Congressional Twitter Analysis

LDA Topic Modeling & K-Means Clustering of Member Tweets

Samuel L. Peoples
IST 736: Text Mining

Objective

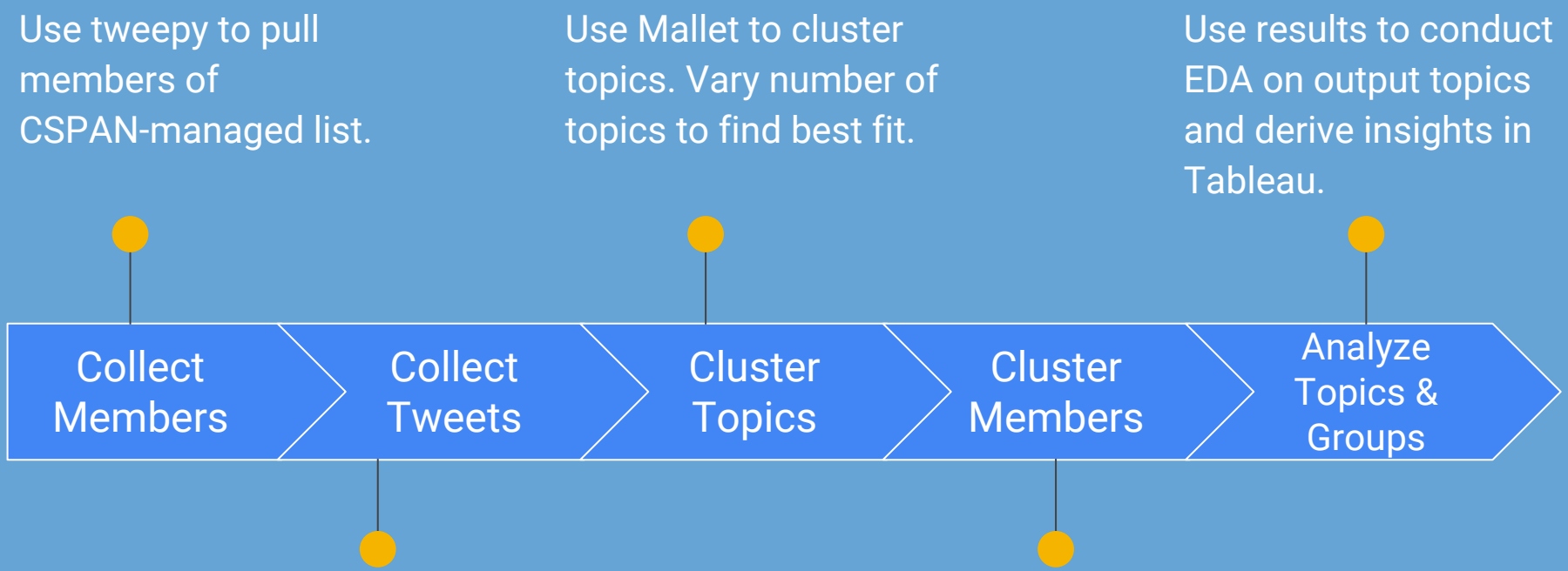
- Lawmakers use social media for various things such as lobbying for legislation, providing support to their constituents, and promoting their own personal publicity.
- This project will gather Tweets from 527[1] of 535 members of the House and Senate to develop insights surrounding how they use this service.
- Member topic clusters will be grouped by seniority, party affiliation, and house of congress[2] to facilitate exploratory data analysis.

---

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

Tools & Methods

- Python will be the primary data collection resource, using **tweepy** to handle API requests, **json** to parse Tweet statuses, and **emoji** to compile a regular expression string to remove emojis from Tweets.
- The returned tweets are written to text files named after the individual twitter handles, and passed into **Mallet** for *LDA analysis*.
- The LDA output is joined with profiling data and grouped via *K-Means Clustering*.
- **Tableau** is used for exploratory data analysis following secondary clustering.


Use tweepy to pull members of CSPAN-managed list.
Use Mallet to cluster topics. Vary number of topics to find best fit.
Use results to conduct EDA on output topics and derive insights in Tableau.
Collect Members
Collect Tweets
Cluster Topics
Cluster Members
Analyze Topics & Groups
Clean data and merge with profiling ballotpedia data. Use tweepy to request 200 tweets for each member.
Join LDA results with profiling data. Use K-Means to group topic clusters.

