Topic Modeling

Kirtimaan Gogna
2019AIM1014@iitrpr.ac.in
Indian Institute of Technology
Ropar

Abstract

In this project we decide to do topic modeling on the dataset of state of the union. The dataset is a csv file which has two columns, the columns are year and speech. The first column gives the year and the second columns give the corresponding speech for that year. We tend to apply natural language prcocessing by trying to understand the dataset and then proceed to find the prominent topics in the speeches. We tend to perform different experiments on the dataset.

1 Preprocessing the data

The dataset given to us is a csv file. It has columns with year and text. There are some years which has more than one speech.

First we will try to read the data with the help of pandas. Then we will try to make a corpus out of it. The corpus will contain all the speeches of different years text speeches appended to make the dataset corpus. Then we will try to pre process the dataset by removing the stop words, lammetization and stemming it. Although these are necessary steps in the text pre processing but we will only remove the stop words from the dataset and make a new dataset devoid of the corpus. The stopwords are the words that do not add any vslue to thr original dataset. Without them we can almost fully retain the inherent meaning of the text. now that we have our dataset pre processed by removing the stop words. We will convert it into the bag of words representation. This will convert each of the speech in a vector. now that we have the speeches in the form of a vector we will be able to leverage the power of linear algebra on it. For creating the bag of words representation we need to go through the dataset twice , first to create the dictionary and then again to create the bag of words representation. Tokenization of the document is also being done implicitly. After this we feed it to the gensim package to generate the tfidf scores.

2 LSI Topic Modeling

After this we apply Latent Semantic Indexing to the tf-idf vectors. We will try to find out the best number of topics. We will try to find it out with the help of coherence values. Although it is not always necessary that the topic highest having the highest coherence value is the best. Therefore we will try to chose some good value for it. From the experimentation and ranging the loop from 2 to 20, we found that the best value of the optimal number of topic for the lsi model turned out to be 4.

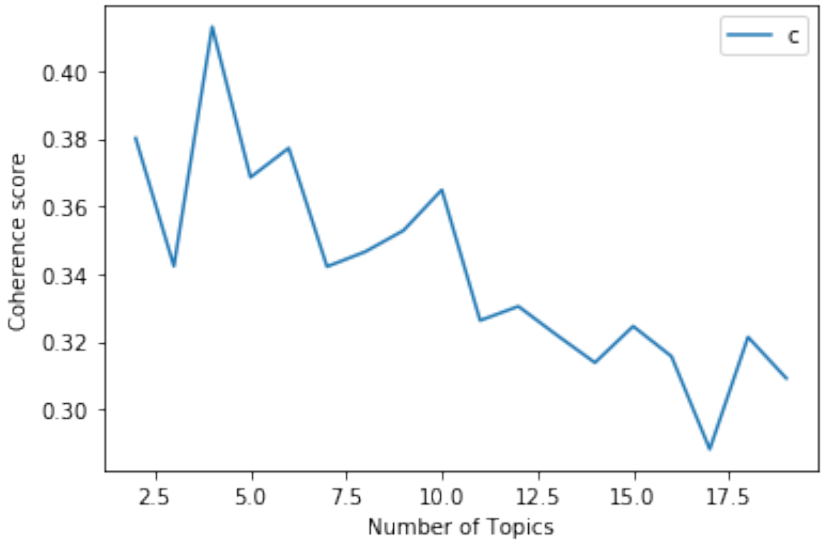

Figure 1: Optimal value for number of topics

Now we try and analyse the topics that came in top 4 if we apply the number of topics to be 4.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.9999	will, united, government, states, upon, may, p...	[state, union, address, george, washington, de...
1	1	1.0	0.9999	will, united, government, states, upon, may, p...	[state, union, address, george, washington, oc...
2	2	1.0	0.9999	will, united, government, states, upon, may, p...	[state, union, address, george, washington, no...
3	3	1.0	0.9999	will, united, government, states, upon, may, p...	[state, union, address, george, washington, de...

