Project Fletcher Summary
News Trends in the Trump Era

Dana Lindquist
November 19, 2018

Project Summary

Our attitudes about our world, politics and business have changed since the 2016 election of Donald Trump to the presidency of the United States. My goal for this project was to quantify these trends through the news written about the country.

Data

The New York Times has an API for downloading articles from their archive. The API provides json files and I used jq to query the json files to extract

- Headline
- Snippet (a very short description of the article)
- Date
- News desk

I collected data from January 1, 2015 to November 16, 2018.

The data was then filtered on the news desk to remove articles from categories such as Classified on Sports, leaving only world news, politics. The news desks that were used were: Business, Foreign, NewsDesk, National, Politics, U.S., U.S. / Politics, U.S. / Election 2016, Washington, World / Europe, World / Middle East, World / Asia Pacific, World / Africa, World / Americas. 274606 articles were downloaded. After filtering, 84,582 remained.

Models

The process for the analysis can be summarized as

1. Create a document by combining the headline and snippet. This gives a descent collection of words to describe the article.
2. Using scikit learn CountVectorizer or TfidfVectorizer, break the documents into words, removing high and low frequency words and changing all to lower case. Ngrams in the range of 1-3 words were added. A sparse matrix with word counts for each document was generated.
3. Bucket the documents based on the word frequencies in the documents. Two methods were used:
   A. Latent Dirichlet Allocation (LDA) was used to bucket the bag of words, creating a probability distribution of each document fitting in each bucket.

B. Non-Negative Matrix Factorization (NMF)

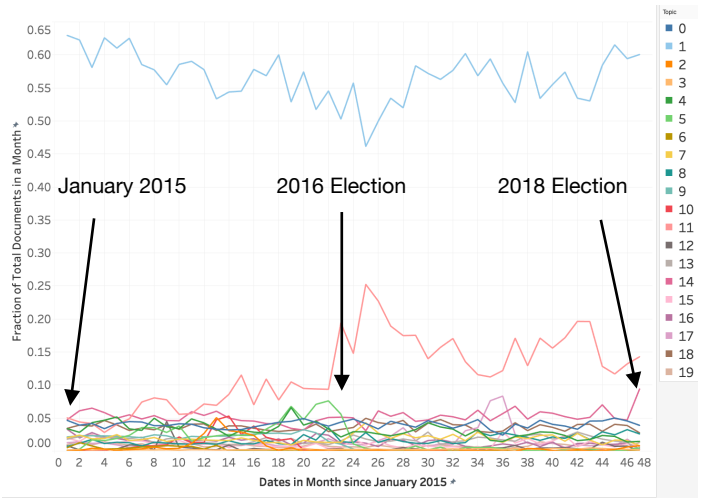
4. K Means was used to cluster like distributions from the buckets into topic clusters. To find the main topics for the clusters the I looked at the documents nearest the cluster centers.
5. To examine the trend over time, the percentage of documents in a month was plotted over the time range of the data.

Results

LDA is a very compute intensive method and was hard to work with. It was able to generate results at a low number of buckets but as the number of buckets increased it was a challenge to get it to converge.

NMF was able to generate interesting results. 50 NMF buckets were used with 20 K-Means clusters.

Shown here is a plot of the 20 clusters as a function of time. Each line represents the fraction of the articles in that month that fall into that topic.



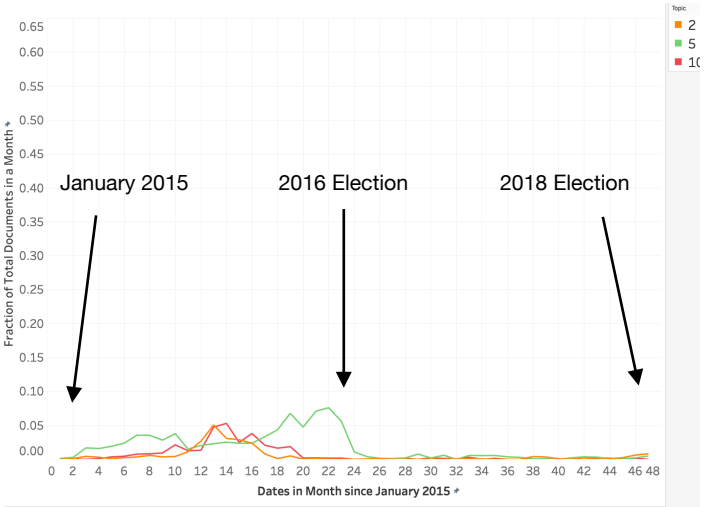
There is one dominant topic, #1. Looking at the documents in this topic this feels like a cluster of anything that was left. More experimentation with the number of NMF buckets might refine the topics in this cluster.

