CONGRESS ON TWITTER

@JohnMcElvenny • @SloanNietert • @AnnieWalker || #DataScience #Spring2018

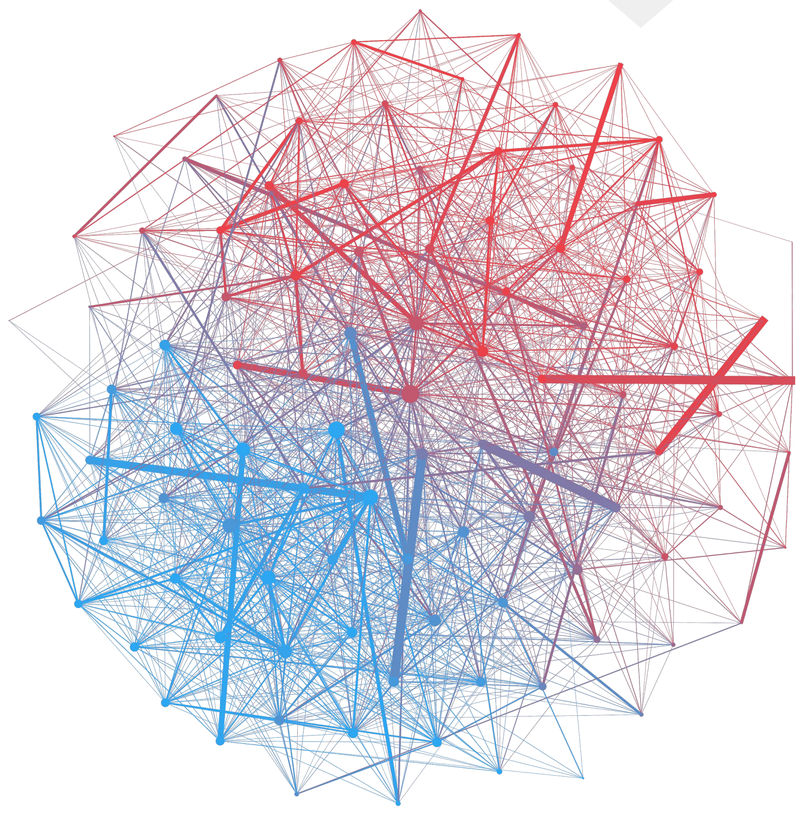
538 MEMBERS OF CONGRESS

Data Collection

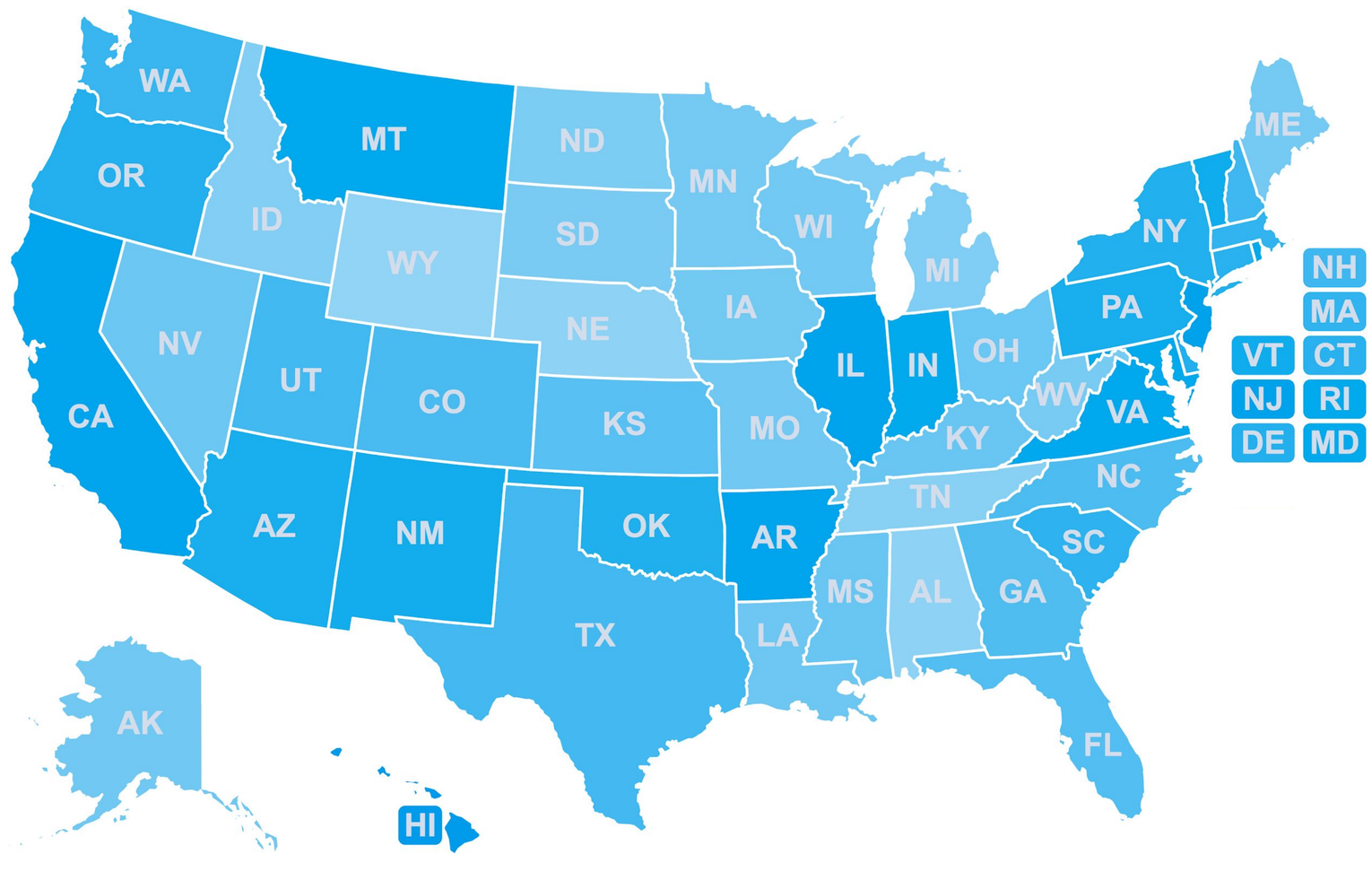
The goal of this project was to build a robust system for collecting and analyzing Twitter data from members of Congress. Since no open source tool currently exists to publicly aggregate congressional tweets, our first step was to collect the data. We developed a Python script which uses the official Twitter API to download the desired tweets, and we implemented an update feature to avoid redundantly repopulating the dataset. Because of memory and rate limit induced time constraints, we ran our scripts on a high-performance CloudLab node. In total, 1,481,457 tweets were collected from 623 Twitter handles.

MOST CONNECTED SENATORS

MITCH MCCONNELL (0.759)
JOHN MCCAIN (0.741)
CHUCK SCHUMER (0.308)
ELIZABETH WARREN (0.225)
CORY BOOKER (0.206)
SHELDON WHITEHOUSE (0.189)
DICK DURBIN (0.175)
EDWARD MARKEY (0.082)



2018 Hot Topic: Immigration



LDA Topic Modeling

The coloring of the states in this illustration represents the amount of Twitter discussion this year among legislators surrounding the topic of immigration. Specifically, we used the collected tweets to create a Latent Dirichlet allocation (LDA) topic model and selected a strong topic of interest which included keywords such as “immigration”, “arizona”, and “trump.” The color of a state is directly proportional to the sum of that topic’s score over all 2018 Tweets from the state’s Senators.

