Topic Modeling

Vibhore Singh
u3248455

Artificial Intelligence Techniques
University of Canberra

Abstract

In this project, we aim to perform topic modelling on state of the union speeches dataset using LSI and LDA algorithms on the dataset. Our goal to find meaningful topics from the topics. Also we try to analyse how the topics have changed over the time.

1 Introduction

Topic Modelling is the process of finding structure in the collection of texts using several statistical methods. The president of the US addresses the people at the start of every year known s state of the union . It defines the agenda of the government for that specific year, which is why it is a good way to judge which direction the country of USA is heading and what are the most important topics of discussion in the countryv on that time. Thus analysing the historical data from the state of the union corpus can be very helpful.

2 Methodology

2.1 Dataset

Our dataset contains two columns one containing the year and other representing the speech for the year. We have speeches from 1790 to 2012. In total we have 226 rows of data.

2.2 Data preprocessing

1. **Lowercase:** The entire text is converted to lowercase to ensure our data is uniform.

2. **Removing Numbers:** Removed all the numbers and digits from the text as they dont add a lot of value to the analuysis.

3. **Removing Punctuation:** Punctuation marks are removed because they are prety much useless in regards of topic modelling.

4. **Splitting:** Some words contain hyphens. so we split them into two different words.

5. **Tokenization :** We split our entire text of the speech in individual words called tokens

6. **Stopwords:** We removeed the stopwords such as "a", "the", "is", "and" etc as they dont add any value to the main content.

7. **Lemmatization :** Lemmatization is done to reduce words to their base form to ensure that all the forms of a same word are treated as same token. Also words occurring less than 2 times were removed.

2.3 Transformation :

1. **Creating a Bag-of-Words (BoW) :** Next we convert our text into numerical vectors. To use the machine learning algorithms on the text data we have to vectorize them and thus as a result bog of words model is used.

2. **TF-IDF Transformation :** to add more refinement and filtering to the bow data , we perform tfidf transformation. It address the issues related to term frequency and importance of the text.

2.4 Latent semantic indexing (LSI)

: Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text[1]. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings.

- First we find the optimum number of topics for LSI model by iterating over a range (5 to 31)of values for the topic number and checking coherence score. We find the optimum value to be 6.

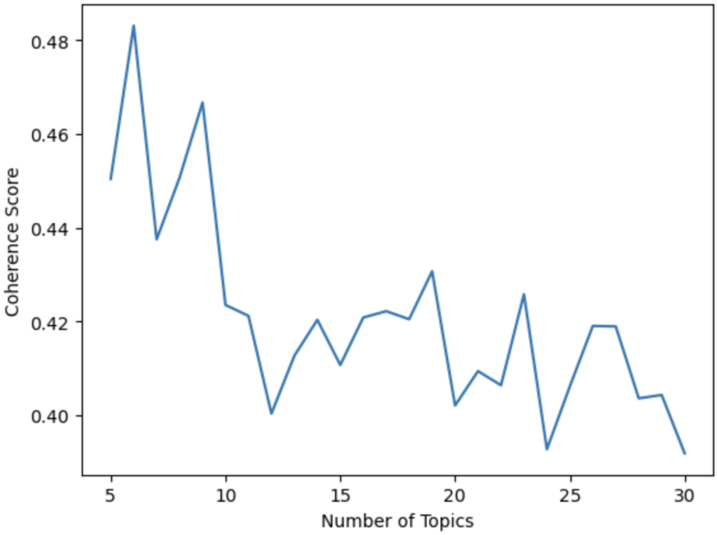


Figure 1: Coherence score of different number of topics

- Then we trained the LSI model on the TF-IDF vectors obtained from the corpus and using optimum number of topics.

