**User Generated Content Analytics (MSITM, Fall 2019)**

**Assignment #2**

Note: While this assignment is about politics and the 2020 Presidential election, upon doing this assignment you will obviously see how the overall approach can 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 5k tweets on the 2020 Presidential election. It is necessary to use general hashtags and search keywords (like 2020 U.S. election or 2020 Presidential Election) rather than “Trump” or “Biden” or “Warren”. Use Tweepy in your Python script, and also ask for the location specified by the user during registration with Twitter.
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 analysis to find out). Replace words if necessary. Issues associated with Mr. Trump may not have a strong association with the democratic candidate and vice versa. So find a couple of issues that are likely to have strong association with Mr. Trump, and a couple for the democratic candidate.
3. Perform ***lift*** and ***sentiment analysis*** on a candidate and an issue:

|  |  |  |
| --- | --- | --- |
|  | **Lift** | **Sentiment Score (if lift > 1)** |
| Issue 1 & Trump | ? | ? |
| Issue 2 & Trump | ? | ? |
| Issue 3 & Trump | ? | ? |
| Issue 4 & Trump | ? | ? |
| Issue 1 & Dem Candidate | ? | ? |
| Issue 2 & Dem Candidate | ? | ? |
| Issue 3 & Dem Candidate | ? | ? |
| Issue 4 & Dem Candidate | ? | ? |

To isolate a candidate and an issue, you have to write a parser that selects tweets that mention a candidate, takes a window around an issue, and chops off everything else, as shown in class.

1. Show the issues and candidates on an MDS plot. Note that a large range if lift values makes it difficult for mds to represent accurately on a 2-dimensional plane. If needed, artificially (and arbitrarily reduce the high values and increase the low values).
2. Use the location data collected in step A to find out **lifts** and **sentiments** regarding the candidates in three states – Michigan, Pennsylvania and Wisconsin (no need to treat them separately, call them “battleground” states).
3. What advice would you give to each of the two candidates above based on your analyses in C, D and E above?

**Guidelines for parts A, B, C, and D**

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 2020 election). Keep in mind that there is some trial and error involved in finding suitable keywords for the Twitter search.
2. You may have to do replacements – e.g., replace Donald with Trump, etc.
3. Run the lifts between the 4 issues and 2 candidates as in the Edmunds assignment
4. For sentiment analysis with issues and candidates (need to do this for candidates & attributes where the lift > 1), extract mentions of a candidate and an issue.
5. For location analysis, get a list of cities and town names from Google for Michigan, Pennsylvania and Wisconsin. Replace these cities/towns by **battleground1**.
6. 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).
7. Perform a lift analysis as follows:

|  |  |  |
| --- | --- | --- |
|  | Trump | Top Dem mentioned |
| Battleground1 | Lift? | Lift? |

1. Create a matrix as follows from the sentiment scores:

|  |  |  |
| --- | --- | --- |
|  | Trump | Top\_dem |
| Battleground1 | Sentiment score? | Sentiment score? |