**Topic Modelling**

Sonam Thinley

**Abstract**

This report provides an in-depth analysis of the State of the Union speeches corpus using topic modelling techniques. The goal of this project is to extract meaningful insights from a dataset containing historical speeches delivered over the past century. Using Python programming, the data is pre-processed and TF-IDF document vectors are generated. Topic modelling is then performed using both Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) methods. Additionally, a method for summarizing changes in speech topics over decades is presented, allowing for a deeper understanding of the evolving themes and their connections to major historical events. The project's key components include data preprocessing, dictionary creation, bag-of-words corpus generation, LSI and LDA topic modelling, and annotation of resulting topics. The project aims to uncover hidden themes and provide valuable insights into the evolving priorities and concerns of the United States government as reflected in these annual presidential speeches.

**1 Pre-processing the data and Generating tf-idf weighted document vectors**

The initial step in the process involves data preprocessing, which encompasses two fundamental techniques: tokenization and lemmatization, with the aid of libraries such as Gensim and NLTK. The entire corpus is traversed twice: the first pass is dedicated to generating a dictionary consisting of unique tokens, and the second pass is focused on transforming each preprocessed document into a bag-of-words representation. Following this preprocessing phase, the entire corpus is converted into a document-term matrix. In this matrix, each column corresponds to a document, and each entry within a column signifies the unnormalized term frequency, essentially rendering each document in a mathematically convenient vector form. This representation is commonly referred to as a "bag-of-words" representation. Finally, the TF-IDF model is applied to transform each bag of words vector into a TF-IDF vector, where TF denotes "term frequency" and IDF stands for "inverse document frequency."

**2. Latent Semantic Indexing (LSI) to tf-idf vectors**

LSI, or Latent Semantic Indexing, is a technique used in natural language processing and information retrieval that works with TF-IDF vectors. By reducing the dimensionality of these vectors, LSI uncovers the semantic structure within a corpus of text. It does this by identifying hidden relationships between words and documents, which leads to a more meaningful representation of textual data.

After implementing the LSI model, topics generated will be annotated with descriptive names, and evaluate how accurately each topic represents a human concept.

**Optimal Number of Topics Selection**

In order to determine the optimal number of topics for LSI, a methodical approach that utilized coherence scores as a guiding metric was utilized. These scores help assess the understandability and excellence of the topics. While selecting fewer than 10 topics will encompass all of the general topics, we have adequate data to also uncover subtopics.

A graph with a line

Description automatically generated

*Fig-1:-Coherence scores vs Number of Topics*

It's best to avoid using more than 30 topics for analysis as it can complicate the process without providing any significant benefits. Based on the plot above, it appears that 25 may be the optimal number of topics as it has the highest coherence score within the range of 12-50.

**Generate Topics using LSI Model and sample 10**

With the optimal number of topics determined, we proceeded to create an LSI model using the TF-IDF. After implementing the LSI model topics, 25 topics are printed as lists of word coefficients, and we can identify the terms most strongly associated with each topic.

The real essence of LSI topic modelling comes to light when we sample ten topics for closer examination and provide descriptive names or phrases that encapsulate their meaning. This process requires referring back to the original documents that exhibit high proportions of a particular topic. In doing so, we gain a deeper insight into the content and nuances of these topics. Some of the topics lack a clear, easily definable concept, highlighting the complexities and nuances of natural language and the challenges of uncovering latent semantic structures.

**Annotation of Topics**

After obtaining the topics, we sampled ten of them for closer analysis. Annotating topics involves reviewing the terms with the highest coefficients in each topic and trying to provide a concise, descriptive name or phrase that captures the essence of that topic. This process requires referring to the original documents that exhibit high proportions of the respective topic.

**Topic 0 : "Economic Policy and Government Programs"**

Topic 0

Topic #0: -0.125\*"program" + -0.101\*"help" + -0.100\*"tonight" + -0.089\*"americans" + -0.086\*"economic" + -0.084\*"budget" + -0.072\*"job" + -0.069\*"today" + -0.064\*"billion" + -0.064\*"school"

Year with the highest proportion of Topic #0: 1911 (Proportion: -0.20052796257337172)

The conversation appears to center around economic policies, government programs, and budget considerations. Key phrases such as "program," "help," "economic," "budget," "job," and "billion" suggest a concentration on economic issues and government initiatives that may be aimed at tackling economic difficulties and introducing new programs to benefit Americans.

