Samuel L. Peoples
IST 736: Text Mining – Final Project
19 March 2019

Congressional Twitter Analysis
LDA Topic Modeling & K-Means Clustering of Member Tweets

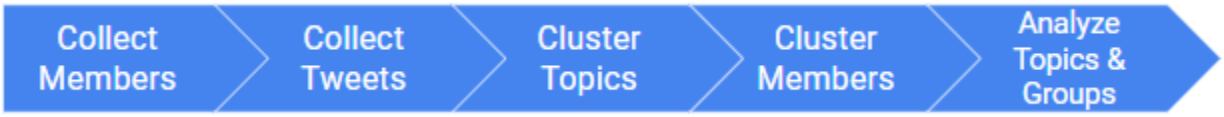
1. Introduction

Lawmakers use social media for various things such as lobbying for legislation, providing support to their constituents, and promoting their own personal publicity. This project aims to identify primary issues discussed on Twitter by members of the United States House of Representatives and Senate. Through the implementation of Latent Dirichlet Allocation topic modeling and K-Means Clustering, the ability to attribute clusters of topics to different groups of lawmakers will provide insight into the behavior of members of our government and contribute to the notion of a more informed voter.

2. Method

Python will be the primary data collection resource, using *tweepy* to handle API requests, *json* to parse Tweet statuses, *re*, and *emoji* to compile a regular expression string to remove special characters from Tweets. *Mallet* is utilized to streamline the LDA analysis, and *SKLearn* is used to conduct the K-Means Clustering. *Tableau* is then leveraged to complete exploratory data analysis and classification of the dual-clustering methods implemented.

The data collection, preprocessing, and analysis is conducted across five steps:



a. Collect Members

Members are identified through a Twitter-based list managed by CSPAN[1]. This list contains members of the House and Senate, as well as Twitter accounts for various government committees. *Tweepy* is used to request the names and user-ids from the list and exported to a CSV file. These values are manually cleaned to match names from a supplemental data file managed by Ballotpedia[2] containing member names, offices, date of assuming office, and party affiliation.

Once this data is joined, the party affiliation is one-hot encoded, generating respective boolean columns for Republicans and Democrats, dropping Independents to avoid the dummy-variable trap. The date of assumed office is transformed into a float value representing the seniority in years, and the member's office is encoded to allow the House to be represented by zeros and the Senate to be represented by ones.

b. Collect Tweets

The collected Twitter handles are passed back into *tweepy*, requesting the latest two hundred tweets from each account. As the tweets are collected, they are stripped of links, special characters, and emojis using the *re* and *emoji* packages. Fifty-seven members returned less than two hundred tweets and were reattempted to ensure that the maximum tweets were collected; two members returned zero tweets and were excluded. The

[1] https://twitter.com/cspan/lists/members-of-congress/members?lang=en
[2] https://ballotpedia.org/List_of_current_members_of_the_U.S._Congress

remaining fifty-five members with incomplete samples were retained to maintain a robust corpus of text and can be attributed to newer members of Congress who recently took office.

The resultant dataset contained data for 527 of 535 members of Congress and saved the latest two hundred tweets for each member to *[twitter_handle].txt* to facilitate analysis within Mallet.

c. Cluster Topics

The corpus of Tweets is then passed into Mallet to develop a topic model, retaining the original sequence and removing stopwords to reduce noise within the data.

```
bin\mallet import-dir
      --input [REDACTED]\congress\
      --output [REDACTED]\congress_twitter.mallet
      --keep-sequence
      --remove-stopwords
```

Topics are then trained, examining the output of ten, fifteen, twenty-five, and fifty clusters, ultimately selecting fifty to observe the greatest diversity between clusters. Although some clusters had repeating topics, the content of the keys revealed varying policy positions and levels of rhetoric.

