Topic Modeling

Ajul Thomas
u3253992@uni.canberra.edu.au

Faculty of Science and Technology
University of Canberra
ACT, Australia

Abstract

This assignment demonstrates the use of Topic Modeling techniques to analyze the State-of-the-Union speeches corpus. The report includes observations on how various subjects have shifted over time with historical events.

1 Introduction

The primary objective of this course activity is to obtain a deeper understanding of topic models. Understanding topics discussed over the years in the State-of-the-Union speeches corpus, which contains speeches dating back to 1790. This research also aims to uncover the hidden patterns within the data through statistical methods such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA).

2 Methodology

2.1 Dataset and Pre-processing

The dataset is downloaded from the unit's course page and is also available for download from the website here. The dataset contains 226 records with year and speech columns. The dataframe, once loaded, has undergone a series of pre-processing steps such as removal of punctuations, tags and stopwords, tokenization and lemmatization. Lemmatization is the process of reducing words to their base form (lemma). It helps in improving text processing tasks[2].

2.2 Generating TF-IDF scores

After lemmatization, the processed documents are used to generate a dictionary, which is then used to create the corpus object representation. The generated corpus object is used as input to get the TF-IDF(Term Frequency-Inverse Document Frequest) score for further modeling. TF-IDF score is a statistical measure used in natural language processing and information retrieval tasks to evaluate the importance of a word in a document relative to a collection of documents(corpus)[1].

2.3 Topic number identification for models

The current analysis uses mathematical techniques such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) for topic modeling. In order to generate a model, we have to supply an optimum number of topics to be identified from the corpus. This is done by calculating the coherence score for each model over a range of values. The value for which the coherence score is highest is used for further processing.

3 Results and Analysis of LSI Model

3.1 The selection of topic number

Based on the results shown in Figure1 the LSI model achieves the highest score at 6 topics. This suggests that 6 topics would provide the optimal balance between model complexity and topic interpretability.

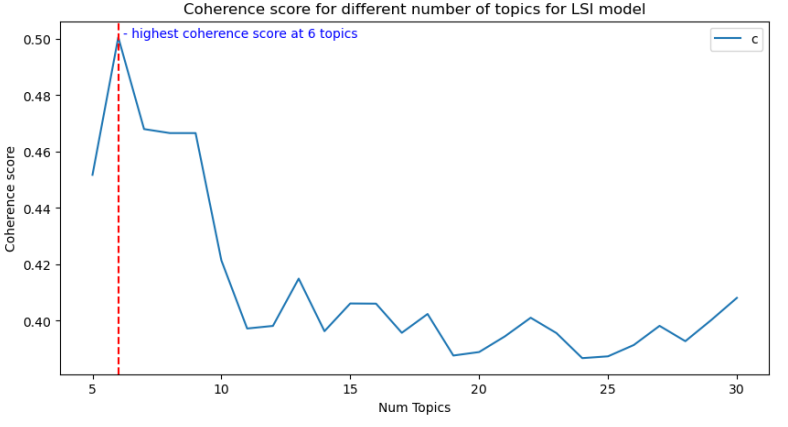

Figure 1: Coherence score versus number of topics for LSI model.

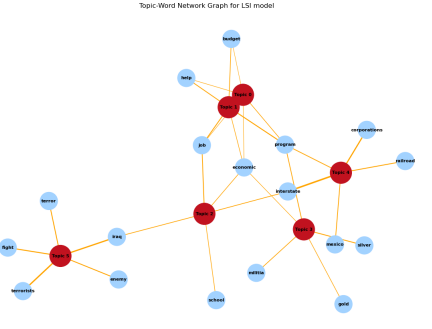

Figure 2: Graph model of identified topics for LSI model.

