Topical Analysis Across United States Using Twitter

Brandon Beylo, Josh Moen

Objective

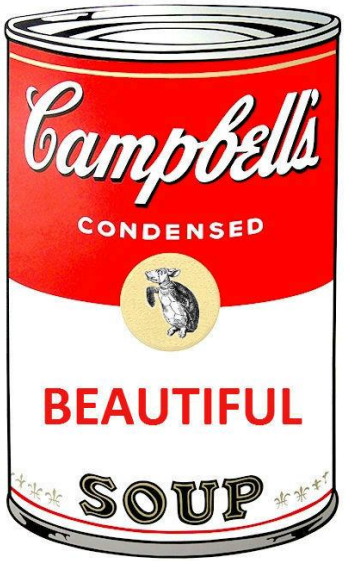
Purpose of Project

- Create a List of 4 Topics (Sports, Politics, Religion, and Finances)
- Mine Twitter for Tweets from various age groups regarding the above Topics.
- Analyze Comparative Distribution of Topics

Why Twitter?

- The perfect place to mine real - time sentimental data.
- Freedom of Expression is more accurate when done online.
- Twitter has very smooth API to create programs with it.

Tools We Will Use



Tools cont.

- Twitter API
  - Will use API to interact with Twitter on the backend to run our Streaming script.

- TextBlob
  - TextBlob will be used to create a Sentiment Analysis based off of the texts of the tweets, from a scale of Positive to Negative.

- Sci-kit Learn
  - We will use Sci-kit Learn to create a Machine Learning algorithm that will be able to predict text based on user data.

Benefits & Uses of Program

Benefits of this Program

- Customizability by small changes in the program based on specific needs.
- Geographical Insights into Most Popular Topics.
- Real Time Changes in Preferences can be Realized.

Potential for Business Use

- Business Owners get in-depth look at customers' preferences.
- Curuate their marketing strategies based on most popular topics.
- Use Age Specific Ranges to cater to target ranges individually.

An Update on Our Progress ...

Economic Literature

Modern Technologies for Big Data Classification and Clustering

- This literature contains a chapter that goes in depth about Twitter data analysis. The author outlines topics about how there are three major classifications of Twitter data analyzation such as Content based, Networked based and Hybrid analysis.

Sentiment Analysis of Twitter Data

- This literature contains information on analyzing the sentiment of tweets. The sentiment analysis is on a scale between -1 and 1. The closer to -1 the value, the more negative the tweet and vice versa.

Word Cloud Examples

Taken from small sample

Various Locations in NY

Cleaning Up The Data

- Classified Data into Specific CSV DataFrames.
  - Tweet_Text.csv is our Tweet File
  - UserLocations.csv is our Location File
  - Enables Easier Use of Various DFs.

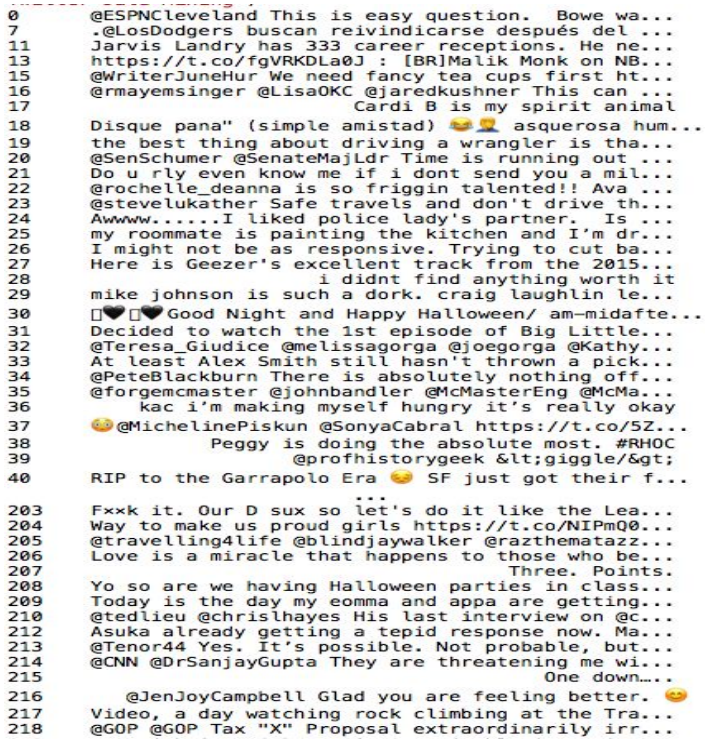
```
# Creating Specific Data Frames

df_user_loc = df['user_location'].dropna(how = 'any')
df_text = df['text']
df_follower_stats = df['followers_count']
df_user_name = df['screen_name']
```

- Removing all Retweets from one Dataset
  - Doing this will create a cleaner and clearer dataset for which to interpret consensus.
  - This is shown in a clearer word cloud.

