Congressional Twitter Search

SI650 Final Report

Jonathan Hartman
University of Michigan - School of Information
Ann Arbor, Michigan
jonhartm@umich.edu

Julie Gilbert
University of Michigan - School of Information
Ann Arbor, Michigan
juliegil@umich.edu

1 PROBLEM

Since the 2016 United States Presidential Election, there has been an increase in political participation from the public, on both the left and the right[6]. This increase has meant many individuals are new to the playing field, and even the continuing participants have increased their scrutiny of current events.

It is therefore important to have sources of information to inform the population of continuing political participants and getting the new ones up to speed. Our project aimed to give this sample population a place to examine which politicians in Congress are most active in various political theme categories.

2 METHOD

2.1 Data Collection

Data was gathered from OnTheIssues.org[4] through a web-scraping process using python and BeautifulSoup. The reported actions of members of the House and Senate in twenty four different categories were retrieved and saved in .json format.

In order to collect the tweets of members of the 115th United States Congress, a public dataset provided by George Washington University[2] was used. This dataset contained the twitter user IDs and user handles for all congresspeople who maintain a twitter account. These user IDs were used in conjunction with the Twitter API in a python script to download the text and tweet ID of all publicly available tweets from those accounts. The time period of tweets covered from January 27th, 2017 through July 20th, 2018.

2.2 Document Indexing

To find the politicians most related to a given topic, a way of searching all of a given politicians tweets at once was considered. As searching through the entire dataset of 1 million tweets and rating each politician on the results would be too slow, the task was split into two parts. First, tweets were combined into 548 âĂIJdocumentsâĂİ based on which account had sent the tweet. A TF-IDF indexer provided by the python ’Woosh’ library[1] was used to identify important words within this set that would be related to a given politician. A second index was created based on each individual tweet, which included keywords extracted from the body of the tweet, such as mentions and hashtags, as well as the twitter user ID of the sender. As part of twitter’s user agreement requires that applications using the API not store the actual text of any particular tweet, the tweet index only associates these terms with a tweet ID, which can be used in conjunction with the twitter API to retrieve the text of the tweet at a later date.

Table 1: An Example of the Related Terms Used Based on System Queries

Query Term	Related Terms via LSA
economy	growing, grow, growth, booming, thriving, economic, manufacturers, sector, creating
education	cte, colleges, schools, technical, educational, college, districts, pell, charter
corporations	billionaires, millionaires, wealthy, giveaways, corporate, massive, wealthiest, rich, breaks

2.3 Querying

As tweets are, by their very nature, short messages, performing a simple BM25 search on an index of tweets will be more likely to return tweets that simply mention the query term multiple times. To adjust for this, a query expansion based on a latent semantic analysis (LSA) was used on the corpus of tweets as a whole. To do this, a TF-IDF matrix of 10,000 words based on the collection of tweets was made, then a singular value decomposition with a k of 70 was performed. This allowed us to locate, based solely on the content of the collected tweets, words that are more likely to appear in close proximity to a given term. (See Table 1 for examples.)