3.2 Topic Annotation and interpretations

Topic modeling of State-of-union speech corpus using LSI model with 6 topics provided the following output, which I have annotated with a descriptive name. The same is illustrated using a graph network in Figure2.

```
Topic 0: Economic Policy and Domestic Spending
-0.114*"program" + -0.087*"help" + -0.082*"job" + -0.080*"
    economic" + -0.075*"budget" + -0.063*"treaty" + -0.059*"
    let" + -0.058*"tax" + -0.058*"mexico" + -0.057*"billion"
```

This topic focuses on economic strategies, government programs, job creation, and budget considerations. The words "Mexico" and treaty suggest an international economic policy dimension.

```
Topic 1: Social Investment and Budget Allocation
-0.173*"program" + -0.157*"job" + -0.147*"help" + -0.124*"
    budget" + -0.102*"economic" + -0.099*"spend" + -0.096*"
    billion" + -0.094*"school" + -0.093*"percent" + -0.088*"
    let"
```

Very similar to Topic 0, but with a stronger emphasis on spending, education, and social support programs. The focus seems to be on government investment in societal development.

```
Topic 2: Education, Security, and Conflict
0.148*"job" + -0.127*"interstate" + 0.125*"iraq" + -0.111*"
    economic" + 0.105*"school" + -0.104*"farm" + 0.104*"
    terrorists" + -0.102*"industrial" + 0.095*"children" +
    -0.092*"program"
```

A complex topic mixing education, national security, and economic concerns. This topic doesn't seems to have captured a real topic. This could be an incorrect grouping of words.

```
Topic 3: Monetary Policy and International Relations
-0.155*"silver" + 0.136*"program" + 0.125*"militia" +
    -0.105*"gold" + 0.092*"economic" + -0.086*"iraq" +
    -0.086*"circulation" + -0.084*"arbitration" + 0.083*"
    soviet" + 0.083*"gentlemen"
```

This topic seems to capture discussions about monetary metals, international arbitration, and Cold War-era geopolitical tensions. The mix of terms suggests complex economic and diplomatic negotiations.

```
Topic 4: Infrastructure and Corporate Regulation
0.269*"interstate" + 0.211*"corporations" + 0.155*"railroad"
    + -0.143*"program" + -0.139*"mexico" + -0.126*"soviet" +
    0.095*"combinations" + -0.094*"communist" + -0.092*"
    treaty" + 0.091*"corporation"
```

Focuses on transportation infrastructure, corporate regulations, and international economic relationships. The presence of "interstate" and "railroad" indicates infrastructure development discussions.

```
Topic 5: War, Terror, and National Defense
0.236*"terrorists" + 0.226*"iraq" + 0.179*"fight" + 0.164*"
    terror" + 0.160*"enemy" + 0.137*"al" + 0.136*"terrorist"
    + 0.136*"iraqi" + 0.134*"enemies" + 0.122*"japanese"
```

Centered on national security, counter-terrorism, and military conflicts, with specific references to Iraq and implied historical conflicts (e.g., "japanese").

4 Results and Analysis of LDA Model

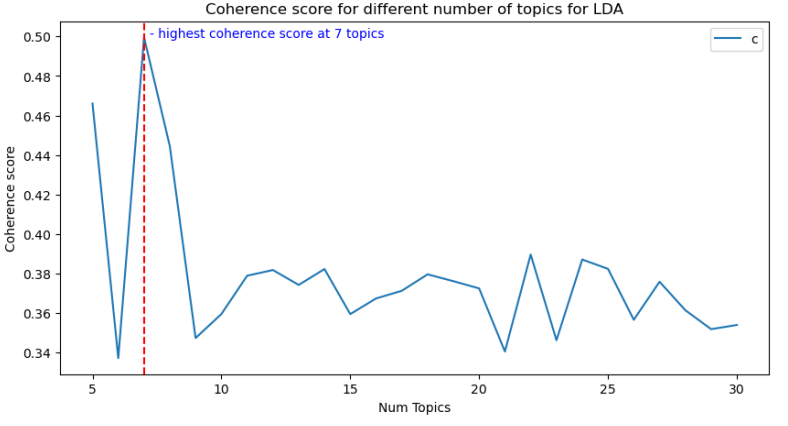

Figure 3: Coherence score versus number of topics for LDA model.