**Topic 2: Taxation and Economic Policy**

Topic #2: -0.219\*"tonight" + 0.133\*"economic" + -0.132\*"iraq" + 0.122\*"program" + 0.116\*"farm" + -0.113\*"job" + 0.112\*"interstate" + -0.111\*"americans" + -0.108\*"terrorists" + 0.105\*"industrial"

Year with the highest proportion of Topic #2: 1926 (Proportion: 0.3595610464392575)

This topic discusses taxation, economic policies, and their impact on the nation's prosperity potentially under Calvin Coolidge's presidency in the 1920s.

**Topic #3: Economic Policy and International Relations**

Topic #3: 0.171\*"silver" + -0.137\*"program" + -0.129\*"militia" + 0.118\*"gold" + 0.094\*"circulation" + 0.093\*"tonight" + 0.091\*"currency" + -0.088\*"economic" + 0.088\*"coinage" + 0.086\*"coin" + 0.086\*"arbitration" + 0.084\*"cent" + -0.083\*"tribes" + 0.082\*"conference" + 0.080\*"cuba" + -0.079\*"gentlemen" + 0.078\*"iraq" + 0.077\*"interstate" + 0.075\*"corporations" + 0.074\*"award"

Year with the highest proportion of Topic #3: 1889 (Proportion: 0.33051805663751366)

This topic primarily discusses economic matters related to currency, coinage, gold, and silver. It also touches upon international relations, diplomacy, and cooperation with other nations. The document from 1889 contains discussions about international conferences, trade, and relations with various countries.

Topic #4: "Regulation of Interstate Corporations and Railroads in 1907"

Topic #4: 0.263\*"interstate" + 0.204\*"corporations" + -0.156\*"program" + -0.151\*"mexico" + 0.140\*"railroad" + -0.123\*"soviet" + 0.106\*"iraq" + -0.102\*"texas" + 0.100\*"combinations" + -0.093\*"communist"

Year with the highest proportion of Topic #4: 1907 (Proportion: 0.46959267783882674)

This text discusses President Theodore Roosevelt's 1907 State of the Union Address, which emphasized the importance of regulating interstate corporations and railroads during a time of significant industrial expansion.

**Topic #5: "Unrelated References to Iraq and Terrorism"**

Topic #5: -0.253\*"iraq" + -0.206\*"terrorists" + -0.197\*"fight" + -0.190\*"enemy" + -0.156\*"iraqi" + -0.149\*"japanese" + -0.144\*"enemies" + -0.133\*"terror" + -0.118\*"victory" + 0.114\*"job

Year with the highest proportion of Topic #5: 1993 (Proportion: 0.25227564527322627)

Although this topic contains keywords related to Iraq, terrorism, fighting, and enemies, it does not seem to be in line with the main themes of the speech given by William J. Clinton in 1993. The speech primarily focuses on economic policy and revitalization efforts, and therefore, this topic appears to include irrelevant references to Iraq and terrorism. As a result, it does not play a central role in the addressed subject matter of the speech.

**Topic #6: "Diplomacy and International Relations in the Americas"**

Topic #6: -0.411\*"mexico" + -0.349\*"texas" + -0.193\*"mexican" + 0.191\*"silver" + -0.138\*"interstate" + -0.131\*"corporations" + -0.131\*"annexation" + 0.121\*"coinage" + 0.099\*"coin" + -0.097\*"california"

Year with the highest proportion of Topic #6: 1889 (Proportion: 0.2657535047380796)

The main theme of this discussion centers around the relationships between countries in North and South America, specifically with a focus on Mexico, Texas, and nearby states. The topic touches on diplomatic affairs, international conferences, and the necessity for improvements in how diplomatic representation is carried out. Additionally, the significance of preserving peaceful relations within the Americas is emphasized.