- The model generated a set of topics, each represented as a list of word coefficients

```
LSI Topics:
[(0,
  '0.111*"program" + 0.080*"tonight" + 0.076*"job"'),
 ('+ 0.074*"economic" + 0.069*"budget" + 0.068*"help"'),
 ('+ 0.065*"america" + 0.061*"child" + 0.057*"treaty"'),
 ('+ 0.056*"today"'),
 (1,
  '-0.155*"tonight" + -0.154*"program" + -0.141*"job"'),
 ('+ -0.110*"help" + -0.107*"budget" + -0.101*"child"'),
 ('+ -0.089*"america" + -0.086*"economic" + -0.085*"school"'
 ('+ -0.084*"billion"'),
 (2,
  '0.187*"tonight" + -0.131*"program" + 0.122*"terrorist"'),
 ('+ -0.122*"economic" + 0.114*"job" + 0.108*"iraq"'),
 ('+ 0.108*"child" + 0.104*"thats" + -0.103*"farm"'),
 ('+ 0.092*"weve"'),
 (3,
  '-0.151*"silver" + 0.126*"program" + 0.108*"militia"'),
 ('+ -0.103*"gold" + -0.099*"interstate" + -0.098*"corporati
 ('+ 0.094*"communist" + -0.085*"cent" + 0.084*"soviet"'),
 ('+ -0.084*"circulation"'),
 (4,
  '0.244*"interstate" + 0.218*"corporation" + 0.141*"railroa
 ('+ -0.138*"mexico" + -0.121*"program" + 0.119*"combination
 ('+ -0.115*"soviet" + -0.097*"communist" + -0.093*"texas"')
 ('+ -0.091*"silver"'),
 (5,
  '-0.348*"terrorist" + -0.254*"iraq" + -0.223*"iraqi"'),
 ('+ -0.158*"terror" + -0.154*"enemy" + -0.142*"al"'),
 ('+ 0.122*"thats" + -0.114*"regime" + -0.111*"afghanistan"'
 ('+ 0.098*"job"')]
```