Data Collection

- **Tweepy** is used to pull list of congressional twitter accounts managed by *CSPAN*
- Names are cleaned to match profiling data from *Ballotpedia* and joined to be passed back into **Tweepy** to pull member tweets.
- Some accounts belong to committees and are excluded, resulting in 527 out of 535 active members of the House and Senate with Twitter accounts.
- 57 had fewer than 200 tweets, and two returned zero.

Data Collection

- **Tweepy** is then used to request the latest 200 Tweets from each member, controlling for API rate limits with **tqdm**.

- Links, emojis, and common symbols are excluded using **re** and **emoji** prior to being written to *[member].txt* using utf-8 encoding.

Topic Modeling

- **Mallet** is used to develop a topic model with stopwords removed.

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

Topic Modeling

- **Mallet** is then used to cluster the topics in 10, 15, 25, and 50 clusters.
- Because the topics are so variant, fifty topics were ultimately selected.

```
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
```

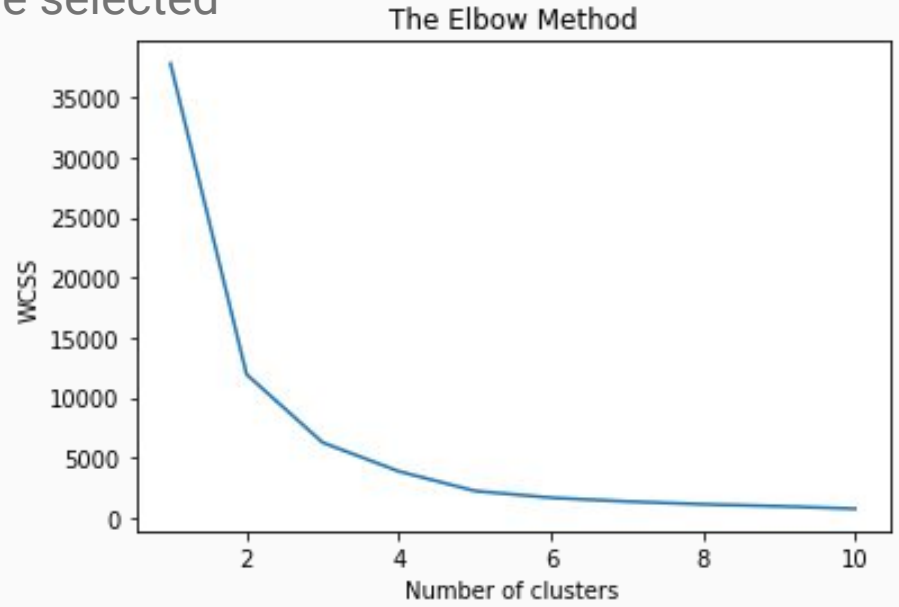
K-Means Clustering

- The output topic weights are then joined on the member's Twitter username to merge the LDA results with the member's:
  - Boolean flag whether or not the member is a Republican
  - Boolean flag whether or not the member is a Democrat
  - The House of Congress to which the member belongs {0: Senate, 1:House}
  - Seniority in years.

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	...

K-Means Clustering

- The (527 x 54) vector is then passed into K-Means to determine optimal clusters.
- Three clusters are selected