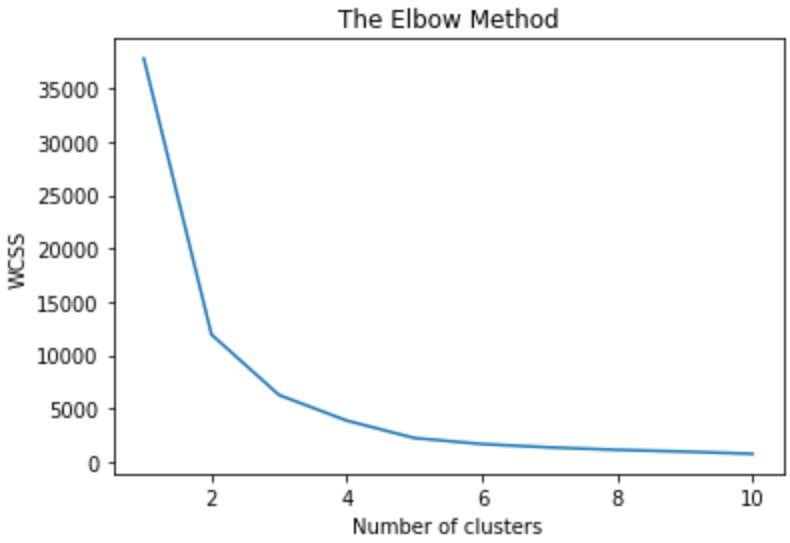
```
bin\mallet train-topics
      --input [REDACTED]\congress_twitter.mallet
      --num-topics 50
      --output-state 50-congress-topic-state.gz
      --output-topic-keys 50-congress-keys.txt
      --output-doc-topics 50-congress-topics.txt
      --optimize-interval 10
```

d. Cluster Members

The topic weights are then joined with the previously collected data, resulting in a (527 x 54) K-Means input vector:

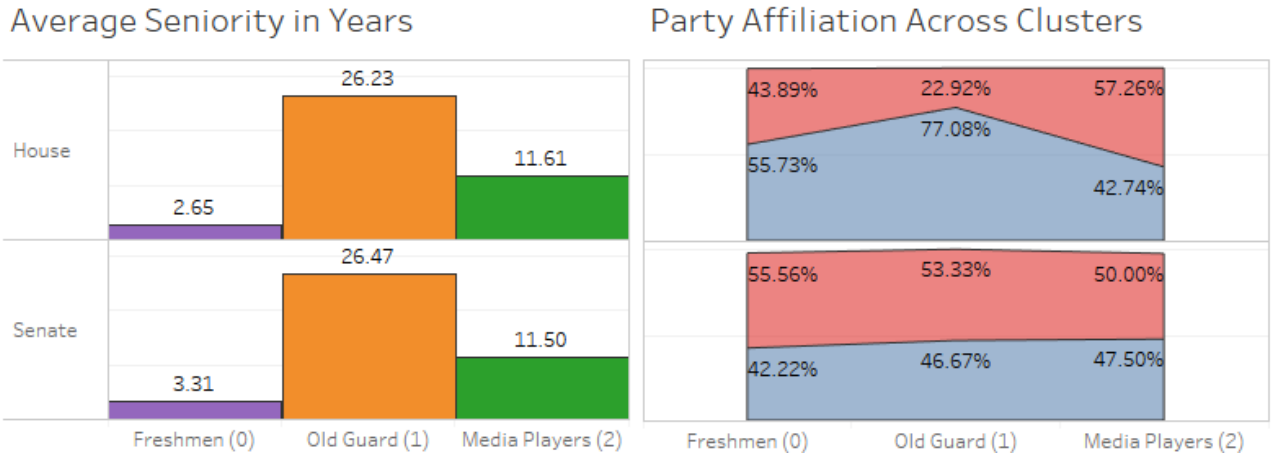
Republican	Democrat	Congress	Seniority	50_c0	50_c1	50_c2	50_c3	50_c4	50_c5	...
1	0	1	8	0.000009	0.000023	0.000007	0.000009	0.000005	0.000011	...
1	0	1	8	0.000013	0.000032	0.000010	0.000012	0.014366	0.000016	...
0	1	1	8	0.000011	0.013607	0.000009	0.000011	0.000006	0.000014	...
1	0	1	16	0.000009	0.000024	0.015297	0.000009	0.000005	0.000012	...
1	0	0	17	0.000013	0.017745	0.000011	0.000013	0.000007	0.000016	...