- we randomly sampled ten topics for closer analysis. here is the output for the sampled topics:

```
('Topic 3: 0.150*"silver" + -0.126*"program" + -0.109*"milit
 '+ 0.100*"interstate" + 0.097*"corporation" + -0.095*"commu
 '0.084*"cent" + -0.083*"soviet" + 0.083*"circulation" + -0.
 '0.078*"coinage" + 0.078*"per" + 0.075*"currency" + 0.075*"
 '0.069*"railway" + -0.068*"gentleman" + -0.068*"economic" +
 '0.067*"conference" + 0.065*"railroad" + -0.064*"object" +
 '0.063*"coin" + 0.063*"company" + 0.062*"bond" + 0.062*"chi
 '0.061*"nicaragua" + -0.061*"defense" + 0.061*"tonight" + 0
('Topic 5: -0.348*"terrorist" + -0.256*"iraq" + -0.225*"iraq
 '-0.158*"terror" + -0.155*"enemy" + -0.143*"al" + 0.123*"th
 '-0.114*"regime" + -0.111*"afghanistan" + 0.096*"job" + -0.
 '0.084*"deficit" + -0.084*"fighting" + -0.081*"weapon" + -0
 '0.080*"cut" + -0.080*"qaida" + -0.079*"qaeda" + -0.079*"hu
 '-0.077*"victory" + 0.076*"weve" + 0.076*"dont" + 0.074*"bu
 '-0.070*"attack" + -0.069*"coalition" + -0.068*"homeland" +
 '+ 0.067*"college" + 0.067*"im" + 0.063*"percent"')
```

```
('Topic 7: -0.236*"mexico" + -0.188*"texas" + 0.135*"silver" +
 '-0.134*"terrorist" + -0.126*"iraq" + -0.124*"mexican" + -0.1
 '0.100*"gold" + 0.098*"coinage" + -0.085*"agriculture" + 0.08
 '0.081*"insurgent" + 0.081*"soviet" + -0.078*"depression" + 0
 '0.077*"coin" + 0.073*"japanese" + 0.070*"enemy" + -0.068*"ta
 '-0.067*"agricultural" + 0.065*"fighting" + -0.065*"construct
 '-0.063*"al" + -0.063*"veteran" + 0.061*"insurrection" + -0.0
 '+ 0.060*"cuba" + 0.059*"circulation" + -0.059*"farmer" + -0
('Topic 1: -0.155*"tonight" + -0.154*"program" + -0.141*"job"
 '+ -0.107*"budget" + -0.101*"child" + -0.089*"america" + -0.0
 '-0.085*"school" + -0.084*"billion" + -0.083*"percent" + -0.0
 '-0.083*"today" + -0.079*"let" + -0.077*"soviet" + -0.076*"sp
 '-0.075*"cut" + 0.075*"mexico" + 0.074*"spain" + -0.071*"goal
 '-0.071*"thats" + -0.068*"challenge" + 0.066*"vessel" + 0.065
 '-0.064*"deficit" + 0.064*"minister" + -0.064*"nuclear" + 0.0
 '+ -0.060*"tax" + 0.060*"claim"')
('Topic 9: -0.154*"enemy" + -0.124*"mexico" + 0.113*"interstat
 '0.108*"corporation" + -0.106*"japanese" + -0.101*"fighting"
 '+ 0.097*"vietnam" + -0.096*"thats" + -0.087*"texas" + 0.087*
 '-0.081*"hitler" + 0.080*"terrorist" + -0.077*"california" +
 '+ 0.069*"gentleman" + -0.067*"cent" + -0.065*"nazi" + -0.065
 '-0.064*"navy" + -0.063*"tank" + 0.062*"nuclear" + 0.061*"com
 '-0.061*"mile" + -0.061*"plane" + -0.061*"get" + -0.060*"axis
 '0.059*"soviet" + 0.059*"communist" + -0.058*"naval"')
('Topic 8: 0.156*"currency" + 0.150*"gold" + 0.147*"silver" +
 '0.119*"circulation" + 0.118*"specie" + 0.115*"coin" + 0.107*
 '0.105*"paper" + 0.101*"constitution" + -0.096*"arbitration"
 '0.092*"coinage" + 0.089*"bank" + -0.082*"interstate" + 0.082
 '0.080*"per" + -0.077*"french" + 0.076*"tender" + -0.075*"chi
 '0.075*"rebellion" + 0.074*"cent" + -0.073*"japanese" + -0.07
 '-0.072*"railway" + -0.071*"chile" + -0.070*"china" + -0.065*
 '0.065*"note" + -0.064*"nicaragua" + -0.063*"france"')
('Topic 6: 0.334*"mexico" + 0.281*"texas" + 0.162*"mexican" +
 '0.128*"interstate" + 0.119*"corporation" + -0.106*"silver" +
 '0.104*"annexation" + -0.100*"terrorist" + 0.094*"california"
 '+ -0.085*"per" + -0.084*"cent" + 0.080*"railroad" + 0.075*"i
 '-0.073*"iraqi" + 0.070*"oregon" + 0.068*"man" + -0.066*"coin
 '0.063*"territory" + -0.062*"militia" + 0.060*"combination" +
 '-0.059*"agriculture" + 0.057*"supervision" + -0.054*"depress
 '0.052*"constitution" + -0.050*"debt" + 0.050*"shipper" + -0
 '+ -0.048*"estimated" + -0.048*"loan"')
('Topic 2: 0.188*"tonight" + -0.131*"program" + 0.122*"terrori
 '-0.122*"economic" + 0.114*"job" + 0.108*"iraq" + 0.107*"chil
 '0.105*"thats" + -0.103*"farm" + 0.093*"weve" + -0.090*"inte
 '-0.089*"communist" + -0.087*"corporation" + 0.085*"school" +
 '-0.084*"industrial" + 0.079*"iraqi" + 0.078*"america" + 0.07
 '-0.073*"problem" + 0.073*"college" + -0.073*"agriculture" +
 '+ 0.069*"drug" + 0.067*"medicare" + -0.064*"price" + 0.064*"
```

```
  '0.064*"let" + 0.063*"im" + -0.062*"objective" + 0.062*"cut
 ('Topic 0: 0.111*"program" + 0.080*"tonight" + 0.076*"job" +
  '+ 0.069*"budget" + 0.068*"help" + 0.065*"america" + 0.061*
  '0.057*"treaty" + 0.056*"today" + 0.056*"tax" + 0.056*"bill
  '0.055*"mexico" + 0.055*"let" + 0.052*"school" + 0.052*"upo
  '0.051*"federal" + 0.051*"problem" + 0.051*"spain" + 0.050*
  '0.049*"soviet" + 0.049*"percent" + 0.048*"vessel" + 0.047*
  '0.047*"claim" + 0.045*"british" + 0.044*"goal" + 0.044*"mi
  '0.043*"treasury" + 0.043*"cut"')
 ('Topic 4: -0.243*"interstate" + -0.219*"corporation" + -0.1
  '0.138*"mexico" + 0.123*"program" + -0.118*"combination" +
  '0.100*"communist" + 0.094*"silver" + 0.093*"texas" + -0.08
  '0.079*"mexican" + -0.075*"militia" + -0.074*"shipper" + -0
  '+ -0.072*"terrorist" + -0.070*"man" + 0.069*"treaty" + 0.0
  '-0.066*"railway" + -0.065*"gentleman" + -0.058*"iraq" + -0
  '0.058*"award" + 0.058*"chinese" + 0.057*"gold" + 0.056*"cu
  '-0.056*"rebate" + -0.054*"employee" + 0.054*"billion"')
```