Network Analysis

With the collected tweet corpus, we constructed a graph of Senators with edges introduced by retweets, replies, and mentions. Specifically, the network was defined by the following adjacency matrix:

$$\begin{aligned} A_{ij} = &(\#\text{ times } j \text{ retweeted } i) \\ &+ (\#\text{ times } j \text{ mentioned } i) \\ &+ 0.5 \cdot (\#\text{ times } j \text{ replied to } i) \end{aligned}$$

The visualization to the left was created using Gephi with proportional edge weights and colors corresponding to GovTrack ideology scores. The nodes were placed according to the Fruchterman Reingold layout algorithm, and a clear partisan split emerged. Additionally, we used NetworkX to rank the Senators by eigenvector centrality (this method is based on the recursive principle that central nodes are those which are connected to other central nodes).

Results

Our team found that partisan behavior is clearly manifested on Twitter, as can be seen with the network visualization. Our graph analysis also demonstrated that the most powerful Senators tend to be well-connected on Twitter. In particular, our list of most central legislators included both party leaders and senior members of both parties. Finally, topic analysis gave reasonable results, with states along the Mexican border displaying particular concern.

In the future, our team is interested in analyzing topic trends over time and in eventually making the dataset publicly available for other researchers.

1,481,457 INDIVIDUAL TWEETS

35,526 SENATE NETWORK CONNECTIONS

7.1 GB OF TWITTER DATA