4.1 The selection of topic number

As shown in Figure-3, the LSA model achieves the highest score with 7 topics, indicating that this number of topics offers the best balance between model complexity and topic interpretability.

4.2 Topic Annotation and interpretations

The annotated topics that were modeled using LDA seem to make more sense in a human context compared to the LSI topics. The figure4 shows the wordcloud of topics modeled using the LDA technique.

4.2.1 Topic 0: Cold War Economic Policy

```
-0.114*"program" + -0.087*"soviet" + -0.082*"communist" +
    -0.080*"economic" + -0.075*"budget"
```

This topic seems to be centered around the economic and foreign policies during the Cold War era. I would classify this as a moderately strong topic capture as the words clearly evokes the economic and geopolitical context of the time.

Figure 4: Coherence score versus number of topics for LDA model.

4.2.2 Topic 1: Economic Infrastructure and International Relations

```
-0.114*"interstate" + -0.087*"job" + -0.082*"program" +
    -0.080*"spain" + -0.075*"corporations"
```

This topics explores areas of economic development, job creation and probable international economic co-operation. The words does not have strong coherence and fails to effectively capture a human concept.

4.2.3 Topic 2: Vietnam War and Budgetary Concerns

```
-0.114*"help" + -0.087*"program" + -0.082*"vietnam" +
    -0.080*"budget" + -0.075*"billion"
```

Clearly relates to the Vietnam War period, focusing on governmental support, program funding, and the significant financial implications of the conflict. This one has better topic capture and better coherence. The words within the topics are better related to each other and could derive at a particular period in US history.

4.2.4 Topic 3: Regional Economic Interests - Gulf Coast

```
-0.114*"mexico" + -0.087*"program" + -0.082*"gentlemen" +
    -0.080*"oil" + -0.075*"texas"
```

This could be referring to economic interactions with Mexico, with special focus on oil industry and Texas region. This is a moderately good capture, and the words are sensibly linked to each other.

4.2.5 Topic 4: International Economic Diplomacy

```
-0.114*"mexico" + -0.087*"program" + -0.082*"treaty" +
    -0.080*"economic" + -0.075*"minister"
```

This topic focuses on international economic negotiations, particularly with Mexico, involving diplomatic treaties and ministerial-level discussions. the topic capture is comparatively good and the words collectively are capable of identifying a idea/concept.

4.2.6 Topic 5: Economic Development and Job Creation

```
-0.114*"program" + -0.087*"economic" + -0.082*"mexico" +
    -0.080*"help" + -0.075*"job"
```

Here, the topics is similar to Topic4 and emphasizes economic development programs, job creation, and potentially international assistance or collaboration. It's has a moderate to strong topic capture and words clearly indicates a focus on economic growth, employment, and supportive government programs.

4.2.7 Topic 6: Social Support and Economic Mobility

```
-0.114*"help" + -0.087*"job" + -0.082*"let" + -0.080*"school"
     + -0.075*"currency"
```

This is a broad topic touching on social support, education, employment, and economic opportunity. It has a low to moderate coherence. While the words suggest themes of economic mobility and social support, the connection feels slightly less defined compared to other topics.

In summary, I find some topics showing strong thematic coherence, while others appears to be more fragmented. The model has well captured significant events, concepts, and political and economic areas related to the USA.

5 Decade Summarization

5.1 Decade summarization algorithm

I have made use of both LDA and LSI techniques to extract the relevant topics from the speeches in each decade. However, I have only used the results from the LDA topics as they seem to better explain the historical events.

The logic used involves filtering the lemmatized speech for each decade into separate dataframes and pass them to the custom function called `get_topics` to extract the topics relevant to that decade.

The `get_topics` function performs topic modeling on a DataFrame of lemmatized speech text using both LSI and LDA. It first creates a dictionary and corpus from the lemmatized text, applies TF-IDF transformation to the corpus, and then generates topics using both LSI and LDA models with 6 and 7 topics, respectively. The function returns the top 5 words for each topic from both LSI and LDA models.