- Annotation of Sampled Topics based on the dominant words weight.

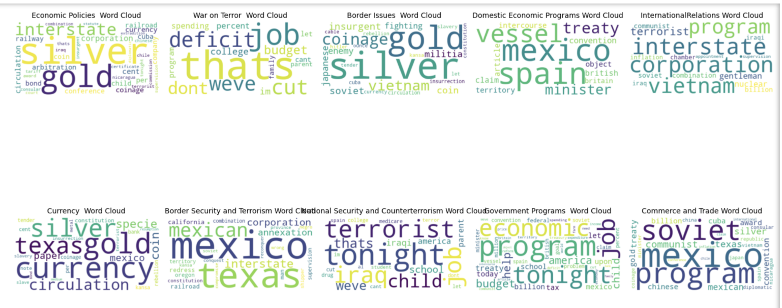
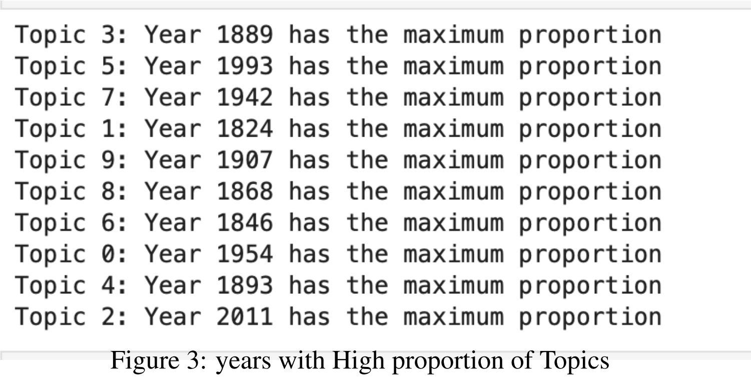


Figure 2: word cloud of random topics with annotations

- We can clearly see some of the topics have a dominant similar theme i.e more interpreteble with real humans concepts. On the other hand some were really random and difficult to interpret.
- We check for the year that has high proportions of the topic and check the speech of that year.

