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

**Assignment #2, Handed out: 10/08 Date Due: 10/22 by 23:59 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. It is preferable to use general hashtags and search keywords (like Texas Senate Race) 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 word frequency analysis to find out). Replace words as necessary (e.g., replace O’Rourke by Beto).
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 | ? | ? |

To do the above analysis, you will have to write a script to parse each tweet and extract parts that pertain to a candidate and attribute. Provide details of how you accomplished this task in a script.

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. The Twitter API gives a maximum of 3000 tweets in a single search. 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 in your script 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. You may have to do replacements – e.g., replace O’Rourke or its variations by **Beto**, Ted or other variations by **Cruz**, etc.
5. Run the lifts between the 4 issues (like healthcare, crime, etc.) and 2 candidates as in the Edmunds assignment
6. For sentiment analysis with issues and candidates (need to do this for candidates & attributes where the lift > 1), write a script to extract mentions of a candidate and an issue. You can assume that people express their emotions within a close proximity of an attribute. I would recommend trying 6-7 words as the proximity or radius.

**Guidelines for E**

1. For location analysis, get a list of Texas city and town names from Google. Replace 5-10 biggest cities by population (e.g., Dallas, Houston, San Antonio, Austin, El Paso, etc.) with **big\_city** and the rest with **small\_town**.
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? |

1. 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? |