THE UNIVERSITY OF TEXAS AT AUSTIN

MIS 184N: SOCIAL MEDIA ANALYTICS

ASSIGNMENT 2

Analyzing U.S. Presidential Speeches Using Latent Dirichlet Allocation and Topic Modeling

Authors
Daxi CHENG, Corey HAINES, Daniel QUIJANO, Jinru SU, Reece WOOTEN

February 20, 2018

Contents

Introduction

In this assignment we examine 632 U.S. presidential speeches and apply topic modeling techniques to determine underlying topics amongst the speeches as well as how similar some presidents are with each other. Given these speeches cover 240 years of U.S. history, we knew estimating the number of topic *a priori* would prove challenging. Therefore, as a starting basis for Latent Dirichlet Allocation (LDA) we started with 5 topics, rationale being all speeches would touch upon Constitutionality of laws, war, the economy, human rights, and/or public welfare (fortunes). However, we recognized that contained within these topics would likely be important subtopics, such as national security in war, commerce in the economy, etc. Therefore, we also empirically tested what would be the best number of topics for LDA by running our analysis from 5 topics to 25 topics. We noticed that after roughly 10 topics little incremental value was added by specifying more topics. Most presidential speeches fell within one or two of 10 different topics. Thus, we decided to use this as the basis for our analysis moving forward with the rest of our analysis.

Latent Dirichlet Allocation

After running LDA on the speeches corpus, specifying 10 topics and 1000 passes, the following topics discovered with format [Topic #: Topic (word loading)]:

> Topic 1: World Wars and Economic Recovery (world, war, peace, production, new [deal])
>
> Topic 2: Family and Health (child, family, health, school, community)
>
> Topic 3: Cold War (Soviet, peace, war, military, defense)
>
> Topic 4: Slavery (slavery, territory, constitution, free, south)
>
> Topic 5: Labor (power, work, purpose, industry, life)
>
> Topic 6: International Trade (treaty, tariff, international, foreign, commerce)
>
> Topic 7: Government Institutions (government, state, president, tax, policy)

Topic 8: Communism (Vietnam, communist, freedom, people, hope)

Topic 9: Nationalism (American, great, united, nation, power)

Topic 10: Founding of Nation and Inalienable Rights (congress, British, treasury, debt, right, law, country)

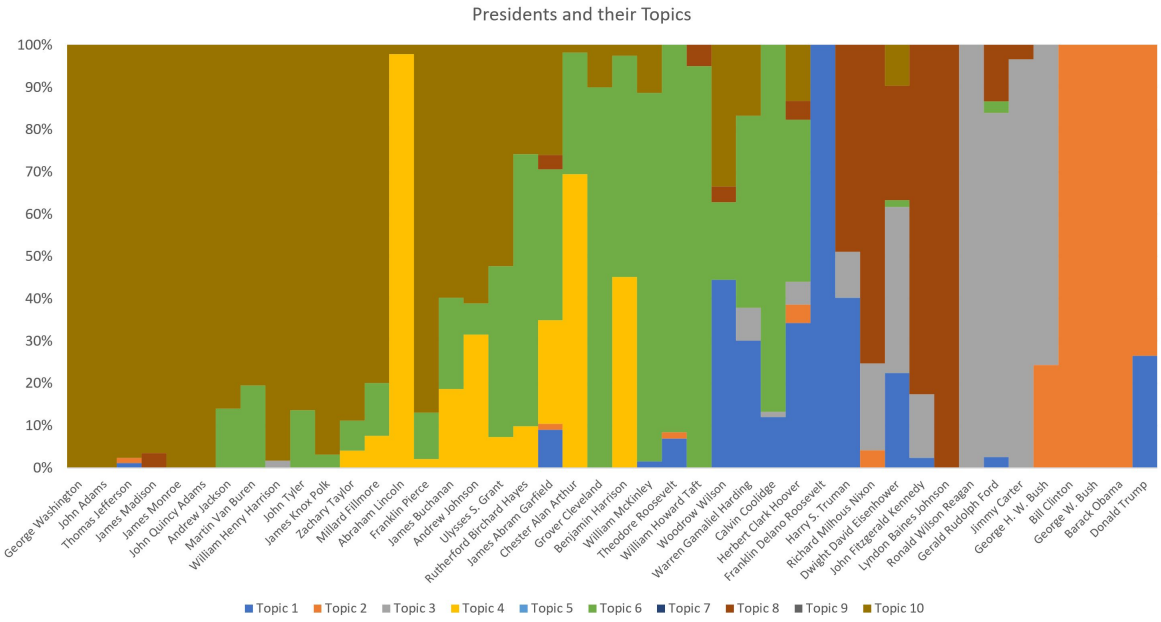


Figure 1: Topic Distribution Over Time

As can be seen above, the topics discussed by presidents change throughout time. In the beginning much of the focus is on establishing the Constitution, government, and rights of the citizens. Then, both slavery and international trade become pressing matters as the nation expands its reach with industrialization in the North and grows its agrarian economy in the South. At the turn of the 19th century the topics turn towards the World Wars and Economic recovery. This is not surprising as the Great Depression was between World War I and World War II. Additionally, there was a post World War II economic boom as American industry blossomed. As the United States and Russia grew into the two world powers, the topics also changed to the Cold War, communism, and nationalism. There was competition with Russia and pride in the nation. Lastly, topics have shifted to

reflect concerns for the every day citizen. As society becomes mroe isolated with technology, people become more concerned with themselves and less about the nation as a whole. This is reflected in recent speeches by past presidents.

