# Assignment 2 - Topic Modeling

Nayan Arora University of Canberra
ID-u3249907 ACT, AUSTRALIA

Abstract

The objective of assignment is to implement the Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) topic modeling algorithms to model topics using the State of the Union dataset provided. The dataset consists of two columns – year of the Speech and text of the speech. The genism library is used for preprocessing, and modeling. The matplotlib library has been used to produce graph plots and the wordcloud library has been used to produce wordclouds. Also note that the natural language toolkit (nltk) was used to produce a list of common stop words. Resources provided in the assignment instructions including this external article were used to complete this assignment. This report covers the steps used to produce the topics from the State of the Union dataset along with visualizations and further analyzations.

## 1. Pre-Processing and TF-IDF Document Vectors

The first step used in this implementation is to pre-process the state of the union dataset which helps in reducing future complications and improve results. We do so by removing redundant and meaningless data points which would not help with the objective of topic modeling. To achieve this the genism and nltk library were used. The techniques applied are categorized as lemmatization and stemming which helped achieve processed documents. Then the entire speech corpus is traversed to generate a dictionary with speech and year of speech (tokenization). Next, we transform each processed document to create a bag of words corpus which indexes the term frequency for each processed document (unnormalized term frequency). Last, each vectorized document element in the corpus is transformed using the TF-IDF model into tf-idf (term frequency-inverse document frequency) vectors.

## 2. Latent Semantic Indexing (LSI) Topic Modeling

For LSI Topic Modeling is used for analyzing the relationships between words in a large corpus of text. It is helpful in finding hidden semantic patterns to improve document retrieval and relevance ranking. LSI uses singular value decomposition to transform the term document matrix which helps in capturing the underlying meaning of the words. This process helps LSI to identify patterns and relations in the document which is the underlying idea in Topic Modeling.

The first step required for topic modeling is to have an intelligent guess for the number of topics that we would feed the algorithm. To infer the number of topics, we calculate and plot coherence scores over number of topics. Topic Coherence score for a topic is basically the degree of sematic similarity between high scoring words in the topic. Topic coherence helps us distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. Below is a plot of topic coherence scores over 50 topics.

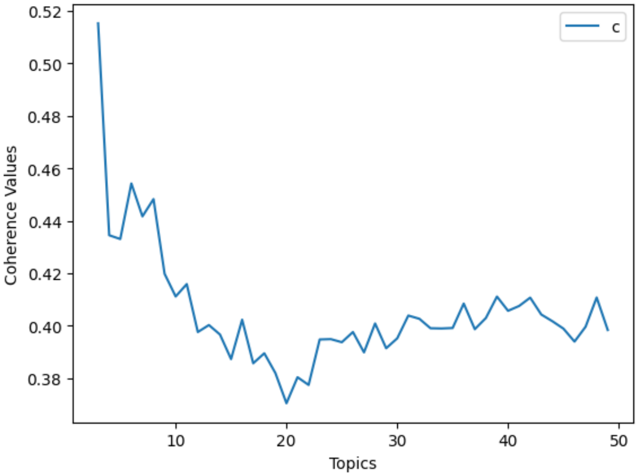

*Figure 1: Coherence Scores vs Number of Topics (LSI)*

So, we see that we have very high topic coherence values for total topics less than 10. But since we have a very large data at our disposal, we can use a higher number of topics to generate subtopics and thus gather better inference. For the LSI model 35 topics were generated.

Now, after generating the topic word cloud library was used to plot word clouds with 50 words for each topic that helped annotate all the topics. Thus, helping infer 'what it is about'. Attached below are all the generated topics:

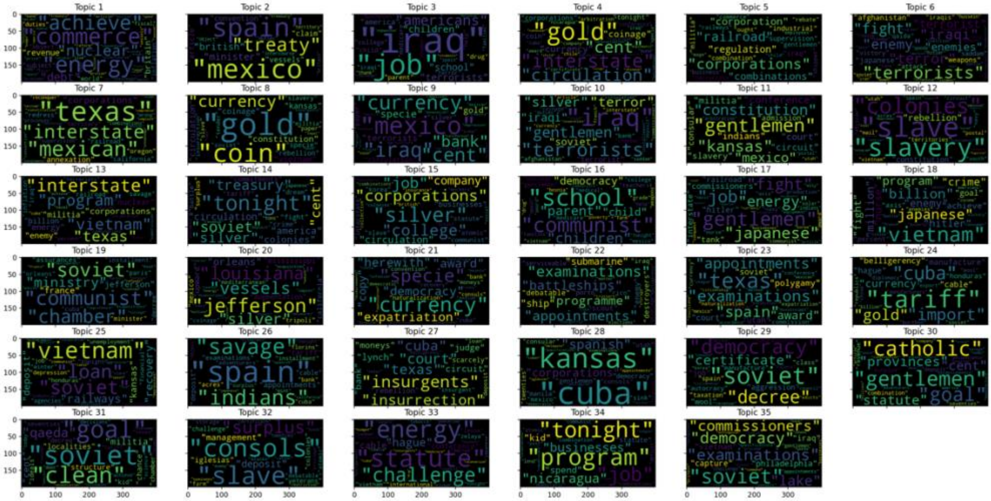

*Figure 2: LSI Topics*

Now, we sample 10 random topics to try and analyze what they are about:

1) Topic 1:-0.115*"program"+-0.087*"help" + -0.086*"tonight" + -0.079*"economic" + -0.077*"americans" + -0.075*"budget" + -0.068*"america" + -0.064*"job" + -0.061*"treaty" + -0.059*"today"

   This topic has a theme of moving forward. The words program, tonight, help, economic, americans, budget, America, job treaty helped me infer that the speech is about a program that is being launched tonight which could help America's economy. This economic boom in the market could increase jobs. This theme is also highlighted by the word cloud as achieve commerce, energy, revenue, nuclear. I was unable to find these proportion in the original corpus. So, I am unsure which year's speech it exactly belongs to. But we follow a similar structure for the rest of the topics.

2) Topic 3: -0.207*"tonight" + 0.121*"economic" + -0.119*"iraq" + 0.114*"program" + -0.109*"job" + -0.107*"americans" + 0.105*"farm" + 0.102*"interstate" + -0.099*"terrorists" + -0.097*"school"

   This topic talks about the terror in Iraq and how it might affect the American economy, the jobs as well as schools.

3) Topic 7: -0.379*"mexico" + -0.336*"texas" + -0.182*"mexican" + 0.149*"silver" + -0.137*"interstate" + -0.125*"corporations" + -0.123*"annexation" + 0.093*"coinage" + -0.085*"california" + -0.084*"railroad"

This topic had no clear theme except for a focus on something to do with Texas, Mexico and their connectivity via railroad and interstate highways. Thus, cannot be clearly annotated to any theme.

4) Topic 10: -0.168*"enemy" + 0.131*"iraq" + 0.119*"terrorists" + 0.105*"gentlemen" + -0.100*"mexico" + -0.096*"army" + -0.094*"japanese" + -0.094*"naval" + 0.084*"silver" + 0.083*"terror"

This topic can be annotated to the war theme. Keywords terror, Iraq, Iraqi, Japanese, enemy all relate to war. This could relate to the time of World War 2.

5) Topic 14: 0.281*"vietnam" + 0.199*"tonight" + -0.142*"spain" + -0.125*"cuba" + -0.117*"gentlemen" + -0.105*"job" + -0.086*"energy" + -0.086*"kansas" + -0.083*"spanish" + -0.082*"businesses"

No clear inference from this topic. Just different names of places which would require a lot of knowledge about the past to infer why all these places could be mentioned in a single line.

6) Topic 16: 0.183*"energy" + 0.177*"job" + -0.146*"school" + -0.120*"communist" + -0.116*"children" + 0.114*"spend" + 0.112*"inflation" + 0.110*"solar" + -0.101*"parent" + -0.089*"child"

This topic could be annotated to the time of the economic crisis. Because the keyword solar is present it cannot be the first one as technology was not that advanced then. Thus, this topic relates to the time period 2008.

7) Topic 18: 0.287*"soviet" + -0.284*"vietnam" + -0.105*"billion" + -0.105*"japanese" + -0.100*"program" + 0.079*"nicaragua" + -0.078*"enemy" + -0.077*"million" + -0.077*"hitler" + -0.072*"crime"

This topic belongs to the time period of cold war with the soviet. Around the same time period US had their army in Vietnam and billions were being spent to tackle war and crime.

8) Topic 22: -0.077*"cable" + -0.072*"examinations" + -0.071*"appointments" + 0.069*"slave" + 0.068*"hitler" + 0.068*"combinations" + 0.067*"japanese" + -0.064*"serviceable" + 0.064*"corporations" + -0.063*"programme"

No clear inference other than corporations and technological advances.

9) Topic 23: 0.137*"acres" + -0.135*"vietnam" + 0.111*"california" + -0.110*"texas" + -0.093*"soviet" + 0.086*"silver" + 0.082*"minister" + -0.078*"currency" + 0.071*"respectfully" + -0.069*"hague"

No clear inference. May belong to the early 1900’s because of the key word hague.