Annotations :

1. Topic 3 was annotated a Economic policies and when we check the year it was highest proportion in 1889 which was an year of economic reforms all aaround the world.
2. Topic 5 was named as War on terror and the year was 1993 which was the year of WTC bombing. The presence of words such as enemy, afganistan , attack, terrorist etc helps us annotate this.
3. Topic 7 had border issues due to the year 1942 which was the middle of world war 2 and there was a huge displacement of people.



```
Topic 3: Year 1889 has the maximum proportion
Topic 5: Year 1993 has the maximum proportion
Topic 7: Year 1942 has the maximum proportion
Topic 1: Year 1824 has the maximum proportion
Topic 9: Year 1907 has the maximum proportion
Topic 8: Year 1868 has the maximum proportion
Topic 6: Year 1846 has the maximum proportion
Topic 0: Year 1954 has the maximum proportion
Topic 4: Year 1893 has the maximum proportion
Topic 2: Year 2011 has the maximum proportion
```

Figure 3: years with High proportion of Topics

4. topic 4 was quite arbitary and unclear.

5. topic 9 was named international telations due to the presence of words such as hitler, soviet, mexico,japanese etc.

6. Topic 8 was annotated to Currency, highest in 1868 due to the several monetary terms but it si quite abstract .

7. Topic 6 was named Border Security and Terrorism due to the similarly themed words.

8. Topic 0 was named to Government Programs as in 1954 speech there was a lot of mention of new programs

9. Topic 4 gave us commerce and trade info and the year was 1893 as president cleveland talked a lot about banks , trade with mexico and other countries , finance treaties with other countries .

10. Topic 2: National Security and Counterterrorism had the most proportion in 2011 as this year President obama talked extensively about Iraq, AlQaeda, Intelligence, Taliban, Afghanistan and Pakistan. This was also the year when Osama was killed.

2.5 LDA :

We followed the similar steps and found the following connections

1. Topic 26: Economic Policy and Finance 1843 as President John Tyler's speech predominantly focused on various economic aspects such as international trade relations, fiscal policies, public debt management, currency stability, and the role of the government in economic matters.

2. Topic 6 as Pres Monroe in 1821 outlined the foreign relations and commercial policies

3. Topic 10 is annotated Foreign policy as in 1974 Nixon Talked about ending the vietnam war , improving relationship with USSR and China.

4. Topic 15 is not clear as the words and the speech didn't match

5. Topic 28 was annotated as National security as President Washington talked in his two addresses about strengthening the country and the progress of the nation

6. Topic 12 is named international relations and is mapped to 1941 as it was the year of world war 2 and Pres Roosevelt talked about war, allies and sacrifices quite extensively.

7. topic 13 is annotated as Social Issues as President Clinton in his 1996 address discusses various social issues, including strengthening families, tackling crime, improving education

8. Although topic 19 is very subtle we named it National defence due the the president Washington's address of Indian hostilities and the violence

9. President Carter talked about the increased employment of black teen, more appointment of coloured people in government jobs and strengthning of minority business owners in his address of 1981. Thus we annotated it as civil rights.

10. Topic 20: President eisenhower in 1953 talked about infrastructure development , economic growth and resource development.