The optimal number of clusters is identified using The Elbow Method, which revealed an optimal number of three clusters. The predicted clusters are stored as a column in the input dataframe, and saved for input to *Tableau*.

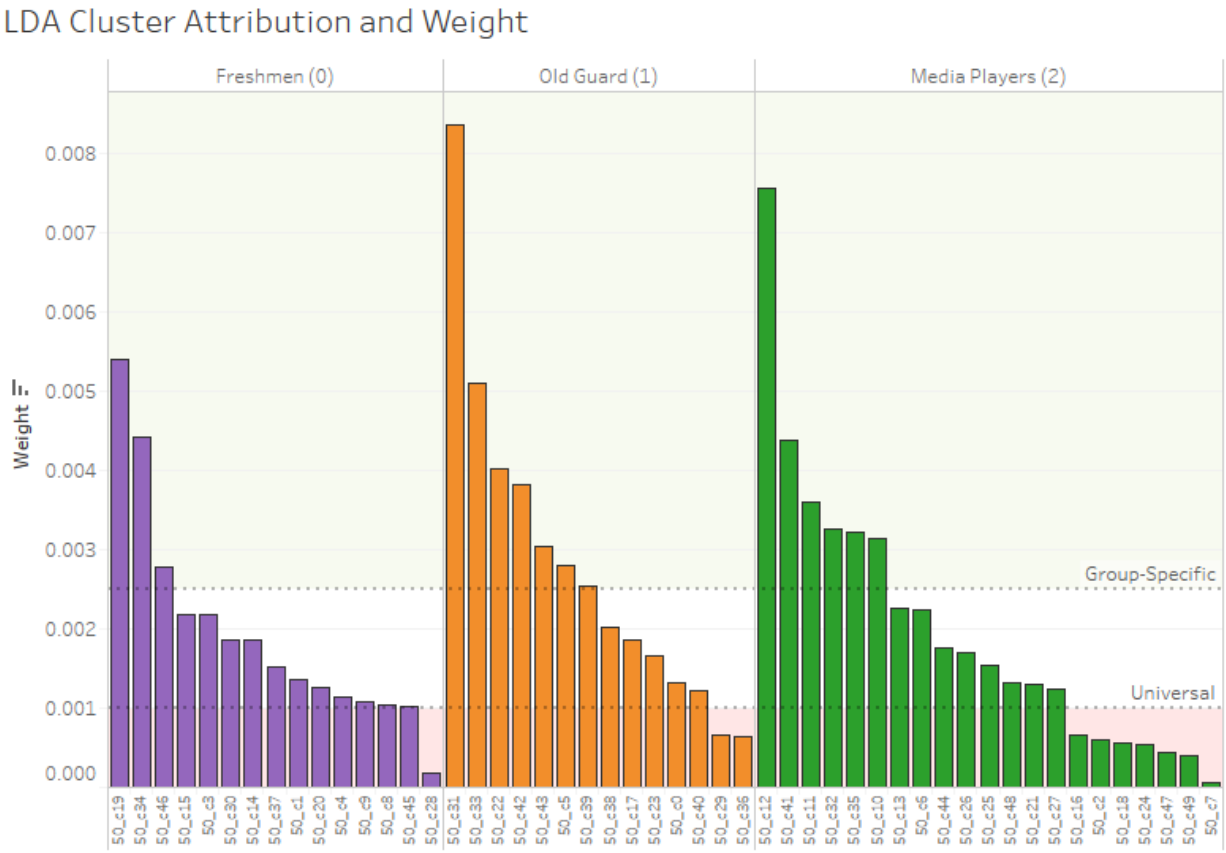


e. Analyze Topics & Groups

The clustered members and output topics are then read into *Tableau*, visualizing the average seniority in years across clusters and between the House and Senate, as well as the party affiliation across the same fields. Three groups are identified, being Freshmen: those with very little experience, Old Guard: members with between twenty and twenty-five years of experience, and Media Players: the career politicians and committee members typically presenting their agenda on national media outlets.