**Topic #7: "Taxation, Economy, and Fiscal Policy in 1926"**

Topic #7: -0.160\*"gold" + -0.158\*"silver" + 0.145\*"iraq" + -0.138\*"currency" + -0.135\*"constitution" + -0.129\*"coin" + -0.114\*"kansas" + 0.112\*"terrorists" + -0.110\*"circulation" + -0.108\*"vietnam"

Year with the highest proportion of Topic #7: 1926 (Proportion: 0.24314210923080554)

The subject matter primarily centers on terms connected to gold, silver, currency, constitution, coins, and circulation. It delves into fiscal policy, taxation, and economic circumstances in 1926, emphasizing the significance of the economy, decreasing national taxes, and the excess funds in the Treasury.

**Topic #8: "Unrelated Topics"**

Topic #8: 0.219\*"mexico" + 0.212\*"texas" + 0.200\*"iraq" + 0.152\*"terrorists" + -0.151\*"corporations" + -0.146\*"interstate" + 0.123\*"currency" + 0.115\*"iraqi" + 0.115\*"bank" + 0.113\*"cent"

Year with the highest proportion of Topic #8: 1868 (Proportion: 0.33904838433570894)

This topic revolves around keywords such as "mexico," "texas," "iraq," "terrorists," "currency," "iraqi," "bank," and "cent." However the speech in 1868 primarily discusses the financial challenges faced by the United States in the aftermath of the Civil War, focusing on issues related to public expenditures, national debt, and fiscal policies. There is a lack of correlation between the topics and the speech.

**Topic #9: "Unrelated Topics"**

Topic #9: 0.193\*"iraq" + 0.179\*"gentlemen" + -0.176\*"enemy" + 0.149\*"terrorists" + -0.140\*"cuba" + -0.122\*"spain" + 0.113\*"iraqi" + -0.102\*"savage" + -0.098\*"japanese" + 0.094\*"terror"

Year with the highest proportion of Topic #9: 1799 (Proportion: 0.23879964179410215)

This topic is centered around keywords like "iraq," ,"enemy," "terrorists," "cuba," "spain," "iraqi," "savage," "japanese," and "terror." It appears to discuss international relations, conflicts, and diplomacy. However the speech refers to an earlier period which does not relate to the topics generated.

**Topic #10: "Spanish-American Relations and Economic Matters"**

Topic #10: 0.236\*"spain" + -0.184\*"kansas" + 0.173\*"gold" + 0.153\*"interstate" + 0.150\*"corporations" + 0.135\*"cuba" + -0.128\*"constitution" + 0.125\*"silver" + 0.117\*"currency" + -0.116\*"slavery"

Year with the highest proportion of Topic #10: 1897 (Proportion: 0.3224036195756279)

Topic being discussed involves Spain, Cuba, gold, silver, interstate commerce, corporations, currency, and economic matters. Most likely, it revolves around the economic relationship between Spain and the United States during the late 19th century. The mention of Cuba indicates that the issue of Cuba is also being focused on, which eventually led to the Spanish-American War.

**Word Cloud of the topics**

A collage of different colored text

Description automatically generated

Fig-2:-WordCloud for LSI Topics

**Comment**

The effectiveness of each "topic" in capturing a real human concept is varying in the topics. Some topics are quite effective at capturing coherent and meaningful concepts, while others appear to be a mix of unrelated terms or lack a clear, distinct concept.

**3. LDA Topic Modeling**

Latent Dirichlet Allocation (LDA) is a powerful natural language processing and machine learning technique used for topic modeling and document clustering. It helps uncover the underlying themes or topics within a collection of documents by probabilistically assigning words to topics and topics to documents. LDA assumes that each document is a mixture of various topics, and each topic is characterized by a distribution of words. Through iterations and statistical inference, LDA identifies these topics, their word distributions, and how they are distributed across documents.

**Annotating Topics:**

The first ten topics below will be selected for analysis and annotation.