10) Topic 30: 0.092*"examinations" + -0.090*"deposit" + 0.083*"appointments" + -0.075*"bank" + 0.074*"billion" + -0.072*"constitution" + -0.069*"consols" + -0.067*"enemy" + -0.066*"challenge" + -0.066*"america"

No clear inference from this topic either. Although it does have a general them of maintaining memorandum. Consitution, banks, billions are all financial related. So it must be from a time when the economy was well.

## 3. Latent Dirichlet Allocation (LDA) Topic Modeling

LDA is a probabilistic model. It assumed that a collection of documents is generated from a mixture of topics, where each topic is a probability distribution over words. The aim of LDA algorithm is to uncover the hidden topics by analyzing and recognizing the pattern in which the words occur in the documents. Using the defined statistical inference of LDA topic modeling the words are assigned to topics and then the documents are assigned to topic mixtures. We use the LDA algorithm for topic modeling using similar steps used for LSI modeling. First step is to infer the number of topics by plotting the coherence scores and number of topics. Below is the plot:

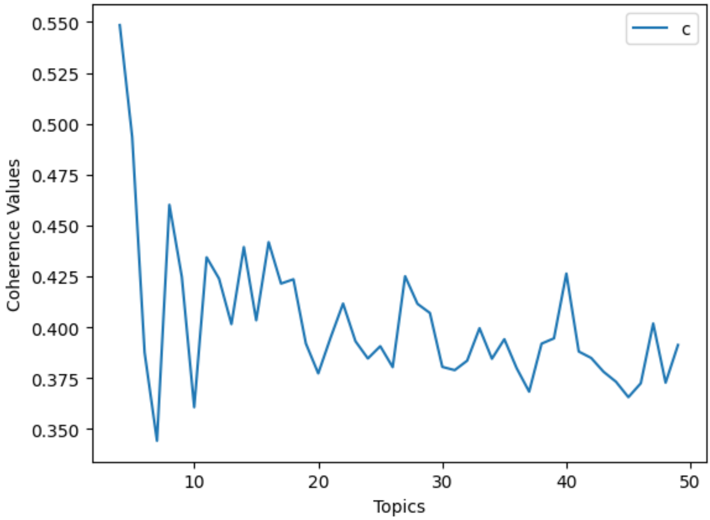

*Figure 3: Coherence Scores vs Number of Topics (LDA)*

On evaluating the coherence scores, we have similar inference as earlier found in LSI. We could go with topics less than 10 as it has a relatively higher coherence value but for a dataset as large as we have, we should opt for a higher number of topics for better inference and subtopic predictions. In this case, we go with 25 number of topics as that is when it peaks and then drops again before it goes to over 40 topics. 40 is a bit high and might add unwanted complexities, so we choose 25 topics.
Attached below are the topics generated by LDA:

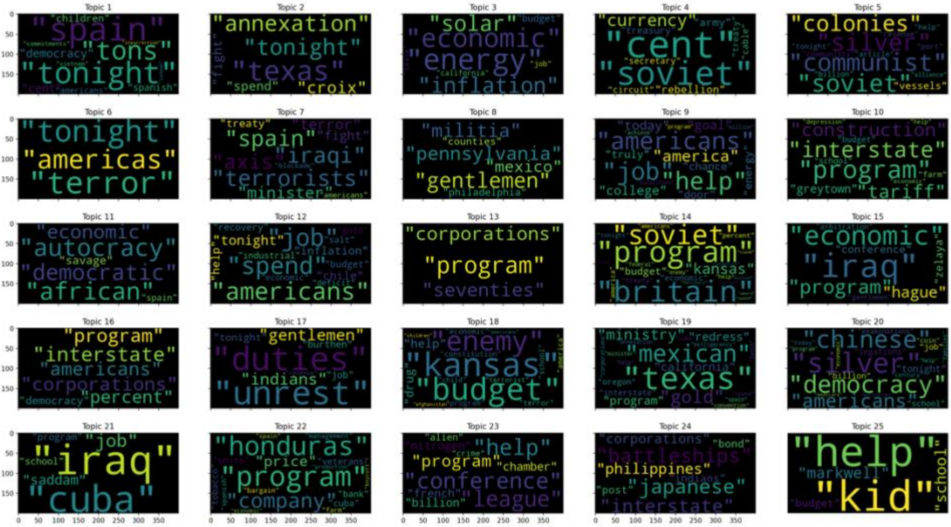

*Figure 4: LDA Topics*

Using the word clouds can help with our inference to annotate the 10 randomly selected topics below:

1) Topic 1: 0.001*"program" + 0.001*"communist" + 0.001*"help" + 0.001*"billion" + 0.001*"tons" + 0.001*"alliance" + 0.001*"budget" + 0.001*"americans" + 0.001*"percent" + 0.001*"storage"

   This topic can be annotated as Communist party and its influence on America.