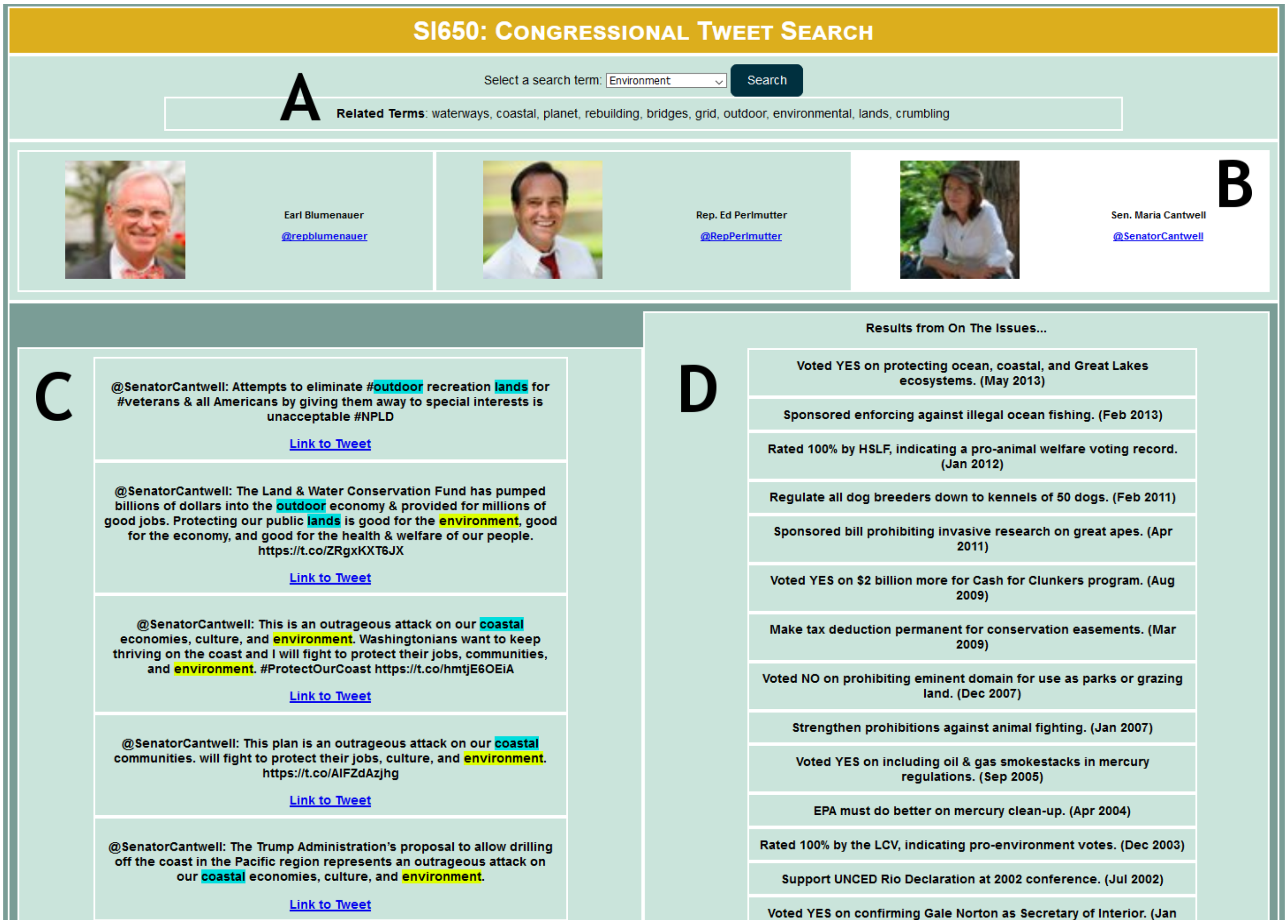
When it comes time to actually search the system for a given term, first related words to the search term from the LSA are collected, and those words are used to expand the query. A BM25F search provided by the python Whoosh library[1] is performed on the smaller index of combined tweets by politician. The ranking returned by this search is, in a general sense, which politicians are most often publicly speaking about regarding a given topic. We can then save the user ID for the top 3 returned accounts for use in the next part of the system. The same method of querying the index is applied to the individual tweet index - the only difference being that the results are restricted to a provided account ID.

2.4 User Interface

Combining all of these, a simple python flask application allows a user to interact with the index. Information not stored in the index is retrieved via the twitter API. This includes profile images, profile links, tweet text, tweet publication dates, and direct links to tweets. Basic functionality connecting the twitter API and the flask application is handled via JQuery’s ajax functions.

For full code and assistant files, please see the public GitHub repository.

Figure 1: Sample Output of the Developed System



SI650: Congressional Tweet Search

Select a search term: Environment

Related Terms: waterways, coastal, planet, rebuilding, bridges, grid, outdoor, environmental, lands, crumbling

Earl Blumenauer
@repblumenauer

Rep. Ed Perlmutter
@RepPerlmutter

Sen. Maria Cantwell
@SenatorCantwell

@SenatorCantwell: Attempts to eliminate #outdoor recreation lands for #veterans & all Americans by giving them away to special interests is unacceptable #NPLD

Link to Tweet

@SenatorCantwell: The Land & Water Conservation Fund has pumped billions of dollars into the outdoor economy & provided for millions of good jobs. Protecting our public lands is good for the environment, good for the economy, and good for the health & welfare of our people. https://t.co/ZRgxKXT6JX

Link to Tweet

@SenatorCantwell: This is an outrageous attack on our coastal economies, culture, and environment. Washingtonians want to keep thriving on the coast and I will fight to protect their jobs, communities, and environment. #ProtectOurCoast https://t.co/hmtjE6OEiA

Link to Tweet

@SenatorCantwell: This plan is an outrageous attack on our coastal communities. will fight to protect their jobs, culture, and environment. https://t.co/AIFZdAzjhg

Link to Tweet

@SenatorCantwell: The Trump Administration's proposal to allow drilling off the coast in the Pacific region represents an outrageous attack on our coastal economies, culture, and environment.