[(0,

'0.001\*"silver" + 0.001\*"conference" + 0.000\*"chinese" + 0.000\*"japan" + 0.000\*"hayti" + 0.000\*"international" + 0.000\*"industries" + 0.000\*"fee" + 0.000\*"extradition" + 0.000\*"methods" + 0.000\*"switzerland" + 0.000\*"competitive" + 0.000\*"utah" + 0.000\*"hawaii" + 0.000\*"fishermen" + 0.000\*"korea" + 0.000\*"consular" + 0.000\*"cent" + 0.000\*"examinations" + 0.000\*"total" + 0.000\*"interoceanic" + 0.000\*"cents" + 0.000\*"railway" + 0.000\*"china" + 0.000\*"copyright"'),

(1,

'0.000\*"cent" + 0.000\*"coin" + 0.000\*"economic" + 0.000\*"industries" + 0.000\*"conference" + 0.000\*"militia" + 0.000\*"mexico" + 0.000\*"mexican" + 0.000\*"california" + 0.000\*"gentlemen" + 0.000\*"kansas" + 0.000\*"interstate" + 0.000\*"silver" + 0.000\*"mercy" + 0.000\*"louis" + 0.000\*"memory" + 0.000\*"merge" + 0.000\*"mechanics" + 0.000\*"master" + 0.000\*"massachusetts" + 0.000\*"mainly" + 0.000\*"mail" + 0.000\*"machine" + 0.000\*"municipal" + 0.000\*"move"'),

(2,

'0.001\*"savage" + 0.001\*"colonies" + 0.001\*"pensacola" + 0.001\*"adventurers" + 0.001\*"catholic" + 0.001\*"gallantry" + 0.000\*"squadron" + 0.000\*"provinces" + 0.000\*"tribes" + 0.000\*"enemy" + 0.000\*"likewise" + 0.000\*"erie" + 0.000\*"militia" + 0.000\*"marys" + 0.000\*"whilst" + 0.000\*"augmentation" + 0.000\*"presume" + 0.000\*"felicity" + 0.000\*"exclusively" + 0.000\*"indies" + 0.000\*"ghent" + 0.000\*"martial" + 0.000\*"jam" + 0.000\*"float" + 0.000\*"madison"'),

(3,

'0.001\*"texas" + 0.001\*"mexico" + 0.001\*"mexican" + 0.001\*"annexation" + 0.001\*"ministry" + 0.000\*"chamber" + 0.000\*"oregon" + 0.000\*"holland" + 0.000\*"ration" + 0.000\*"corporate" + 0.000\*"compromise" + 0.000\*"note" + 0.000\*"wilderness" + 0.000\*"deposit" + 0.000\*"fluctuations" + 0.000\*"deference" + 0.000\*"dissatisfaction" + 0.000\*"specie" + 0.000\*"chili" + 0.000\*"granada" + 0.000\*"rice" + 0.000\*"invade" + 0.000\*"unexampled" + 0.000\*"obviously" + 0.000\*"naples"'),

(4,

'0.001\*"gentlemen" + 0.001\*"seal" + 0.001\*"insurgents" + 0.001\*"whatsoever" + 0.001\*"croix" + 0.001\*"chinese" + 0.001\*"militia" + 0.001\*"silver" + 0.001\*"counties" + 0.001\*"philadelphia" + 0.001\*"burthen" + 0.001\*"nicaragua" + 0.001\*"gold" + 0.001\*"recess" + 0.001\*"coinage" + 0.001\*"tranquillity" + 0.001\*"holland" + 0.001\*"kentucky" + 0.001\*"catholic" + 0.000\*"commissioners" + 0.000\*"george" + 0.000\*"coin" + 0.000\*"cuba" + 0.000\*"passamaquoddy" + 0.000\*"cent"'),

(5,

'0.001\*"seal" + 0.001\*"gold" + 0.001\*"chinese" + 0.000\*"nicaragua" + 0.000\*"insurgents" + 0.000\*"silver" + 0.000\*"gentlemen" + 0.000\*"legations" + 0.000\*"cent" + 0.000\*"coin" + 0.000\*"hawaii" + 0.000\*"cuba" + 0.000\*"peking" + 0.000\*"conference" + 0.000\*"santiago" + 0.000\*"salvador" + 0.000\*"missionaries" + 0.000\*"china" + 0.000\*"arbitration" + 0.000\*"international" + 0.000\*"manila" + 0.000\*"bullion" + 0.000\*"tons" + 0.000\*"coinage" + 0.000\*"chilean"'),