Figure 2: Topics with words

```
[(0,
  '0.082*"-" + 0.071*"upon" + 0.054*"economic" + 0.049*"treaty" + 0.047*"help" + 0.045*"treasury" + 0.045*"silver" + 0.045*"per" + 0.044*"program" + 0.
d"'),
 (1,
  '-0.119*"help" + -0.116*"tonight" + -0.103*"economic" + -0.101*"budget" + -0.101*"americans" + -0.097*"--" + -0.092*"program" + -0.091*"programs" + -
oviet" + -0.087*"jobs"'),
 (2,
  '-0.943*"-" + -0.055*"+" + -0.044*"--" + -0.038*"five-twenties" + -0.035*"consols" + -0.026*"1882," + -0.026*"consular" + -0.026*"1864" + -0.023*"188
0.022*"jobs"'),
 (3,
  '-0.162*"--" + -0.106*"tonight" + 0.089*"interstate" + 0.077*"economic" + -0.076*"tonight," + 0.071*"industrial" + -0.069*"americans" + -0.069*"jobs"
*"farm" + 0.068*"program"')]
```

Figure 3: Topics with words

We will try to make the word cloud for the number of dominant keywords in topics

After performing this experimentation.

We will try to sample 10 topics randomly from the dataset. These are the topics that we get. We will try to analyse these topics.