5.2 1900s

From the historical facts, the 1900s were marked by rapid industrialization, social upheaval, and significant geopolitical changes. In US, racial violence and lynching were a big social issue causing deep rooted racial tensions. During this time period, there were a huge number

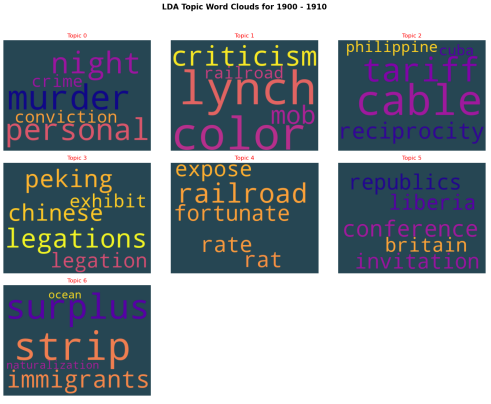

Figure 5: Topics extracted from speeches from 1901 to 1910.

of immigrants coming to America, which led to serious discussions around naturalization and assimilation. The consequences of Spanish-American war were also major topics of this time period, where we can see keywords such as `Cuba`, `Philippine`.

The words `peking` and `chinese` refer to the boxer rebellion in China, was an anti-foreign, anti-imperialist, and anti-Christian uprising in North China. Meanwhile, economic growth driven by industries like railroads was often accompanied by exploitation and corruption, sparking social and political criticism, which were quite evident from the keywords `surplus`, `railroad`, `expose`, `rat`, `fortunate` e.t.c

5.3 1910s

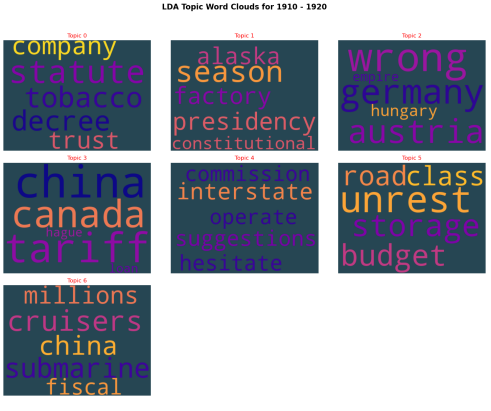

Figure 6: Topics extracted from speeches from 1911 to 1920.

The 1910's was a transformative decade marked by global conflicts such as world-war 1 (keywords `germany, hungary, Austria, wrong, empire`), rapid industrialisation fuelled by expanded transportation networks in US, anti-trust efforts targeting monopolistic companies like American Tobacco Company e.t.c Economic concerns, such as tariffs, loans, and budget issues, became prominent, particularly in relation to China, Canada, and

the broader geopolitical landscape. Technological advancements in naval warfare (keywords `submarine, cruisers`), reflected growing military focus, while internal unrest, industrial growth, and constitutional changes shaped the political discourse. The topics modelled are as shown in figure 6 well captures these characteristics of the decade.

5.4 1920s

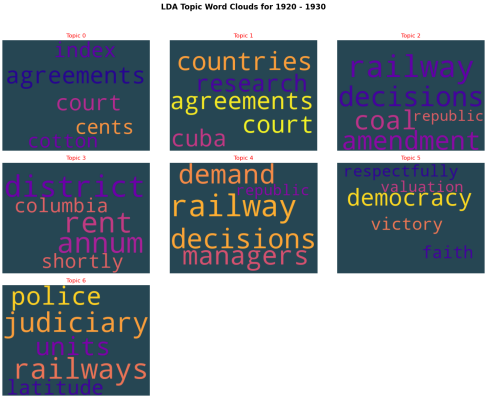

Figure 7: Topics extracted from speeches from 1921 to 1930.