(6,

'0.001\*"chamber" + 0.001\*"colonies" + 0.001\*"ghent" + 0.001\*"redeemable" + 0.001\*"upward" + 0.001\*"installment" + 0.001\*"interdict" + 0.000\*"discriminate" + 0.000\*"chesapeake" + 0.000\*"colonial" + 0.000\*"porte" + 0.000\*"naples" + 0.000\*"flourish" + 0.000\*"cumberland" + 0.000\*"delaware" + 0.000\*"maxims" + 0.000\*"rice" + 0.000\*"ayres" + 0.000\*"monarch" + 0.000\*"relaxation" + 0.000\*"throne" + 0.000\*"affaires" + 0.000\*"recess" + 0.000\*"squadron" + 0.000\*"tonnage"'),

(7,

'0.001\*"kansas" + 0.001\*"california" + 0.000\*"mexico" + 0.000\*"rebellion" + 0.000\*"mexican" + 0.000\*"honduras" + 0.000\*"slavery" + 0.000\*"acres" + 0.000\*"utah" + 0.000\*"texas" + 0.000\*"kentucky" + 0.000\*"loyal" + 0.000\*"slave" + 0.000\*"electors" + 0.000\*"juan" + 0.000\*"oregon" + 0.000\*"jefferson" + 0.000\*"nicaragua" + 0.000\*"election" + 0.000\*"postal" + 0.000\*"mineral" + 0.000\*"nebraska" + 0.000\*"circuit" + 0.000\*"coin" + 0.000\*"seacoast"'),

(8,

'0.001\*"jefferson" + 0.001\*"household" + 0.001\*"upward" + 0.001\*"edicts" + 0.001\*"thomas" + 0.001\*"tripoli" + 0.001\*"chesapeake" + 0.001\*"mediterranean" + 0.001\*"militia" + 0.001\*"barbary" + 0.001\*"boat" + 0.000\*"orleans" + 0.000\*"burthen" + 0.000\*"dispositions" + 0.000\*"shew" + 0.000\*"embargo" + 0.000\*"identify" + 0.000\*"louisiana" + 0.000\*"posture" + 0.000\*"impost" + 0.000\*"cannon" + 0.000\*"council" + 0.000\*"tunis" + 0.000\*"messrs" + 0.000\*"revoke"'),

(9,

'0.000\*"silver" + 0.000\*"cent" + 0.000\*"coinage" + 0.000\*"rebellion" + 0.000\*"expatriation" + 0.000\*"gold" + 0.000\*"coin" + 0.000\*"cuba" + 0.000\*"postal" + 0.000\*"naturalization" + 0.000\*"herewith" + 0.000\*"award" + 0.000\*"japan" + 0.000\*"specie" + 0.000\*"reconstruction" + 0.000\*"vienna" + 0.000\*"woods" + 0.000\*"depreciate" + 0.000\*"manufactories" + 0.000\*"medium" + 0.000\*"holders" + 0.000\*"capita" + 0.000\*"industries" + 0.000\*"strife" + 0.000\*"dollar"'),

(10,

'0.000\*"chamber" + 0.000\*"colonies" + 0.000\*"redeemable" + 0.000\*"ghent" + 0.000\*"installment" + 0.000\*"porte" + 0.000\*"chesapeake" + 0.000\*"upward" + 0.000\*"cumberland" + 0.000\*"superintendence" + 0.000\*"colonial" + 0.000\*"interdict" + 0.000\*"commissioners" + 0.000\*"delaware" + 0.000\*"indies" + 0.000\*"naples" + 0.000\*"cents" + 0.000\*"withstand" + 0.000\*"flourish" + 0.000\*"maxims" + 0.000\*"affaires" + 0.000\*"francs" + 0.000\*"throne" + 0.000\*"monarch" + 0.000\*"revolutions"'),

**Annotations**

Based on the topics generated and the associated speech that has the highest proportion of the topics, the following topics have been annotated:

Topic 0: "Education and Economic Support"

Topic 0: tonight, school, child, drug, children, job, help, college, parent, reclamation, americans, constitution, budget, article, valuation, students, program, affordable, agriculture, invest, duties, economic, tariff, dispositions, blockade"

Year with the highest proportion of Topic #0: 1804 (Proportion: 0.6690909266471863)

Topic 1: "International Relations and Diplomacy"

Topic 1: soviet, examinations, interstate, partisan, bargain, administrative, veterans, program, duties, management, official, copy, herewith, commence, appointment, energy, dispute, afghanistan, collective, jurisdictional, island, legal, mexico, atomic, persons"