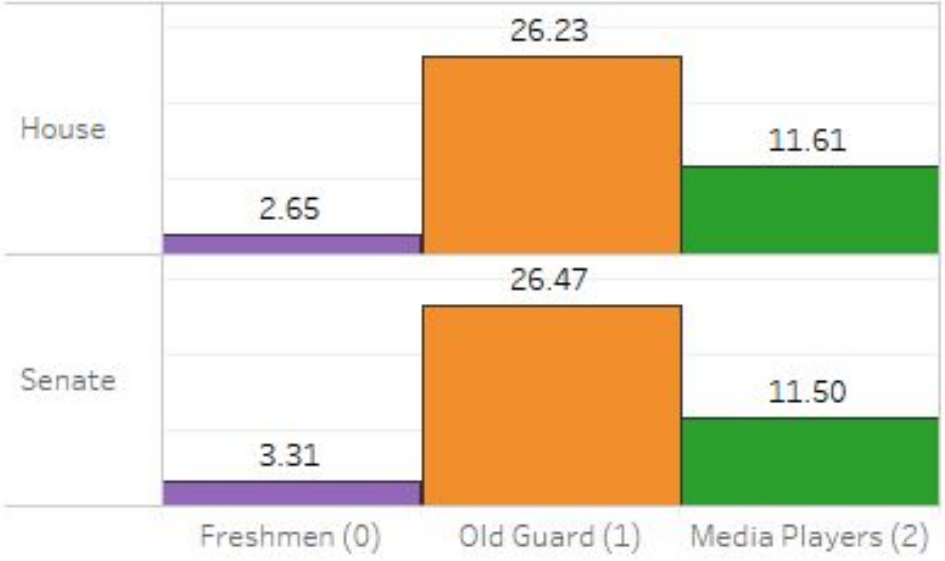


Profiling

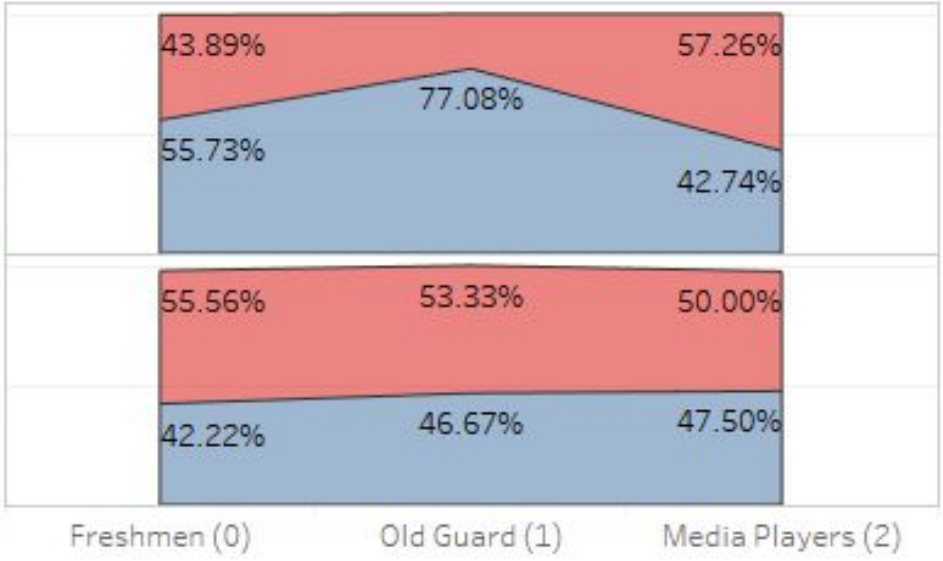
- The three clusters reveal stark differences in seniority across across groups, but there is less variance between the House and Senate.
- The Senate has fairly uniform proportions across clusters, while the House has varying proportions of Republicans and Democrats.
- Three distinct groups are identified:
  - Cluster 0: Freshmen - Congress members with very little experience
  - Cluster 1: Old Guard - Representatives that average twenty five years of experience
  - Cluster 2: Media Players - Members who discuss hot topics and engage in debate