Link to Tweet

Results from On The Issues...

Voted YES on protecting ocean, coastal, and Great Lakes ecosystems. (May 2013)

Sponsored enforcing against illegal ocean fishing. (Feb 2013)

Rated 100% by HSLF, indicating a pro-animal welfare voting record. (Jan 2012)

Regulate all dog breeders down to kennels of 50 dogs. (Feb 2011)

Sponsored bill prohibiting invasive research on great apes. (Apr 2011)

Voted YES on $2 billion more for Cash for Clunkers program. (Aug 2009)

Make tax deduction permanent for conservation easements. (Mar 2009)

Voted NO on prohibiting eminent domain for use as parks or grazing land. (Dec 2007)

Strengthen prohibitions against animal fighting. (Jan 2007)

Voted YES on including oil & gas smokestacks in mercury regulations. (Sep 2005)

EPA must do better on mercury clean-up. (Apr 2004)

Rated 100% by the LCV, indicating pro-environment votes. (Dec 2003)

Support UNCED Rio Declaration at 2002 conference. (Jul 2002)

Voted YES on confirming Gale Norton as Secretary of Interior. (Jan

3 RESULTS

This project resulted in a guided search-retrieval system. To begin, a user chooses a topic from the drop-down menu labelled "A" in Figure 1. In order to properly provide comparable results from OnTheIssues, queries were restricted to the twenty four topics listed via a dropdown. Once selected, in section "B" of Figure 1, the related terms as determined by LSA are shown, along with the three accounts determined to be most relevant to the query. Clicking on the link below a politician's name will take the user to that politician's twitter page. Clicking on a politician's photograph will bring up sections "C" and "D" in Figure 1. Section "C" will provide the text of the top five tweets the selected politician has made about the selected topic, along with an html link to that tweet. Within the text of the tweet, query terms are highlighted (the selected search term in yellow, and any related terms in blue) in order to indicate to the user why these tweets were selected. Section "D" is populated with information on the politician's actions regarding the selected topic as pulled from OnTheIssues.org[4].

4 EVALUATION

Quantitative evaluation for our system is difficult due to two main factors. First, the public tweets are unlabeled for any sentiment or categorization. Second, that both agreement and disagreement between tweets and actions are interesting findings. Agreement shows that the politicians public stances match their political actions. Disagreement shows that the politician makes legislative choices that differ from their public stance. Both scenarios are true possibilities.

Though we recommend that further attempts at evaluation be made, we did compare our system to some available score systems.

Planned Parenthood has a 'Congressional Scorecard', which ranks Congress members based on their agreement on Planned Parenthood's stances when it comes to votes on legislation[5]. A 100% score means that the Congress member always votes on the Planned Parenthood position for items of legislation[5]. The top three politicians that appear in our system when the 'Abortion' category is selected were Kirsten Gillibrand, Lois Frankel and Dianne

Table 2: Precision of Tweets Overall

Precision@K	Score
Precision@5	0.740
Precision@10	0.710
Precision@20	0.695

Feinstein. All three individuals do appear in the Planned Parenthood Scorecard system, and each of them have received a 100% score[5].

The League of Conservation Voters has a scorecard system as well, calculated by counting a member's pro-environment votes divided by the total number of environment legislative votes considered for a given year[3]. A higher score is equivalent to more pro-environment votes. The three top individuals who appear in the system when the 'Environment' category is selected are Earl Blumenauer, Ed Perlmutter and Maria Cantwell. They are all present in LCV's scorecard system, and received scores of 96%, 86%, and 92%, respectively[3].

The American Civil Liberties Union (ACLU) provides a scorecard[7] based on how often a politician votes in alignment with the ACLU on civil rights and civil liberties over the past legislative session, which can be taken to identify congresspersons with a stronger Civil Rights record. For the query "Civil Rights", our system returns Gwen Moore, Adam Smith and Mark Takano, for which the ACLU scored 93%, 96%, and 93% respectively.

It is interesting that the top politicians that show up in our system for these topics also receive high scores in the selected scorecard systems. It suggests that those that speak often about these topics are often 'on the same side' as the large institutions that are often the face of these topics.