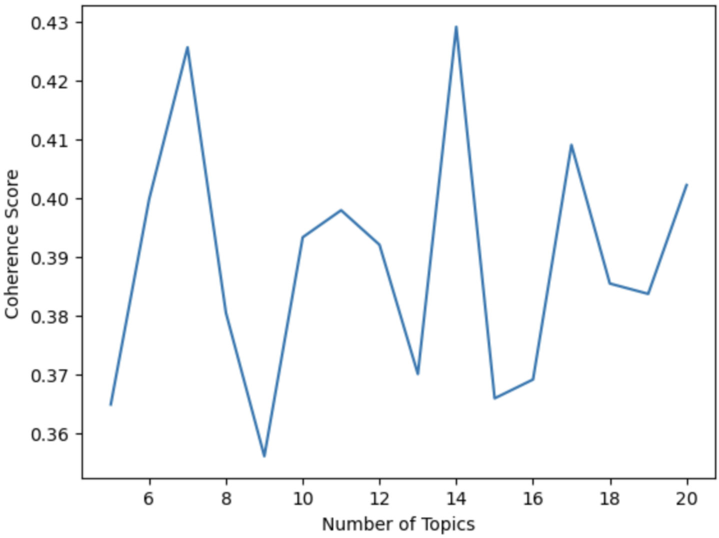


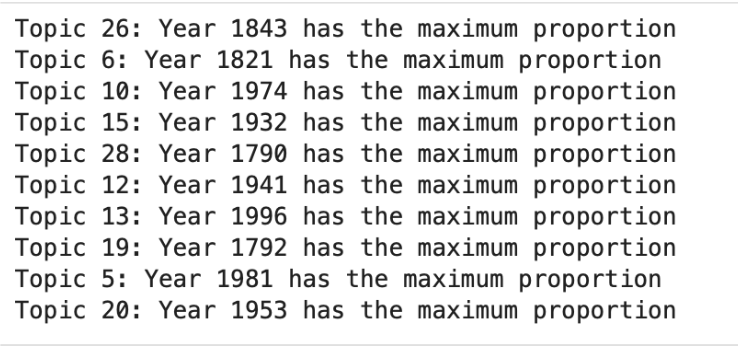
Figure 4: Coherence score of different number of topics

2.6 Difference between LDA and LSI

1. LDA aims to find the topics discussed in a set of documents while LSI aims to find the relationship between text.

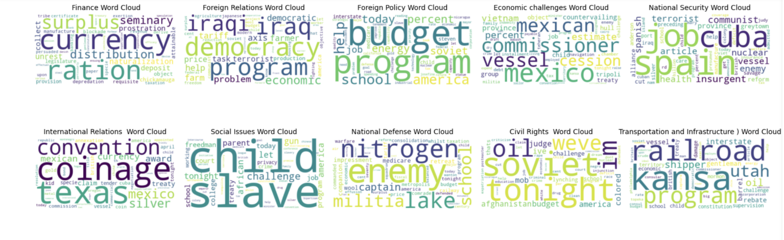
2. LSI works faster than LDA

3. Lsi workd on SVD(singular value decomposition) to. transform the high dimensional matrix into low dimensional where as LDA works on bayesian approach and every doccument is cosidered as a mixture of different topics and topis as a set of different words. These words define what a topic is about.



```
Topic 26: Year 1843 has the maximum proportion
Topic 6: Year 1821 has the maximum proportion
Topic 10: Year 1974 has the maximum proportion
Topic 15: Year 1932 has the maximum proportion
Topic 28: Year 1790 has the maximum proportion
Topic 12: Year 1941 has the maximum proportion
Topic 13: Year 1996 has the maximum proportion
Topic 19: Year 1792 has the maximum proportion
Topic 5: Year 1981 has the maximum proportion
Topic 20: Year 1953 has the maximum proportion
```

Figure 5: years with High proportion of Topics



4. In our program LDA worked better than LSI as LDA inferred much clearer information.

3 Decade Summarization:

In this part we take the data from 1900 to 2012 and group them into decades. Then we perform LDA on the decade text to try to understand what was the major issues and major discourse in that decade and we can study how the speeches have changed over the last century. Decade :1900 represents the decade starting from 1900 and ending with 1909

We generate topics for each of the decade and choose the most dominant one as the main theme of that decade and anotate them accordingly.

1. The decade of 1900 has shown 84.9% probability for topic 3. The topic 3 has words like arbitration, decree, tribunal etc all of which focus on legal and International relations. This refer to the fact that USA signed **legal treaties** with Haiti, Panama, Cuba, Urugay,Nicaragua, Portugal etc in the first decade of 20th century. Along with that first and second hague conventions also happend around that time.