2) Topic 4: 0.001*"mexico" + 0.001*"program" + 0.001*"coinage" + 0.001*"statute" + 0.001*"tonight" + 0.001*"economic" + 0.001*"silver" + 0.001*"america" + 0.001*"federal" + 0.001*"bank"

   Clear focus on currency. Bank, federal, silver all relate to funds. On going through the original data, it can be inferred that this topic relates to the early years after independence.

3) Topic 5: 0.001*"silver" + 0.001*"help" + 0.001*"program" + 0.001*"terrorists" + 0.001*"weapons" + 0.001*"school" + 0.001*"terrorist" + 0.001*"america" + 0.001*"afghanistan" + 0.001*"terror"

   This topic has a clear theme of war.

4) Topic 6: 0.001*"mexico" + 0.001*"texas" + 0.001*"mexican" + 0.001*"honduras" + 0.001*"french" + 0.001*"bank" + 0.001*"france" + 0.001*"redress" + 0.001*"treaty" + 0.001*"help"
   No clear annotation can be interpreted for this.

5) Topic 8: 0.001*"tonight" + 0.001*"inflation" + 0.001*"spend" + 0.001*"americans" + 0.001*"job" + 0.001*"help" + 0.001*"reform" + 0.001*"vessels" + 0.001*"iraq" + 0.001*"program"

   This topic relates to the time of economic crisis and inflation casuing more government spending and requiring reforms to make up for lost jobs.

6) Topic 9: 0.001*"interstate" + 0.001*"corporations" + 0.001*"battleships" + 0.001*"spend" + 0.001*"combinations" + 0.001*"economic" + 0.001*"scout" + 0.001*"help" + 0.001*"destroyers" + 0.000*"job"

   This topic relates to the time of industrialization as it talk abouts interstate corporations and how it helps the economy but may have led to destroying jobs for some.

7) Topic 16: 0.001*"job" + 0.001*"program" + 0.001*"budget" + 0.001*"tonight" + 0.001*"americans" + 0.001*"democracy" + 0.001*"help" + 0.001*"recovery" + 0.001*"economic" + 0.001*"spend"

   This topic talks about the budget and how it can help in economy;s recovery.

8) Topic 21: 0.001*"americans" + 0.001*"terrorists" + 0.001*"tonight" + 0.001*"fight" + 0.001*"vietnam" + 0.001*"help" + 0.001*"cuba" + 0.001*"terror" + 0.001*"axis" + 0.001*"america"

   This topic can be annotated to the Vietnam war.

9) Topic 24: 0.001*"ministry" + 0.001*"job" + 0.001*"program" + 0.001*"economic" + 0.001*"help" + 0.001*"chamber" + 0.000*"depression" + 0.000*"spain" + 0.000*"minister" + 0.000*"france"

This topic can be annotated to the time after World War 2 when the people may have suffered from depression because of the chambers used during world war 2.

10) Topic 25: 0.001*"soviet" + 0.001*"salt" + 0.001*"railroad" + 0.001*"nuclear" + 0.001*"loan" + 0.001*"inflation" + 0.001*"reconstruction" + 0.001*"corporation" + 0.000*"conference" + 0.000*"help"

    This topic is related to the time after cold war when in a conference a nuclear treaty may have been signed between America and soviet union which could further help reconstruct the economy.

Some of the Key differences noted between LDA and LSi are as follows:

- Topics were more short, crisp and to the point in LDA. LSI had more longer topics which would not exactly make sense each time.
- The plotted word clouds with 50 most relevant words each also show more to the point focus for each topic tile in LDA while there is a focus on multiple unrelated words in LSI.
- LDA is a more complex intuition based generative model and thus takes longer to train than LSI which is based on more statistical learning through matrix decomposition.
- As mentioned earlier, LDA is a probabilistic model while LSI is a deterministic model.
- LDA helps uncover the hidden topic distributions while LSI only captured the semantic relationships between the terms and documents.

## 4. Change in Speeches Over Time

The goal of this step is to summarize the changes in the State of the Union speech in each decade of the 20th and 21st centuries. First similar steps are used from all earlier stages to create a new dataset but this time the dataset created starts from the year 1901 and splits at 1910 while concatenating all speech of the year entries for the entire decade. The next entry would be for 1911 to 1920 and so on until 2011 – 2012 as 2012 is the last entry. This is done by visualizing the data and doing a manual split by using index values of the dataset and saved as a new data dictionary with key values corresponding to each decade.
Once we have the split dataset for all the decades of the 20th and 21st century, we redo the pre-processing steps on this new dataset using the genism library. After pre-processing the dataset, we create the tf-idf vectors as we did earlier and then use the LDA algorithm to perform the topic modeling on this new decade dataset.