The topics identified for this decade as shown in figure 7 are mostly regarding post-war diplomacy, economic growth, industrialization, and social reforms. If we look at actual historical events, we can see that one of the major change was the 19th Amendment granting voting rights to women. During the initial years of this decade there was significant economic growth and prosperity.

5.5 1930s

The 1930s were characterised by the aftermath of the Wall Street crash in 1929 and the great depression after that. This decade saw economic hardships, high unemployment and economic reforms by Franklin D Roosevelt. The major keywords that appear across the topics for this decade as shown in figure 8 are `loan, stability, emergency`, which captures the essence of this decade accurately.

5.6 1940s

The major world events during the 1940's were the World War II, in which USA enter's the war after the attack on Pearl Harbor in 1941, which ended with the use of atom bombs on Hiroshima and Nagasaki in 1945. From the word cloud shown in figure 9, the words such as `dictator, Japanese, guard, victory, enemy, ...` clearly captures the war efforts and subsequent global reorganization.

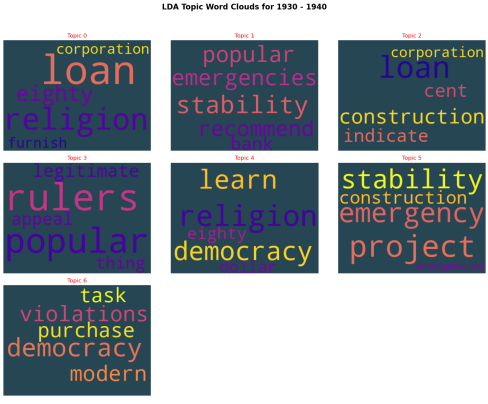

Figure 8: Topics extracted from speeches from 1931 to 1940.

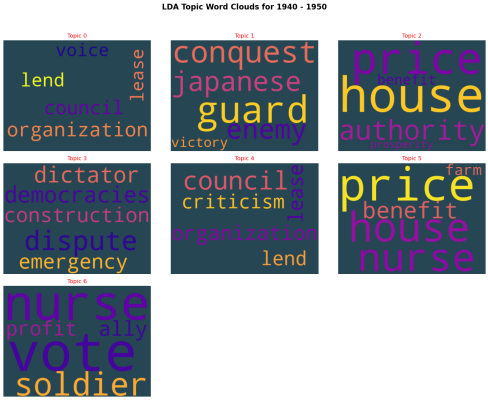

Figure 9: Topics extracted from speeches from 1941 to 1950.

5.7 1950s

1950s, the decade following the end of World War II marked the beginning of the long standing Cold-War era and the arms race. This historical event can be clearly identified from the pattern drawn by the modeled topics as shown in figure 10, through the words such as `russian, missile, science, soviets, ballistic`.

5.8 1960s

This era was heavily characterized by the Vietnam War and increased US involvement in it, leading to widespread protests and a growing anti-war movement. These characteristics of this decade were very well captured by the modeled topics, which is evident from figure 11. Furthermore this decade was characterized by a period of rapid growth and social programs, included tax cuts, increased government spending, and initiatives like the "Great Society" aimed at addressing poverty and inequality. The words `relief, bill, reduction, deb` could be indicate of these policies.

Figure 10: Topics extracted from speeches from 1951 to 1960.

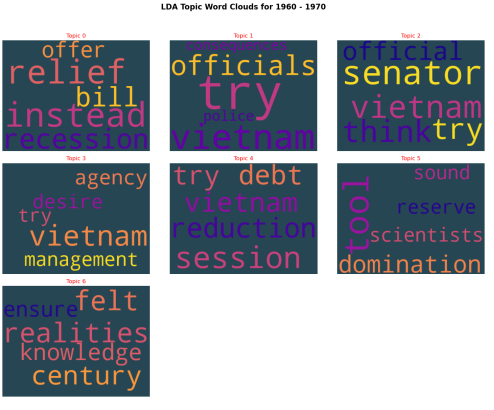

Figure 11: Topics extracted from speeches from 1961 to 1970.