Profiling

Average Seniority in Years



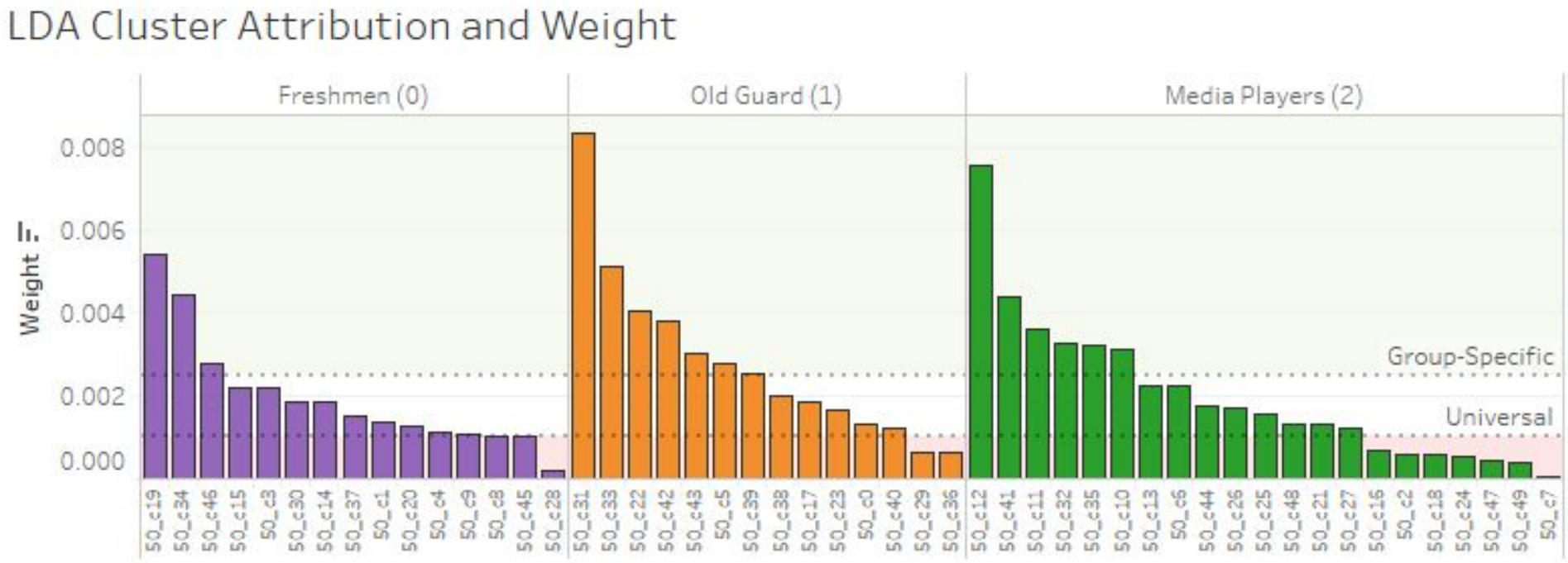
Party Affiliation Across Clusters



Profiling

- Attribution of each topic cluster is then applied to each k-means cluster.
- Weights are calculated by subtracting the average LDA weight across the cluster members from the average across all the members.
- This reveals topics that are important within and between groups.