2. 1910s had topic 8 as the most dominant topic with 78% probability. Topic 8 had dominance of words like railway, development, infrastructure etc. we annotated it as Infrastructure development. This could be due to the fact that US made **Panama canal** in 1914

3. In 1920 Topic 13 was dominant which represented International conflict and terrorism

due to the presence of words such as communist, armament etc. This was the time of rise of **facism** in europe and and post war world in 1920s was plagued with **global unrest** and violence.

4. For 1930s the topic 0 with 89% was major topic representing the World war 2 , Dictatorship and economic unrest. The USA was reeling through **the great depression** of 1929 and also the **world war 2** had started in europe. Words such as hitler, axis etc were used.

5. Again for 1940s as well the topic 0 with 88% was prominent as it was a decade of war, **Holocaust**, **atomic bomb**, **end of colonialism** etc.

6. In 1950s , topic 5 was the main topic of speeches Words such as soviet, communist, nuclear give us the theme as Geographical Tensions . This was the decade of post war tension and security issues. **Cold war** started in this decade, **Korean war** was during these

7. Topic 10 clearly shows words such as communist , missile, soviet, vietnam making it easier to identify it as Cold war period. In 1960 the **vietnam war** was on full fledge, **Cuban missile crisis** happened in 1962, **East - West Germany issues**

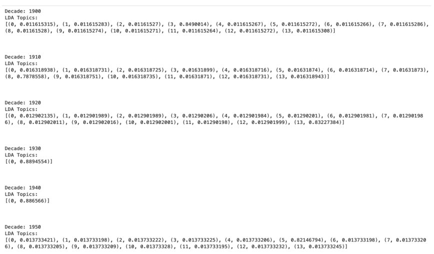
8. In 1970s, energy crisis(**oil crisis 1973**) was a big issue in US as well as the infamous**watergate**. These were covered in topic 12 with words such as powerplants, nuclear, solar, involvement, confrontation etc.

9. 1980s most dominant topic was topic 5 again. This is due to the USSR's **Afghanistan invasion** of 1989 and the **Iran-Iraq war**

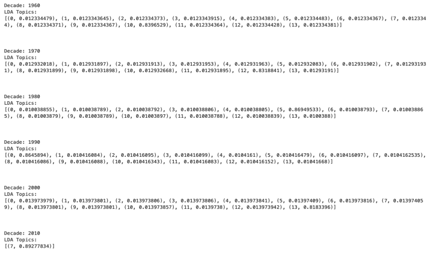
10. In 1990 The modt dominant topic was topic 0 again having 86% due to the presence of words like Iraq, Saddam, Invasion etc. This is due to the **Gulf War** in 1990, Breaking up of **USSR** and **Yugoslavia**

11. 2000s were topic 13 dominant again with 81%named International conflict and terrorism as this was the decade of **9/11 attack**, **Afghanistan war**, **Us invasion of Iraq** etc.