5.9 1970s

The major markers of the 1970's in US were the Watergate Scandal, stagflation (economic stagnation and high inflation) and the energy crisis caused by disruptions in Middle Eastern oil exports due to the Yom Kippur War. Patterns relating to economic crisis, energy crisis and political scandals are evident in the topics modeled as shown in figure 12.

5.10 1980s

The 1980s was a transformative decade for US space exploration, dominated by the Space Shuttle program and marked by both remarkable achievements and profound challenges. Lebanon war was also a major global conflict that was featured in the modeled topics. The patterns relating to these events and changes are captured to a great extent by the model. The word cloud shown in figure 13 reflects this.

Figure 12: Topics extracted from speeches from 1971 to 1980.

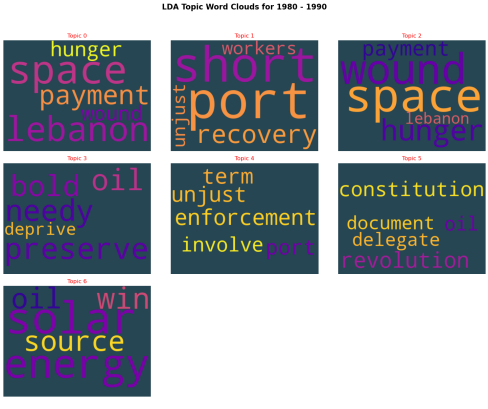

Figure 13: Topics extracted from speeches from 1981 to 1990.

5.11 1990s

The 1990s represented a pivotal moment of technological, economic, and geopolitical transformation. The United States emerged as the world's sole superpower, experienced unprecedented economic growth, and underwent a digital revolution that would reshape society in the coming decades[3, 5]. We can see the keywords such as `internet, standards, company` drawing these changes and patterns in figure 14. The attack on the twin towers and other terrorist attacks were a major concern during this period, however we could find these patterns in the modeled topics.

5.12 2000s

This decade was characterized by the war on terrorism by the US after the 9/11 attack on World Trade Center and the Pentagon. The Afghan war, Iraq war and the global financial crisis of 2008 were the major events of the early 21st century[4]. The armed conflicts and economic crisis is well captured by the modeled topics as shown in figure 15. Keywords from the topics include `saddam, qaeda, hussein, extremists, bank, lend`.

Figure 14: Topics extracted from speeches from 1991 to 2000.

Figure 15: Topics extracted from speeches from 2001 to 2010.

6 Conclusion

This detailed analysis shows how State of the Union topics are not just words, but a linguistic snapshot of national priorities, challenges, and transformations across decades. The LDA topics extracted from State-of-the-Union speeches from 1900 to 2010 reveal distinct patterns and connections to the major historical events of the corresponding decades.

References

[1] GeeksforGeeks. Understanding tf-idf (term frequency-inverse document frequency), 2025. URL https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency Accessed: 28 Mar. 2025.

[2] IBM. Stemming and lemmatization, 2025. URL https://www.ibm.com/think/topics/stemming-lemmatization. Accessed: 28 Mar. 2025.

[3] OpenAI. Chatgpt: Analysis of major economic events relating to the usa (1900s - 2010), 2025. URL https://chat.openai.com/. Accessed: 2025-03-28. Includes topics such as the Great Depression, post-WWII economic growth, the 1970s oil crisis, and the 2008 financial crisis.

[4] OpenAI. Chatgpt: Analysis of major military events relating to the usa (1900s - 2010), 2025. URL https://chat.openai.com/. Accessed: 2025-03-28. Includes topics such as World Wars I II, the Korean War, Vietnam War, Gulf Wars, and the War on Terror.

[5] OpenAI. Chatgpt: Analysis of major technological events relating to the usa (1900s - 2010), 2025. URL https://chat.openai.com/. Accessed: 2025-03-28. Includes topics such as the Space Race, computer revolution, internet development, and advancements in AI.