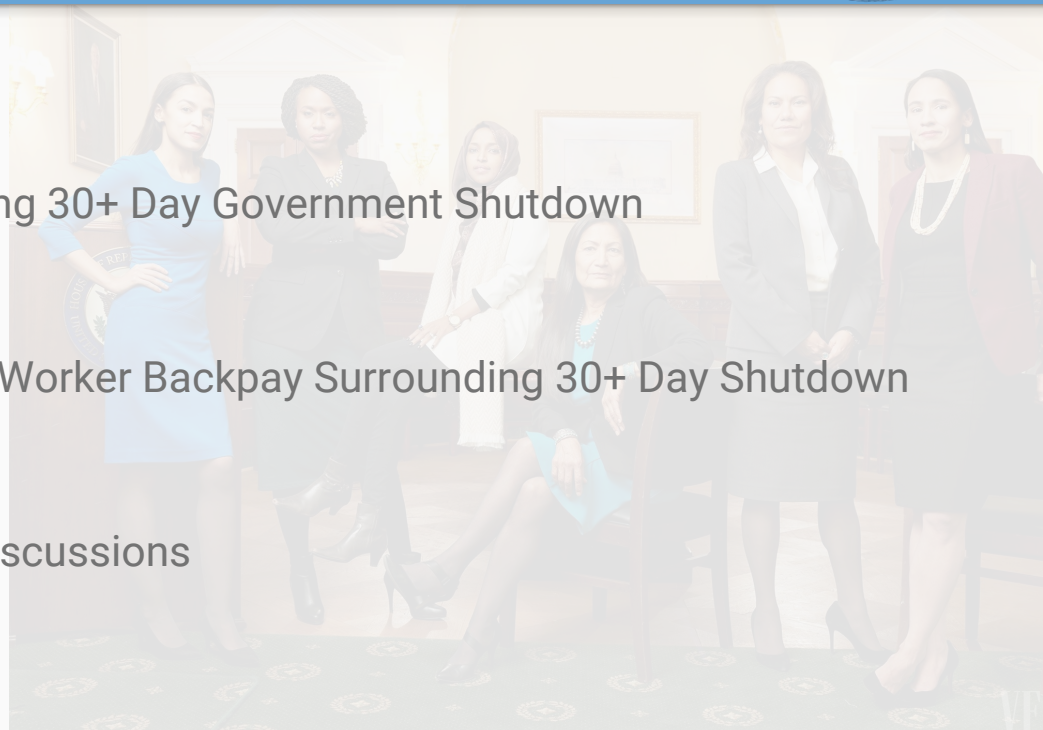
Profiling


LDA Cluster Attribution and Weight
Freshmen (0)
Old Guard (1)
Media Players (2)
Weight
0.008
0.006
0.004
0.002
0.000
Group-Specific
Universal
50_c19
50_c34
50_c46
50_c15
50_c3
50_c30
50_c14
50_c37
50_c1
50_c20
50_c4
50_c9
50_c8
50_c45
50_c28
50_c31
50_c33
50_c22
50_c42
50_c43
50_c5
50_c39
50_c38
50_c17
50_c23
50_c0
50_c40
50_c29
50_c36
50_c12
50_c41
50_c11
50_c32
50_c35
50_c10
50_c13
50_c6
50_c44
50_c26
50_c25
50_c48
50_c21
50_c27
50_c16
50_c2
50_c18
50_c24
50_c47
50_c49
50_c7

Universal Topics (Weight < 0.001)

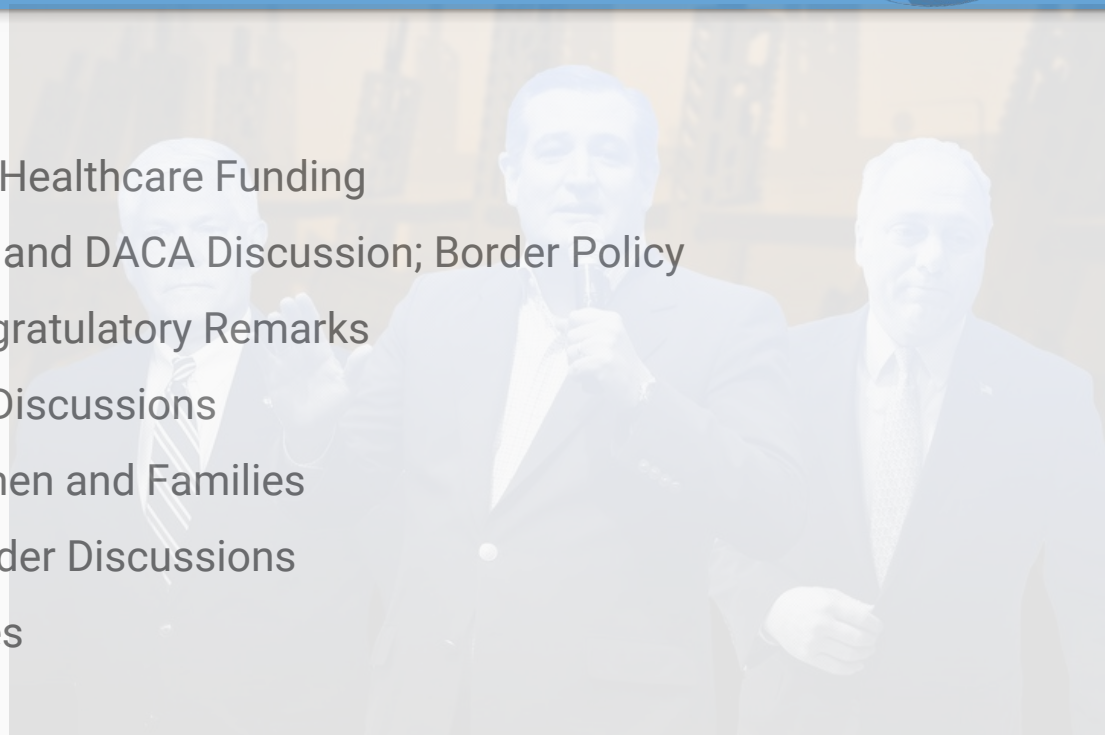
- Freshmen preferred:
  - 50_c28: State and Federal Environmental Policy
- Old Guard preferred:
  - 50_c29: Ohio Job and Trade Growth
  - 50_c36: 9/11 Victim's Compensation Renewal
- Media Players preferred:
  - 50_c24: General Inter-Member Debate
  - 50_c16: Border & Cartel Drugs (Fentanyl)
  - 50_c7: 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 Parkland County, FL