12. The 2010s represent the rise of social media, internet usage, and coming out of **2008 financial depression** . The dominant words were technology, internet, trillion, millionaire which helps us to annotate as Technological and economical policies with 89% probability



```
Decade: 1900
LDA Topics:
[(0, 0.011615315), (1, 0.011615283), (2, 0.01161527), (3, 0.8490014), (4, 0.011615267), (5, 0.011615272), (6, 0.011615266), (7, 0.011615286),
(8, 0.01161528), (9, 0.011615274), (10, 0.011615271), (11, 0.011615264), (12, 0.011615272), (13, 0.011615308)]


Decade: 1910
LDA Topics:
[(0, 0.016318938), (1, 0.016318731), (2, 0.016318725), (3, 0.01631899), (4, 0.016318716), (5, 0.01631874), (6, 0.016318714), (7, 0.01631873),
(8, 0.7878558), (9, 0.016318751), (10, 0.016318735), (11, 0.01631871), (12, 0.016318731), (13, 0.016318943)]


Decade: 1920
LDA Topics:
[(0, 0.012902135), (1, 0.012901989), (2, 0.012901989), (3, 0.01290206), (4, 0.012901984), (5, 0.01290201), (6, 0.012901981), (7, 0.01290198
6), (8, 0.012902011), (9, 0.012902016), (10, 0.012902001), (11, 0.01290198), (12, 0.012901999), (13, 0.83227384)]


Decade: 1930
LDA Topics:
[(0, 0.8894554)]


Decade: 1940
LDA Topics:
[(0, 0.886566)]


Decade: 1950
LDA Topics:
[(0, 0.013733421), (1, 0.013733198), (2, 0.013733222), (3, 0.013733225), (4, 0.013733206), (5, 0.82146794), (6, 0.013733198), (7, 0.01373320
6), (8, 0.013733205), (9, 0.013733209), (10, 0.01373328), (11, 0.013733195), (12, 0.013733232), (13, 0.013733245)]
```



```
Decade: 1960
LDA Topics:
[(0, 0.012334479), (1, 0.0123343645), (2, 0.012334373), (3, 0.0123343915), (4, 0.012334383), (5, 0.012334483), (6, 0.012334367), (7, 0.012334
4), (8, 0.012334371), (9, 0.012334367), (10, 0.8396529), (11, 0.012334364), (12, 0.012334428), (13, 0.012334381)]


Decade: 1970
LDA Topics:
[(0, 0.012932018), (1, 0.012931897), (2, 0.012931913), (3, 0.012931953), (4, 0.012931963), (5, 0.012932083), (6, 0.012931902), (7, 0.01293193
1), (8, 0.012931899), (9, 0.012931898), (10, 0.012932668), (11, 0.012931895), (12, 0.8318841), (13, 0.01293191)]


Decade: 1980
LDA Topics:
[(0, 0.010038855), (1, 0.010038789), (2, 0.010038792), (3, 0.010038806), (4, 0.010038805), (5, 0.86949533), (6, 0.010038793), (7, 0.01003886
5), (8, 0.01003879), (9, 0.010038789), (10, 0.01003897), (11, 0.010038788), (12, 0.010038839), (13, 0.0100388)]


Decade: 1990
LDA Topics:
[(0, 0.8645894), (1, 0.010416084), (2, 0.010416095), (3, 0.010416099), (4, 0.0104161), (5, 0.010416479), (6, 0.010416097), (7, 0.0104162535),
(8, 0.010416086), (9, 0.010416088), (10, 0.010416343), (11, 0.010416083), (12, 0.010416152), (13, 0.01041668)]


Decade: 2000
LDA Topics:
[(0, 0.013973979), (1, 0.013973801), (2, 0.013973806), (3, 0.013973806), (4, 0.013973841), (5, 0.01397409), (6, 0.013973816), (7, 0.01397405
9), (8, 0.013973801), (9, 0.013973801), (10, 0.013973857), (11, 0.0139738), (12, 0.013973942), (13, 0.8183396)]


Decade: 2010
LDA Topics:
[(7, 0.89277834)]
```

4 Conclusion

1. The speech themes are evolving from infrastructure and legal advancements to technical advancements.
2. Topics such as Terrorism , International tensions and Wars remain similar throughout the corpus
3. The events happening good or bad impact heavily to the president's address
4. It also infers that the events happening in the last decade also affect the next decade discourse.

References

[1] Latent semantic analysishttps://en.wikipedia.org/wiki/Latent_semantic_analysis#:~:text=Latent%20semantic%20indexing%20(LSI)%20is,an%20unstructured%20collection%20of%20text.

[] https://radimrehurek.com/gensim/auto_examples/core/run_topics_and_transformations.html

[] https://github.com/kapadias/medium-articles/blob/master/natural-language-processing/topic-modeling/Evaluate%20Topic%20Models.ipynb

[] https://investigate.ai/text-analysis/introduction-to-topic-modeling/#Attempt-two:-State-of-the-Union-addresses