Mary Anderson  
Text Mining Acquire & Analyze  
November 8, 2020

**Repository Link:** <https://github.com/maryeand/TM_Twitter_McConnell-McGrath>

**Executive Summary:** Twitter profile descriptions can contain a brief snapshot as to the type of person who owns that account. I pulled the descriptions for all followers of Kentucky’s Republican Senator Mitch McConnell and Amy McGrath, the Democratic candidate who recently lost to McConnell in the election. Analysis of their followers’ descriptions shows strong political party loyalty by looking at the top unique words between the two candidates. I looked at the frequency of these unique words per candidate and used this information to create a word cloud.

**Introduction:** The recentKentucky senate race between Mitch McConnell and Amy McGrath was in the national spotlight as McConnell is a Republican and the current Senate Majority Leader. Amy McGrath unsuccessfully ran against him as a Democrat. This specific Senate race caught my attention as Mitch McConnell has held this seat since 1985 and his role as Senate Majority Leader is a significant and important role for our country. With the current political climate, I wondered if Amy McGrath might be able to win or at least come close to winning.

Additionally, McConnell and McGrath vary widely in their political beliefs and I felt this would make for interesting analysis of their Twitter followers.

**Dataset information:**

This data was scraped from Twitter using Tweepy. The data is in two text files, one for McConnell and McGrath each. The information scraped includes the follower’s handle, location, follower count (how many people follow them), friend count (how many people they follow), and their profile description.

**Descriptive statistics:**

Size of Data: Two text files, one for McConnell (149 MB, 1,938,905 rows) and one for McGrath (62 MB, 593,986 rows).

Twitter specific characters: When looking at the initial stats of the data, I did basic cleaning of the text but wanted to keep in hashtags. Instead of just keeping only alphanumeric characters, I made sure to still include #’s with the following code: text\_clean = [w for w in text\_clean if w.isalpha() or '#' in w]. Emojis are still present in the text data as well.

|  |  |  |
| --- | --- | --- |
|  | **McConnell** | **McGrath** |
| **# of Followers** | 1,938,904 | 593,985 |
| **Tokens** | 9,846,150 | 4,133,492 |
| **Unique Tokens** | 1,555,600 | 635,767 |
| **Avg. Token Length** | 6.33 | 6.49 |
| **Lexical Diversity** | 0.16 | 0.15 |
|  | | |
| **Top 10 Words in Followers’ Descriptions** | | |
| 1 | USA, 6.5% of followers | USA, 10.2% of followers |
| 2 | Love, 3.9% | New, 4.1% |
| 3 | United, 2.4% | ca, 4.0% |
| 4 | States, 2.2% | Love, 3.7% |
| 5 | New, 2.0% | #Resist, 3.1% |
| 6 | Trump, 2.0% | Proud, 2.8% |
| 7 | God, 1.6% | United, 2.7% |
| 8 | Proud, 1.6% | States, 2.5% |
| 9 | #MAGA, 1.6% | ny, 2.4% |
| 10 | Life, 1.5% | Mom, 2.0% |

*Table 1: Descriptive Stats*

**Methods:**

I used the code from the Twitter assignment to create this data. To begin the analysis of this text, I first had to clean and tokenize the text files. Next, I created a function to look at the descriptive statistics of the files.

From there, I looked at the top 1,000 words by frequency for each candidate. Seeing that some of these words were shared, I subtracted the sets from each candidate to just see the top unique words. Finally, I created a word cloud based on the top unique words for each candidate.

**Results:**

Expanding beyond the initial descriptive statistics of the dataset, I was interested in looking at the differences between McConnell’s and McGrath’s followers.