Year with the highest proportion of Topic #1: 1916 (Proportion: 0.3983463943004608)

Topic 2: "International Diplomacy and Trade"

Topic 2: slave, nicaragua, arbitration, international, program, japan, consular, soviet, china, african, boundary, chinese, convention, british, diplomatic, seal, economic, mexico, afghanistan, republics, lease, minister, budget, britain, lend"

Year with the highest proportion of Topic #2: 1881 (Proportion: 0.17923541367053986)

Topic 3: "Social Challenges and Economic Issues"

Topic 3: program, vietnam, percent, americans, help, crime, billion, challenge, tonight, children, corporations, start, wage, soviet, deficits, partnership, spend, social, heroism, anarchy, million, job, needy, workers, medicare"

Year with the highest proportion of Topic #3: 1901 (Proportion: 0.15112580358982086)

Topic 4: "Agriculture and Economic Concerns"

Topic 4: savage, farm, california, group, spain, program, price, mexico, democratic, income, miles, soldier, production, coast, economic, help, sailors, combination, selfish, plenipotentiary, units, farmers, tribes, territory, enemies"

Year with the highest proportion of Topic #4: 1811 (Proportion: 0.3727635443210602)

Topic 5: "International Trade and Fiscal Policy"

Topic 5: spain, kansas, program, tonight, communist, budget, silver, economic, award, soviet, help, minister, majesty, claim, currency, fiscal, article, tribunal, democracy, billion, britain, naturalization, treasury, school, territory"

Year with the highest proportion of Topic #5: 1894 (Proportion: 0.9507661461830139)

Topic 6: "Economic Recovery and Education"

Topic 6: silver, help, democracy, job, recovery, circulation, farm, americans, school, program, company, tonight, modern, college, today, economic, problems, minister, task, teachers, majesty, britain, bank, education, million"

Year with the highest proportion of Topic #6: 1937 (Proportion: 0.5269790887832642)

Topic 7: "Militia and Foreign Affairs"

Topic 7: militia, gentlemen, counties, challenge, spain, belligerent, suspension, cuba, pennsylvania, indian, children, commissioners, tribes, decree, kentucky, achieve, economic, requisite, indians, program, france, embargo, object, democratic, telegraph"

Year with the highest proportion of Topic #7: 1790 (Proportion: 0.7890498638153076)

Topic 8: "Economic Initiatives and Space Exploration"

Topic 8: tonight, program, budget, help, today, americans, spend, economic, chamber, energy, dream, school, billion, percent, job, initiative, space, kid, children, california, strategic, soviet, communist, families, drug"

Year with the highest proportion of Topic #8: 1985 (Proportion: 0.5335350036621094)

Topic 9: "Economic Challenges and Agricultural Focus"

Topic 9: inflation, farmer, texas, economic, annexation, job, program, agriculture, flood, unemployment, group, mexico, ahead, today, surplus, bank, propaganda, emergency, problem, farm, reduction, tonight, minister, major, canal"

Year with the highest proportion of Topic #9: 1940 (Proportion: 0.27662068605422974)

**How LDA differed from LSI**

To summarize, the topics generated by both LSI and LDA are in line with the prevalent themes present in the documents, including education, international relations, diplomacy, social issues, and agriculture. However, LDA topics are easier to interpret due to their probabilistic nature and word distribution, whereas understanding the meaning behind LSI topics requires a closer examination of the associated keywords.

There were more cases using the LSI models where the topics did not match to the speech with the high proportion of the topics. For example a topic with words “terrorism”, “Iraq”, “military” did not match with the speech. However, the same topics in LDA matched with the speech. For example the, above topics matched to a 2003 speech given by President Bush.

When it comes to identifying topics, LSI and LDA use different methods. LSI uses singular value decomposition, which can be less clear and easy to understand, while LDA uses a probabilistic generative model, resulting in more coherent and interpretable topics. LDA provides clear and theme-based topics that reflect the structure of the dataset, while LSI requires a more in-depth analysis of keyword associations. Overall, LDA provides a more user-friendly and insightful way to extract topics, making it easier to understand the content and themes found in State of the Union speeches.