President Donald Trump

Speech Similarity with Past Presidents

In order to determine which past presidents had the greatest similarity with President Donald Trump, we followed two approaches [both based on cosine similarity]. First, we took the speeches corpus and vectorized it by taking the entire set of words across all speeches, and then, for each president determined whether each word was present in their individual speech corpus. Subsequently, we populated the cosine similarity matrix with each pairwise comparison between presidents, i.e., found the cosine angle between each pairs of presidents. Using this approach, the three presidents with the cosine angles closest to 1 were identified as being most similar [in their speeches] to Donald Trump.

1. Bill Clinton: 0.734771
2. Barack Obama: 0.708722
3. Lydon B. Johnson: 0.69874

The second approach we took involved taking the topic loadings on each president vector and finding the cosine similarity between each presidents' topic loading vector and President trump's topic loading vector. Using this approach, the three presidents with the cosine angles closest to 1 were:

1. George W. Bush: 0.94
2. Bill Clinton: 0.94
3. Barack Obama: 0.94

Both methods revealed Donald Trump being close to Bill Clinton and Barack Obama. On the surface this may seem surprising. However, upon further consideration it makes sense given both Bill Clinton and President Trump have emphasized the economy, and President Trump continually talks about how bad Obama's policies were [thus mentioning them]. We

believe that George W. Bush is probably closer to President Trump than Lyndon B. Johnson because George W. Bush was the last conservative president in the White House before Trump.

Speech Patterns through Presidency

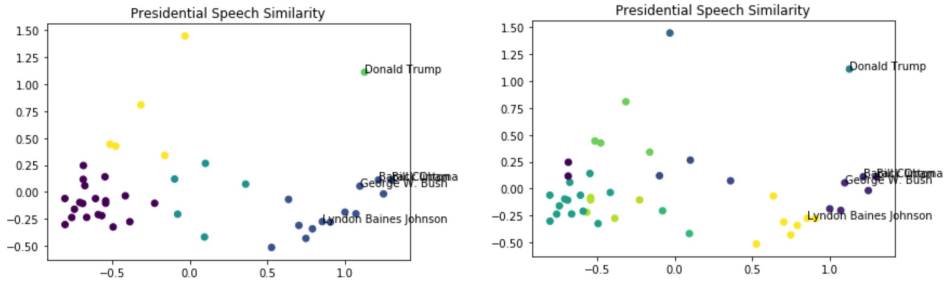
In order to determine what topics President Trump has focused on during his presidency, we ran LDA with 5 topics and 1000 passes on his speech corpus. Our results showed that President Trump has only focused on two topics:

> Topic 1: Nationalism and Immigration (glorious, triumph, righteous, success, shine, unstoppable, visa, protection, refugee, banned, immigrant, prejudice)
>
> Topic 3: Tax Reform and Job Creation (tax, reform, billion, million, regulation, cut, business, work, job, [bring] back, money)

The other topics barely loaded onto any of his speeches, which after investigation was unsurprising as there was much overlap with the two aforementioned topics.

K-means Clustering and Multidimensional Scaling

To verify our similarity results obtained through LDA and topic modeling, we also performed K-means clustering on the cosine similarity matrix. Because there are only 44 presidents, it concerned us that having 10 clusters would result in non-meaningful groups. Therefore, we decided to run K-means clustering with both 5 and 10 clusters, thereby increasing our chance of discovering an interesting grouping.



(a) (b)

Figure 2: K-means clustering on cosine similarity with (a) 5 clusters and (b) 10 clusters

As seen above, all four presidents identified in task C were part of the same cluster. However, President trump was in his own cluster when running K-means with 5 or 10 clusters. Therefore, while some presidents are more similar in the content of their speeches with President trump than others, President Trump is a clear outlier. This can be further emphasized when looking at an MDS map of only President trump and those presidents with high cosine similarity scores.

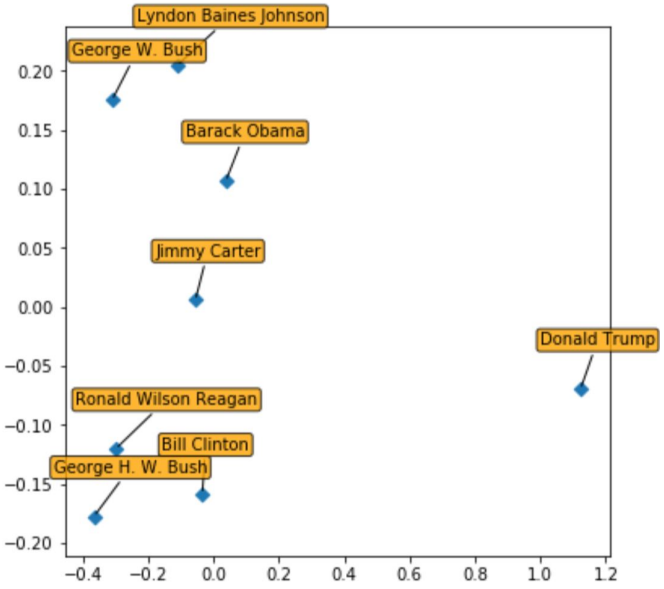


Figure 3: President Trump is an Outlier even Amongst Those Most Similar to Him