Examples of Data Cleaning

```
0     @ESPNCleveland This is easy question.  Bowe wa...
1     RT @PDXStephenG: There goes @russellokung show...
2     RT @MPearce6: Ep 174 - Matt has a PSA for all ...
3     RT @TheHockeyNews: #NHL goalies: Even-strength...
4     RT @ProCityHoops: Eric Gordon! For the win! #N...
5     RT @C_MoralesD: FINALISTAS GUANTE DE ORO (Jard...
6     RT @PronosRush: #NHL\n\nPittsburgh + Capitals ...
7     .@LosDodgers buscan reivindicarse después del ...
8     RT @JeffGSpursZone: Ginobili on Rdy Gay's play...
9     RT @All_SportNews: ¡George Springer CLUTCH HOM...
10    RT @LethalShooter__: Social media has 0 chill ...
11    Jarvis Landry has 333 career receptions. He ne...
12    RT @pastormarkburns: The #NFL are Politically ...
13    https://t.co/fgVRKDLa0J : [BR]Malik Monk on NB...
14    RT @adnazteca: #TipicoQueConFrio hay #NFL anál...
Name: text, dtype: object
```

What a Cleaner Dataset Looks Like



```
0      @ESPNCleveland This is easy question.  Bowe wa...
7      .@LosDodgers buscan reivindicarse después del ...
11     Jarvis Landry has 333 career receptions. He ne...
13     https://t.co/fgVRKDLa0J : [BR]Malik Monk on NB...
15     @WriterJuneHur We need fancy tea cups first ht...
16     @rmayemsinger @LisaOKC @jaredkushner This can ...
17                              Cardi B is my spirit animal
18     Disque pana" (simple amistad) 😂👨‍🦱 asquerosa hum...
19     the best thing about driving a wrangler is tha...
20     @SenSchumer @SenateMajLdr Time is running out ...
21     Do u rly even know me if i dont send you a mil...
22     @rochelle_deanna is so friggin talented!! Ava ...
23     @stevelukather Safe travels and don't drive th...
24     Awwww......I liked police lady's partner.  Is ...
25     my roommate is painting the kitchen and I'm dr...
26     I might not be as responsive. Trying to cut ba...
27     Here is Geezer's excellent track from the 2015...
28                           i didnt find anything worth it
29     mike johnson is such a dork. craig laughlin le...
30     □❤□❤Good Night and Happy Halloween/ am-midafte...
31     Decided to watch the 1st episode of Big Little...
32     @Teresa_Giudice @melissagorga @joegorga @Kathy...
33     At least Alex Smith still hasn't thrown a pick...
34     @PeteBlackburn There is absolutely nothing off...
35     @forgemcmaster @johnbandler @McMasterEng @McMa...
36        kac i'm making myself hungry it's really okay
37     🙄@MichelinePiskun @SonyaCabral https://t.co/5Z...
38                Peggy is doing the absolute most. #RHOC
39                     @profhistorygeek &lt;giggle/&gt;
40     RIP to the Garrapolo Era 😔 SF just got their f...
                                    ...
203    F××k it. Our D sux so let's do it like the Lea...
204    Way to make us proud girls https://t.co/NIPmQ0...
205    @travelling4life @blindjaywalker @razthematazz...
206    Love is a miracle that happens to those who be...
207                                          Three. Points.
208    Yo so are we having Halloween parties in class...
209    Today is the day my eomma and appa are getting...
210    @tedlieu @chrislhayes His last interview on @c...
212    Asuka already getting a tepid response now. Ma...
213    @Tenor44 Yes. It's possible. Not probable, but...
214    @CNN @DrSanjayGupta They are threatening me wi...
215                                        One down….
216       @JenJoyCampbell Glad you are feeling better. 😊
217    Video, a day watching rock climbing at the Tra...
218    @GOP @GOP Tax "X" Proposal extraordinarily irr...
```

Looking Forward ...

This Program Has Scalability

- Since this program works on small data sets, it will now be able to be scaled to larger datasets.

- This will enable us to create a clearer picture of preferences.

Machine Learning & Sentiment Analysis

- Next Step is to develop a Sentiment Analysis Script
  - This will rank each tweet text from Positive to Negative
- The benefit of this will be to see an emotional layout of topics across the U.S.

- Develop a Machine Learning algorithm to predict type of user based on tweet and location data.
  - This will be created through the use of Sci-Kit Learn.

Further Goals

1. Increase Database to Include multiple days of data.

2. Plot cities against each other by certain criteria (such as follower count)

3. Create Word Clouds for all major colleges in Maryland area.