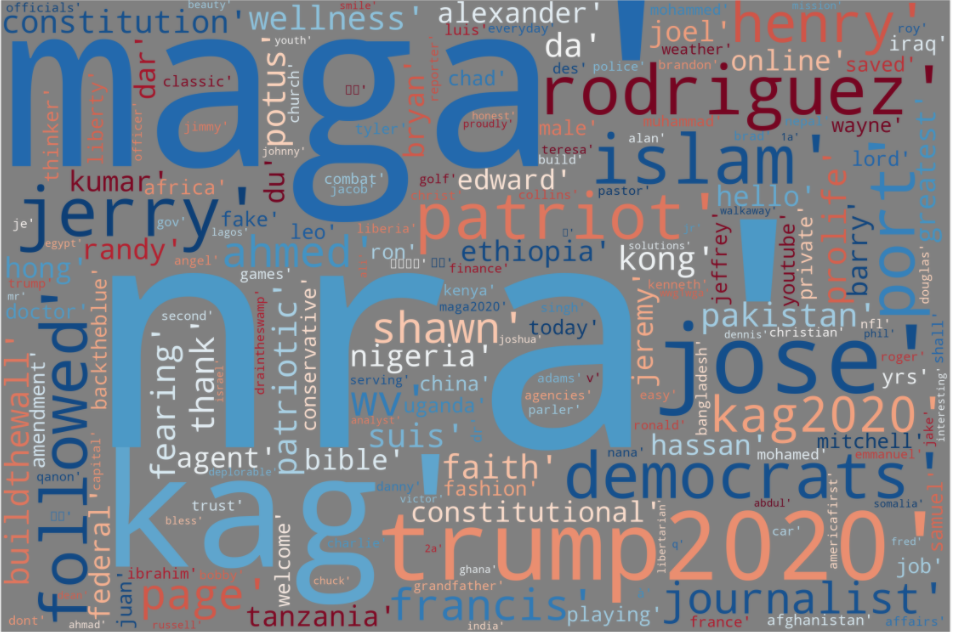
As some top words are shared between the two candidates, I next removed the similar words to just look at the top unique words. Strong themes are present between the two sets of text, clearly demonstrating a political party affiliation. I had to look up several terms that appeared in each of the candidate’s top words, and these words or acronyms were typically strongly representative of a political party. For example, #KAG was not immediately obvious to me. This is from Donald Trump’s Presidential campaign and means “Keep America Great.” I was also unfamiliar with “#FBR,” which means “Follow back resistance.” The intent of this hashtag is to create noise for individuals who are part of the Trump resistance movements. When someone sees this hashtag, they will follow each other and share the follow request with their Twitter followers.

From the top unique words, the following themes stood us as being indicative of political affiliation:

* Mitch McConnell - Republican
  + #maga, maga
  + #trump2020, #trump
  + #kag (keep America great), #kag2020
  + #wwg1wga (where we go one, we go all – associated with Qanon), #qanon
  + #conservative
  + #buildthewall
  + #prolife
  + Deplorable
* Amy McGrath - Democrat
  + #blm, blm, #blacklivesmatter
  + #bidenharris2020, #bidenharris, #biden2020
  + #fbr (Follow Back Resistance)
  + #theresistance, #resistance, resist, #resister, resistance, resister
  + #voteblue, #votebluenomatterwho, #bluewave2020
  + Progressive
  + Mask, #wearamask
  + Climate
  + Feminist, Nasty, #metoo
  + #notmypresident, #fucktrump, #nevertrump
  + Gay, #lgbtq

I created word clouds of the top unique words for each candidate based on frequency (Figures 1 and 2).

  
Figure 1: Wordcloud Amy McGrath’s Followers’ Descriptions Top Words

  
Figure 2: Wordcloud of Mitch McConnell’s Followers’ Description Top Words

I thought that McConnell’s followers might be more spread out across political parties given his role as Senate Majority Leader. The unique top words for both McConnell and McGrath are very clear indicators that their Twitter followers share political beliefs with the candidate and each other.

**Resources:**

<https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5>

<https://matplotlib.org/3.2.1/tutorials/colors/colormaps.html>