Now we analyze the data to infer its relationship with historical events.

1. 1900-1910

Topic 1: 0.006*"people" + 0.005*"government" + 0.004*"work" + 0.004*"state" + 0.003*"world" + 0.003*"nation" + 0.003*"need" + 0.002*"congress" + 0.002*"years" + 0.002*"increase"

We see a theme of uplift – time after independence and before the war.

2. 1911-1920

Topic 2: 0.000*"rightist" + 0.000*"rocket" + 0.000*"reinvigorate" + 0.000*"renegotiation" + 0.000*"replacement" + 0.000*"resourcefulness" + 0.000*"retail" + 0.000*"retribution" + 0.000*"revenge" + 0.000*"rhine"

Time right before independence theme of rightist, retribution, resourcefulness all can be inferred as foreign policies.

3. 1921-1930

Topic 3: 0.000*"government" + 0.000*"people" + 0.000*"state" + 0.000*"world" + 0.000*"great" + 0.000*"work" + 0.000*"need" + 0.000*"nation" + 0.000*"free" + 0.000*"power"

’great’ could be inferred as a reference to the great economic depression around 1929.

4. 1931-1940

Topic 4: 0.000*"government" + 0.000*"public" + 0.000*"people" + 0.000*"state" + 0.000*"increase" + 0.000*"work" + 0.000*"congress" + 0.000*"need" + 0.000*"world" + 0.000*"power"

Policies and dealing with economic reforms.

5. 1941 - 1950

Topic 5: 0.000*"people" + 0.000*"government" + 0.000*"work" + 0.000*"state" + 0.000*"years" + 0.000*"nation" + 0.000*"world" + 0.000*"increase" + 0.000*"america" + 0.000*"soviet"

Soviet, increase people – all of it related to the war that started in 1939.

6. 1951-1960

Topic 6: 0.000*"work" + 0.000*"government" + 0.000*"people" + 0.000*"state" + 0.000*"need" + 0.000*"job" + 0.000*"know" + 0.000*"years" + 0.000*"nation" + 0.000*"time"

Foreign policies that the government would form for its citizens.

7. 1961- 1970

Topic 7: 0.000*"world" + 0.000*"people" + 0.000*"government" + 0.000*"military" + 0.000*"state" + 0.000*"free" + 0.000*"nation" + 0.000*"need" + 0.000*"increase" + 0.000*"program"

This decade had the Vietnam war going on. Which is why everything relates to government and reforms.

8. 1971-1980

Topic 8: 0.000*"government" + 0.000*"state" + 0.000*"public" + 0.000*"power" + 0.000*"people" + 0.000*"work" + 0.000*"great" + 0.000*"necessary" + 0.000*"nation" + 0.000*"result"

Again, most words relate to government and its reforms which was the case in reality as well.

9. 1981 - 1990

Topic 9: 0.000*"people" + 0.000*"government" + 0.000*"world" + 0.000*"state" + 0.000*"power" + 0.000*"work" + 0.000*"production" + 0.000*"increase" + 0.000*"nation" + 0.000*"great"

Similar data.

10. 1991-2000

Topic 10: 0.000*"people" + 0.000*"government" + 0.000*"world" + 0.000*"nation" + 0.000*"work" + 0.000*"need" + 0.000*"congress" + 0.000*"years" + 0.000*"state" + 0.000*"year"

11. 2001 - 2010

Topic 11: 0.000*"work" + 0.000*"people" + 0.000*"government" + 0.000*"world" + 0.000*"years" + 0.000*"support" + 0.000*"time" + 0.000*"tonight" + 0.000*"state" + 0.000*"know"

12. 2011 - 2020

Topic 12: 0.000*"rightist" + 0.000*"rocket" + 0.000*"reinvigorate" + 0.000*"renegotiation" + 0.000*"replacement" + 0.000*"resourcefulness" + 0.000*"retail" + 0.000*"retribution" + 0.000*"revenge" + 0.000*"rhine"

Summarizing the findings: We can perform further fine tuning of the LDA algorithm hyperparameters to get more accurate results. But we can definitely see proportions of adta that relates to each decade reflecting what was going on in reality in that decade. Most presidential speeches reflect the policies around everything happening during that time which we were able to capture using the LDA Topic Modeling.