Attribution is then made to the labeled clusters by averaging the LDA weights across the cluster and subtracting that from the average of the weights within the cluster. This provides the magnitude of weight that is attributed from each topic. Thresholds are defined to identify universal topics, which have an attribution weight less than 0.001, and group-specific topics, which are significantly weighted towards the respective clusters with an attribution weight greater than 0.0025. The results in ten universal topics, and sixteen significant group-specific topics.



3. **Result**

   a. **Universal Topics**
      - Freshmen preferred:
        - 50_c28: State and Federal Environmental Policy
      - Old Guard preferred:
        - 50_c29: Ohio Job and Trade Growth
        - 50_c36: 9/11 Victim's Compensation Renewal
      - Media Players preferred:
        - 50_c24: General Inter-Member Debate
        - 50_c16: Border & Cartel Drugs (Fentanyl)
        - 50_c7: North Carolina Hurricane Recovery
        - 50_c2: Vaccine, Health, and Veteran Outreach
        - 50_c47: Michael Cohen Hearings & Sentencing
        - 50_c18: Polyfluoroalkyl Water Contamination; Mississippi & Michigan
        - 50_c49: Gun Control Expansion Debate Following Shooting in FL.

   b. **Freshmen Topics** (in weighted order)
      - 50_c19: Town halls Surrounding 30+ Day Government Shutdown
      - 50_c34: Demands for Federal Worker Backpay Surrounding Shutdown
      - 50_c46: New England SALT Discussions

   c. **Old Guard Topics** (in weighted order)
      - 50_c31: Veteran and Child Healthcare Funding
      - 50_c33: Family Separation and DACA Discussion; Border Policy
      - 50_c22: New Year's & Congratulatory Remarks

- 50_c42: Midwest Election Discussions
- 50_c43: Protection of Women and Families
- 50_c5: Texas Energy & Border Discussions
- 50_c39: 2018 Utah Wildfires

d. Media Player Topics (in weighted order)

- 50_c12: Border Security, Trump & the Wall
- 50_c41: Socialism, Maduro & Humanitarian Crisis
- 50_c11: Personal plug to watch the member on a cable network
- 50_c32: Fake Emergency, Trump's Shutdown
- 50_c35: AG Barr Nomination and Hearings
- 50_c10: Midwest Small Business Policy

4. Conclusion

How a lawmaker chooses to engage on social media provides valuable insight into the policies which are important to them. Using this analysis, constituents may be armed with insights into which topics are commonly debated in different groups of Congressional representatives, as well as where members are classified based on the content of their Tweets.

Some topics are less politically driven and are universally discussed, with slight preference to some groups over others. Freshmen discuss environmental policy negligibly more than the rest of the groups, while a similar pattern is observed for the 9/11 Victims Compensation Fund, which is weighted slightly towards the Old Guard. The Media Players are slightly preferred for topics such as Fentanyl being found at the border, or PFMS water contamination in Mississippi and Flint, Michigan. Although these topics are slightly preferred by one group over the others, there is a negligible observation of the magnitude of weight across the cluster, indicating these issues are important to the majority of representatives.

Some groups are more likely to engage on social media and focus on passionate topics, such as the Freshmen. This group includes low-experience members who are primarily Democrats in the House and Republicans in the Senate, such as Alexandria Ocasio Cortez, or Ilhan Omar. Important topics to this group include expanding state and local taxes in New England, and the recent government shutdown.

Others are more careful with the content of their Tweets and tend to remain positive and congratulatory in their responses, such as the Old Guard. This group is heavily weighted by House Democrats and tends to have roughly twenty-five years of experience; members in this group include Steve Scalise, Ted Cruz, and Ted Lieu. Important topics to this group include veteran and child healthcare, border policy, and energy policy.

Some topics are highly rhetorical and inflammatory, and frequently include requests to follow up with content on a national news network. Members in this group tend to have around ten years of experience and are roughly similar in party affiliation; this includes members such as Matt Gaetz, Nancy Pelosi, and Maxine Waters. Important topics to this group include mentioning the President, Debate surrounding Socialism and Venezuela, as well as personal requests to find them on a national news network, such as Fox or CNN.

The methods implemented provide insight into the behavior of Congressional representatives, topics discussed within groups of members, as well as the most important topics to members within the different groups. Clustered members were easily profiled with the selected data, and clear topics were identified using LDA analysis.