Some of the clusters are of particular interest. The following clusters have a direct relationship to the 2016 Election.

#2 is about Ted Cruz and John Kasich.

#5 is about Hillary Clinton

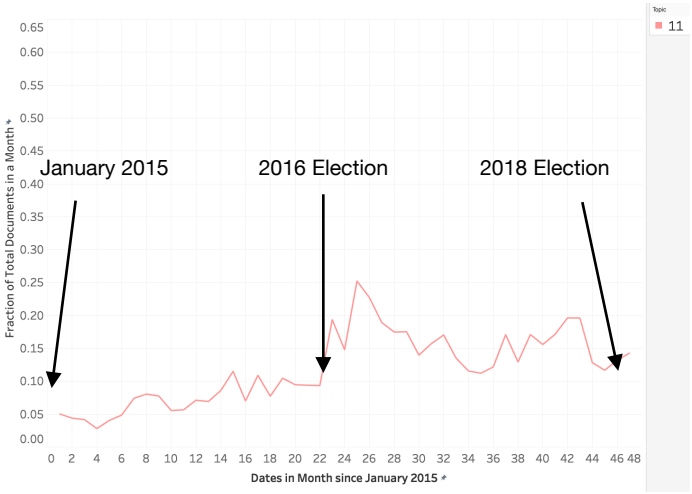
#10 is about Bernie Sanders



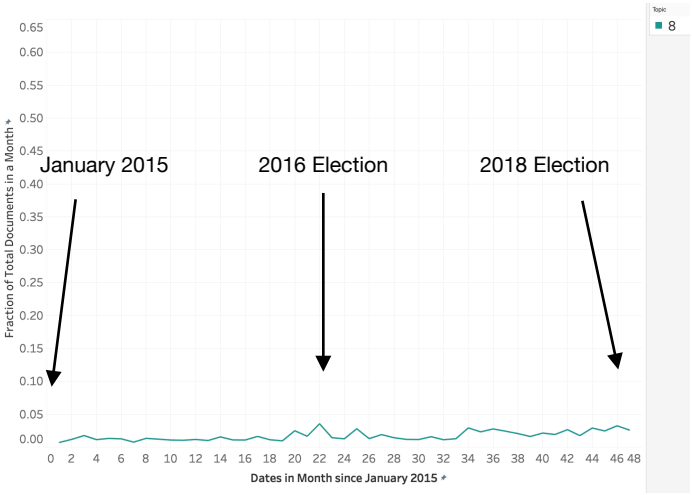
It's interesting to see how similar the primary contenders were and how dramatically the Hillary Clinton topic

drops off after the election.

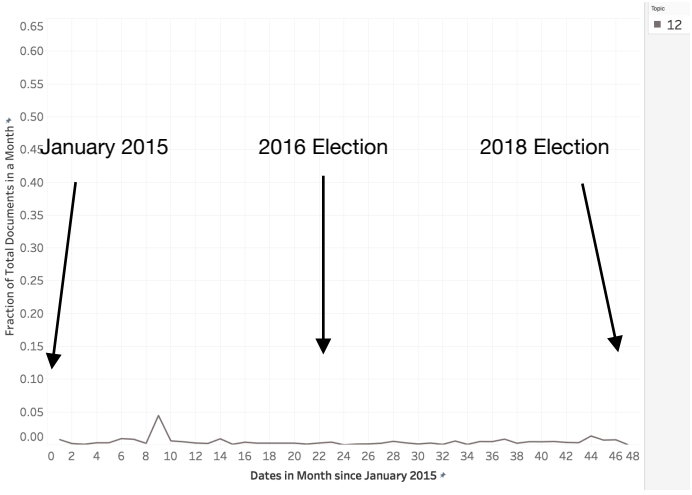
Topic #11 is dedicated to Donald Trump and you can see how it has increased since the 2016 Election.



Topic #8 is about Women, Harvey Weinstein, Me Too with a sprinkling of women's fertility. This has been on a gradual rise over the 4 years of this data.



And finally, Topic #12 is dedicated to Pope Francis. The peak in August 2015 was his trip to Cuba. Clearly this was an important visit as the analysis created a separate topic for this trip.



Conclusions and Future Extensions

I believe that increasing the number of NMF/LDA buckets would help pull out more information from this dataset. It would be nice to get more refinement from Topic 1. It is computationally expensive to use LDA so I would continue the analysis with NMF.

I also believe there is more work that can be done in word filtering to help refine the important topics. Remove the similarities between the documents could improve the analysis as well.