Group-Specific Topics (Weight > 0.0025)



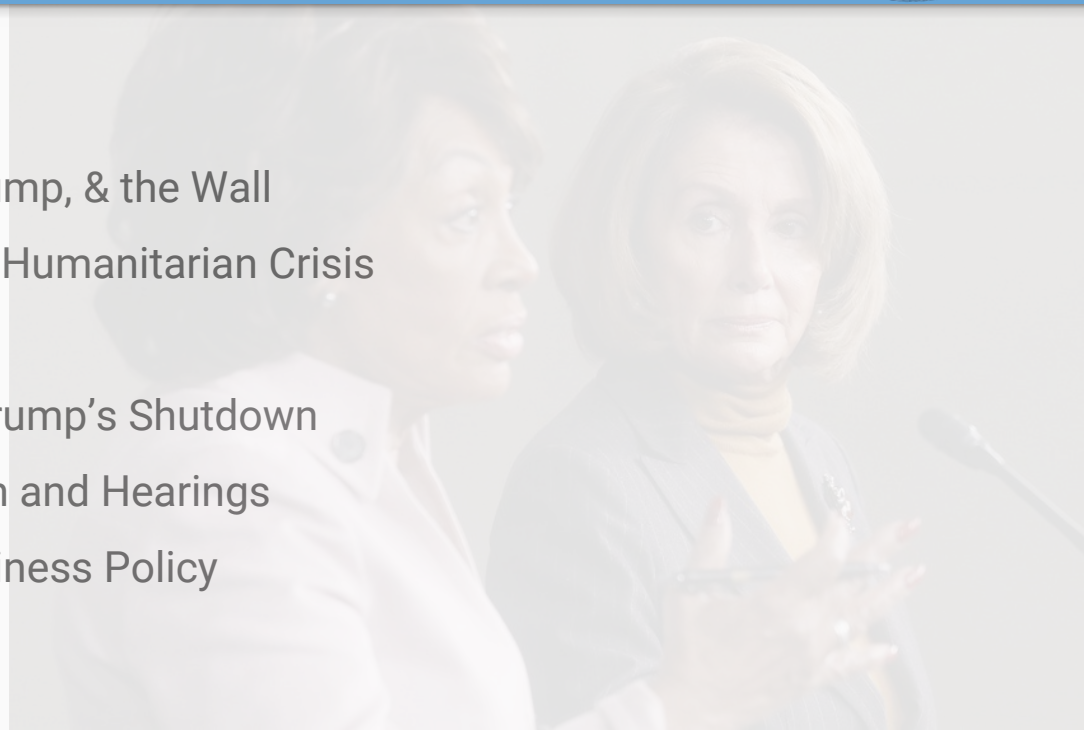
- Freshmen
  - 50_c19: Town halls Surrounding 30+ Day Government Shutdown
  - 50_c34: Demands for Federal Worker Backpay Surrounding 30+ Day Shutdown
  - 50_c46: New England SALT Discussions

Group-Specific Topics (Weight > 0.0025)



- Old Guard
  - 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

Group-Specific Topics (Weight > 0.0025)



- Media Players
  - 50_c12: Border Security, Trump, & the Wall
  - 50_c41: Socialism, Maduro, Humanitarian Crisis
  - 50_c11: Watch Fox News!
  - 50_c32: Fake Emergency, Trump’s Shutdown
  - 50_c35: AG Barr Nomination and Hearings
  - 50_c10: Midwest Small Business Policy

Conclusion

- A Lawmaker's Tweets are a valuable insight into the issues with which they choose to engage. Constituents can use these insights to determine where their representatives stand on various topics
- Some groups discuss more heated topics with a greater level of rhetoric, such as the finger-pointing observed from the media players.
- Some groups are more careful with their posting and are more likely to speak with praise; for example, the congratulatory remarks made by the Old Guard.
- Some topics are universal for members, indicating which topics are the least politically driven.

Thanks!

Image Sources:

https://goo.gl/sNZu9F
https://goo.gl/bUFYKY
https://goo.gl/2byXtM
https://goo.gl/a6DYYK
https://goo.gl/Y4Yv1B