**Word Cloud for LDA model topics**

A collage of different colored text

Description automatically generated

**4. Decade Summarisation Algorithm**

The decade summarization algorithm organizes textual data into 12 decades, based on the year they were published. It then processes each decade's text data by making everything lowercase, getting rid of numbers, and removing common words that don't add much value. Once the text corpus is compiled and filtered for each decade, it creates a dictionary and changes the data into a bag-of-words format.

Next, the algorithm applies TF-IDF transformation technique to weigh the importance of words. The Latent Dirichlet Allocation (LDA) topic modeling is used, which detects the major topics within each decade. The algorithm prints out these topics, their top words, and their probabilities, providing valuable insights into the most important themes during different time periods. This approach allows us to analyze history and discover how topics have evolved or become more significant over the years.

**Decade Topic Annotations:**

**Topic 0: "International Relations”**

This topics contains words such as international affairs, trade, or conferences,

**Topic 1: "Economic Matters"**

This topics contains words such as cent, industry, economic, etc.

**Topic 2: "Military Engagements"**

Some of the words in this topic that stand out include "savage," "colonies," "pensacola," "adventurers," "catholic," and "gallantry." These words collectively suggest a topic related to historical exploration, colonization efforts, and potentially conflicts or interactions with indigenous populations.

**Topic 3: "Territorial Disputes"**

The words in this topic and their associated probabilities suggest a theme related to historical and geopolitical aspects, particularly focusing on the United States and its interactions with Mexico and other nations.The prominent words in this topic include "texas," "mexico," "mexican," "annexation," "ministry," and "oregon." These words collectively point towards a topic that might involve historical events, such as the annexation of Texas, diplomatic relations between the United States and Mexico, and possibly territorial disputes.

**Topic 4: "Government and Coinage".** This topic could be related to US-Chinese policy, regional discussions, trade and economy.

**Topic 5: "International Diplomacy".** This topic includes a mix of international and economic terms. It might relate to discussions or policies involving trade, diplomacy, or international relations with countries like China, Nicaragua, Cuba, and others.

**Topic 6: "Colonial Affairs”.** This topic includes words that refer to cononialism, monarch, and some historical events.

**Topic 7: "Slavery and Territories".** This topic seems to revolve around various geographical and political aspects, including U.S. states like "kansas," "california," and "texas,". It also talks about terms related to politics such as "rebellion," "electors," and "election." The inclusion of "slavery" and "slave" suggests discussions related to the contentious issue of slavery.

**Topic 8: "Barbary Wars".** The topic may be associated with the early history of the United States, possibly during the Jeffersonian era, and may touch upon matters related to foreign policy, military affairs, and international relations, particularly with reference to the Barbary Wars and naval operations in the Mediterranean.

