attributes and candidates on an MDS plot (Using **mds.py** script).
2. Use the location data collected in step A to find out **lifts** and **sentiments** regarding the candidates in large versus small cities/towns in Texas.
3. What advice would you give to each of the two candidates based on your analyses in C, D and E above?

**Guidelines for parts A, B, C, and D – Please read carefully before running the scripts.**

1. Before you collect tweets, find suitable hashtags and keywords (go to Twitter and also search on Google to find out which hashtags and keywords are popular in describing the Texas senate race). Keep in mind that there is some trial and error involved in finding suitable keywords for the Twitter search. As a last resort, you can do two searches, one with Beto and one with Cruz, but first you should try to find more general keywords about the Texas senate race.
2. We just found out that Twitter is not giving more than 3000 tweets in a single search. In fact if you ask for more than 3000, you will get an error. My suggestion is to do multiple searches, each for 3000 tweets, spaced at least by 12 hours (otherwise Twitter will simply send more duplicate tweets). Make sure the output file name for GetTweets.py is changed before running the script each time. Then combine the multiple csv files into a single data file with copy and paste in Excel.
3. Once you have the required tweets, check for duplicate tweets and eliminate them (in Excel you can search for duplicate rows and eliminate them – Google it if you are not familiar with it. If Twitter gives too many duplicates, try searching again after a day or so. Remember, retweets (RT) are NOT duplicates. A row of data is a duplicate if it completely matches another row.
4. Make sure the text of the tweets is in the **third column** (most of the scripts were written for the Edmunds data, and hence expect the text to be in column 3). Also ensure columns 1 and 2 are not blank. Just delete the first few columns so that the tweets are in column 3.
5. You may have to do replacements – e.g., replace O’Rourke or its variations by **Beto**, Ted or other variations by **Cruz**, etc. (Using find\_and\_replace.py script and the models.csv file with appropriate changes)
6. Run the lifts between the 4 issues (like healthcare, crime, etc.) and 2 candidates as in the Edmunds assignment (using **lift.py** and Edmunds\_key\_pair file with suitable changes). So you Edmunds\_key\_pair.txt file should contain: beto,cruz,issue1,issue2,issue3,issue4
7. For sentiment analysis with issues and candidates (need to do this for candidates & attributes where the lift > 1), use the **parserforsentiment.py** script to extract mentions of a candidate and an issue. The script uses the tweets file as input, and asks you for (i) brand (beto or cruz in this case), (ii) attribute (e.g., healthcare in this case) and the number of words to retain on the right and left hand sides of the issue (the assumption being that people express their emotions within a close proximity of the attribute). I would recommend trying 7 or 8 words. **Note that both inputs (brand and attribute) must be typed in lowercase.** The output file should be labeled carefully, e.g., beto\_healthcare.csv. **Make sure the output file is renamed each time before running the script.** Now make this output file the input to **senitment.py**. **For each candidate and each issue, you will run sentiment.py once. So for beto and cruz and 4 attributes, you will be running parserforsentiment.py and sentiment.py 8 times.**

**Guidelines for E**

1. For location analysis, get a list of Texas city and town names from Google. Use the **models.csv file** and **find\_and\_replace.py** script to replace the 10 biggest cities by population (e.g., Dallas, Houston, San Antonio, Austin, El Paso, etc.) with **big\_city** and the rest with **small\_town**. In the models.csv file, add big\_city and small\_town in the left column and the real city/town names of Texas in the right column, and run **find\_and\_replace.py**. **To do this replacement, temporarily make the location field (in your input file) the 3rd column, since the find\_and\_replace.py script works on the third column**.
2. Once the replacements are done, **merge the location and the text columns** into one with the concatenate command in Excel (Google it if not familiar). Make sure this is the third column in the data file.
3. Perform a lift analysis as follows:

|  |  |  |
| --- | --- | --- |
|  | Beto O’Rourke | Ted Cruz |
| Big\_city Texas | Lift? | Lift? |
| Small\_town Texas | Lift? | Lift? |

Obviously the Edmunds\_pair\_keys.txt file should contain: beto,cruz,big\_city,small\_town

1. Use parserforsentiment.py again on this dataset, but now brand=candidate beto or cruz, attribute=big\_city or small\_town and # words = 20. The output of this script should be a file which has tweets on a candidate from big\_city or small\_town. Now pass this output file through sentiment.py to get the overall sentiment of big\_city or small\_town residents about a candidate. **You will run the parserforsentiment.py and sentiment.py 4 times.**

Create a matrix as follows from the sentiment scores:

|  |  |  |
| --- | --- | --- |
|  | Beto O’Rourke | Ted Cruz |
| Big\_city Texas | Sentiment score? | Sentiment score? |
| Small\_town Texas | Sentiment score? | Sentiment score? |