Although without a fully labeled set of tweets, it's not possible to calculate recall, we can manually check the Precision of the returned results by looking at the top tweets returned by our system and scoring them based on their relevance to the search term. This was done for ten randomly selected account/key term pairs the results of which are in Table 2. The precision of tweets varies greatly by topic, however. This is likely a result of the query expansion adversely affecting queries when the component words of the query have a more ambiguoius meaning. For example, the query "Environment" expands with terms such as "coastal", "planet" and "outdoor", whereas "Government Reform" adds terms like "tax", "partisan", and "shutdown" - terms related to the individual terms "Government" and "Reform", but not to the concept of "Government Reform".

5 WHAT WAS LEARNED

Primarily, the most difficult hurdle to get over was figuring out how to index and search tweets effectively. The first few iterations of the system worked on a single term search, which essentially only returned tweets which used the same term multiple times. The LSA approach to query expansion, though with clear drawbacks, does seem to give both more varied and relevant results in most cases.

Additionally, there's an extra level of difficulty added in when dealing with documents provided by a large and diverse group of people. Some accounts tweet up to twenty times a day, whereas others may tweet barely a dozen times a year. These lesser used accounts often serve a different purpose, mainly to post photo ops or legislative achievements rather than publicly state a political stance. As a result, it is questionable whether some accounts should be included in the dataset. There is also a difference in how some users use the platform in general - in one particular instance, the majority of tweets by an account are photos of posts from the politician's Facebook page. As the system is not equipped to perform optical character recognition on images included in tweets, the content of these messages is not accounted for in the current system.

Table 3: Precision of Tweets By Topic

Query: Environment	
Precision@5	0.733
Precision@10	0.733
Precision@20	0.750
Query: Government Reform	
Precision@5	0.067
Precision@10	0.133
Precision@20	0.100

6 NEXT STEPS

The most important next step for improving this project would be to work to classify the tweet library in order to establish better evaluation metrics. Manual labelling would need to be applied to a subset of the collected tweets in order to determine the proper key word query it would most apply to, which would allow us to calculate recall and in turn, mean average precision.

The system currently indexes every tweet published by a member of the most recent congressional session, which results in quite a sizable index and by extension, long query times. The current implementation takes between 2.5s and 6.8s to retrieve all of the relevant tweets. A certain amount of this time is certainly a result of having to retrieve tweet data from the twitter API, however the query alone averages around 1.4s. Reducing the time span of tweets in the dataset to the dates of the last legislative session will likely reduce the query time, as well as shrinking the index (currently about 1GB) to a size that can be easily hosted on an inexpensive service.

As this current system is only concerned with the amount a particular politician publicly speaks on an issue, another direction to take this project would be to apply sentiment analysis to the tweets. This could attempt to identify, based on a collection of tweets, whether or not a particular account is 'supportive' or 'non-supportive' for each topic category.

Other areas for expansion could be to add an element of network analysis to examine the sentiments of each politician, along with who they follow and who is mentioned in tweets. This would provide additional information to inform the 'support' or 'non-support' tag assigned by sentiment analysis. We could also add the ability for this system to update based on election results, search larger periods of time for tweets, or look at additional political hierarchies, such as local government.

REFERENCES

[1] Matt Chaput. 2007–. Whoosh: a fast, pure Python search engine library. (2007–). https://bitbucket.org/mchaput/whoosh/wiki/Home [Online; accessed 2018-12-14].

[2] GWU Libraries Dataverse. 2018. 115th U.S. Congress Tweet Ids. (2018). https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UIVHQR

[3] League of Conservation Voters. 2018. LCV National Environment Scorecard. (2018). http://scorecard.lcv.org/

[4] OnTheIssues.org. 2018. OnTheIssues.org - Candidates on the Issues. (2018). http://www.ontheissues.org/

[5] Planned Parenthood. 2018. 2018 Congressional Scorecard. (2018). https://www.plannedparenthoodaction.org/congressional-scorecard

[6] Laura Sydell. 2017. On Both The Left And Right, Trump Is Driving New Political Engagement. (2017). https://www.npr.org/2017/03/03/518261347/on-both-left-and-right-trump-is-driving-new-political-engagement

[7] American Civil Liberties Union. 2018. ACLU Congressional Scorecard. (2018). https://www.aclu.org/sites/default/files/field_document/acl18002_legislative_report_card_v2.pdf