**Topic 9: "Currency and Economic Issues".** Mainly talks about currency, finance, economy.

**Topic 10: "Colonial Policies".** Talks about the monarch, colonization

**Topic 11: "Industrial and Economic Development"** Mainly talks about US interstate discussions and currency.

**Topic 12: "Economic Challenges".** Includes some topics involving Mexico and interstate.

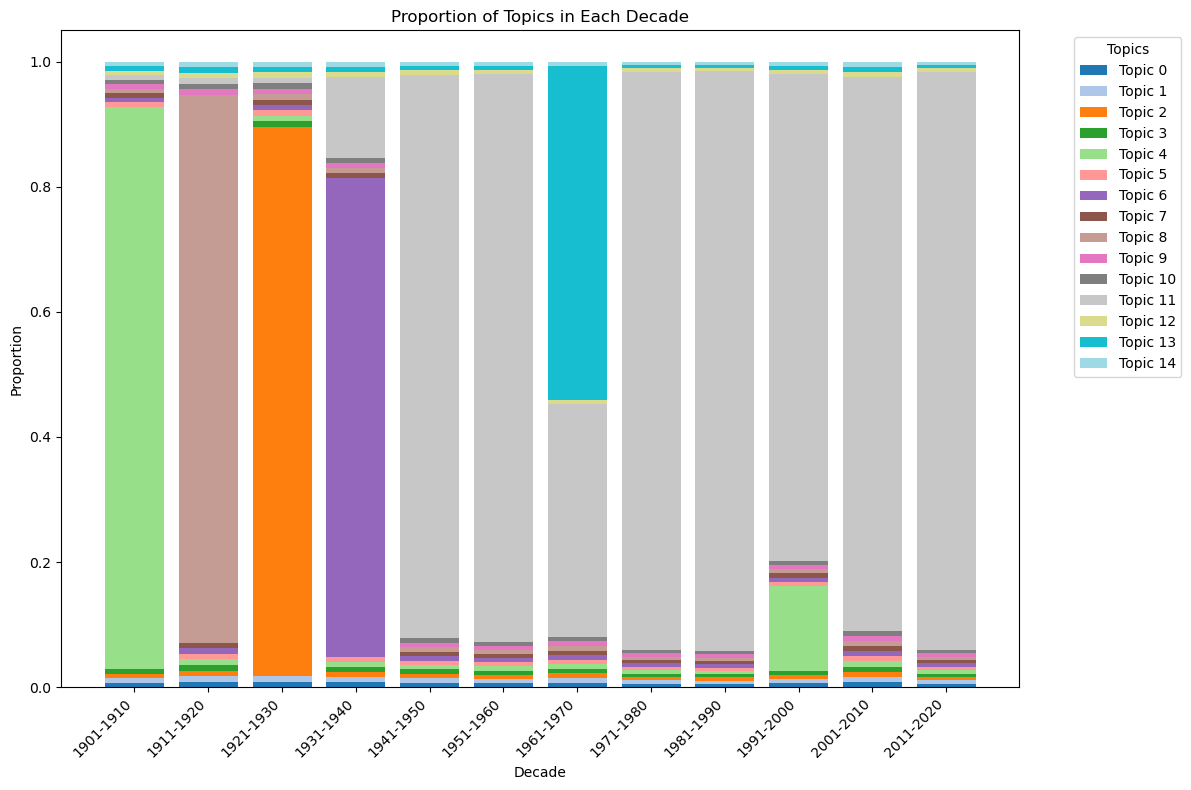
**Topic 13: "Corporations and Labor"**Topics include slavery, interstate transportation, monarch and foreign policy.

**Topic 14: "Labor and Corporate Policies".** Talks about interstate matter, corporations, wage and relations with Mexico.

**Proportion of Topic in each decade.**

Below is the proportion of each topic in each decade.

These proportions provide insights into the dominant themes within each decade, with Topic 11 ("Industrial and Economic Development") being prevalent in most decades, especially in the later years. Other topics have a relatively low proportion in each decade, indicating that they are less prominent in the corpus during those time periods.



|  |  |
| --- | --- |
| Decade | Topic Proportion |
| 1901-1910: | Topic 4: 89.88% (Government and Coinage)  Other Topics: Approximately 0.72% each |
| 1911-1920: | Topic 8: 87.57% (Barbary Wars)  Other Topics: Approximately 0.89% each |
| 1921-1930: | Topic 2: 87.85% (Military Engagements)  Other Topics: Approximately 0.87% each |
| 1931-1940: | Topic 6: 76.51% (Colonial Affairs)  Topic 11: 12.98% (Industrial and Economic Development)  Other Topics: Approximately 0.81% each |
| 1941-1950: | Topic 11: 90.08% (Industrial and Economic Development)  Other Topics: Approximately 0.71% each |
| 1951-1960: | Topic 11: 90.74% (Industrial and Economic Development)  Other Topics: Approximately 0.66% each |
| 1961-1970: | Topic 13: 53.28% (Corporations and Labor)  Topic 11: 37.19% (Industrial and Economic Development)  Other Topics: Approximately 0.73% each |
| 1971-1980: | Topic 11: 92.38% (Industrial and Economic Development)  Other Topics: Approximately 0.54% each |
| 1981-1990: | Topic 11: 92.67% (Industrial and Economic Development)  Other Topics: Approximately 0.52% each |
| 1991-2000: | Topic 11: 77.89% (Industrial and Economic Development)  Topic 4: 13.59% (Government and Coinage)  Other Topics: Approximately 0.66% each |
| 2001-2010: | Topic 11: 88.50% (Industrial and Economic Development)  Other Topics: Approximately 0.82% each |
| 2011-2020: | Topic 11: 92.41% (Industrial and Economic Development)  Other Topics: Approximately 0.54% each |
|  |  |
|  |  |