2.1 Topics

Topic 1

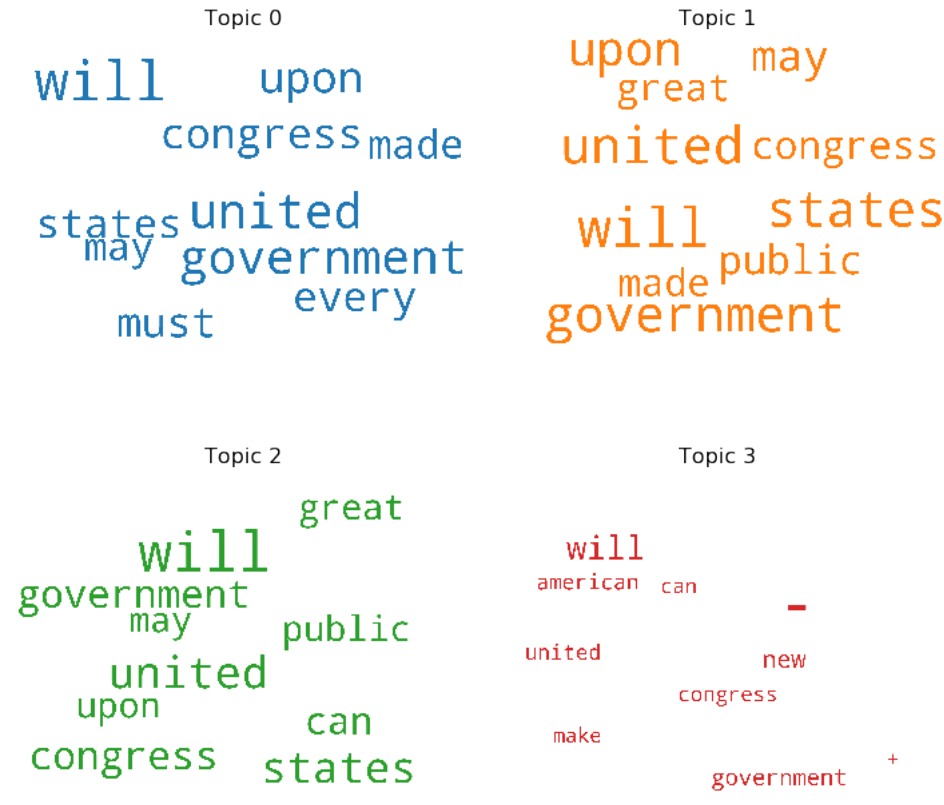

Figure 4: Word clouds for the topics

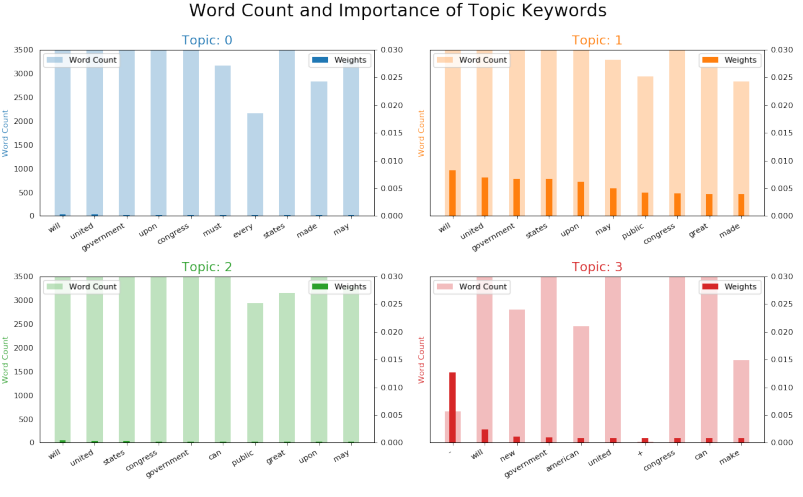

Figure 5: importance of words in topics

’0.082*"-" + 0.071*"upon" + 0.054*"economic" + 0.049*"treaty" + 0.047*"help" + 0.045*"treasury" + 0.045*"silver" + 0.045*"per" + 0.044*"program" + 0.043*"gold"

Description: To improve the economic condition use help of treaty by making program for silver and gold. By looking at the original speech we can see that the topics captured tell a little different idea than what our model has captured.

Topic 2

’-0.119*"help" + -0.116*"tonight" + -0.103*"economic" + -0.101*"budget" + -0.101*"americans" + -0.097*"–" + -0.092*"program" + -0.091*"programs" + -0.089*"soviet" + -0.087*"jobs"

Description: Programs for the economic budget for the jobs for americans and the soviet.When looking at the original document we can see that the speech does has some similar reforms and programs for inmproving the economy for which it has applied various reforms.

Topic 3
’-0.943*"-" + -0.055*"+" + -0.044*"–" + -0.038*"five-twenties" + -0.035*"consols" + -0.026*"1882," + -0.026*"consular" + -0.026*"1864" + -0.023*"1883," + -0.022*"jobs"’

Description: Twenty five British government securities without redemption date and with fixed annual interest relating to the consul or consulate in a foreign city. The topics captured by the model is a little vague and straight from looking at it , we can see it does not make any sense as such. But after looking at the original speech we can capture the true meaning with the help of these words.

Topic 4
-0.162*"–" + -0.106*"tonight" + 0.089*"interstate" + 0.077*"economic" + -0.076*"tonight," + 0.071*"industrial" + -0.069*"americans" + -0.069*"jobs" + 0.068*"farm" + 0.068*"program"

Description: Reforms made for industries for americans to create jobs that can be put in action between states. This topic keywords does capture the true meaning of the speech.

Topic 5

’-0.219*"–" + 0.122*"-" + -0.112*"silver" + 0.110*"program" + 0.105*"soviet" + 0.104*"economic" + 0.101*"communist" + -0.087*"gold" + 0.079*"atomic" + -0.068*"per"

Description: Program for soviet economic communist gold. From the document there is some mention of these words. But it is difficult to capture the true meaning from the document.

Topic 6

’0.253*"–" + -0.138*"silver" + -0.128*"gold" + 0.126*"interstate" + -0.111*"soviet" + -0.081*"notes" + 0.077*"forest" + -0.072*"specie" + -0.069*"mexico" + -0.069*"programs"’

Description: Reference to the trade between states. Trading of silver and gold.
Topic 7

'0.588*"–" + -0.137*"terrorists" + -0.120*"iraqi" + -0.109*"terrorist" + -0.089*"iraq" + -0.080*"saddam" + 0.076*"jobs" + -0.076*"terror" + -0.070*"al" + -0.062*"iraqis"

Description: Terrorism in Iraq headed by Saddam. Saddam heading the iraqis terrorists. This topic does capture the important keywords very well. This is the only topic I found that truly captures the meaning of the speech. By looking at the speech, the speech has some sentences related to it which our model captured very well.

Topic 8

'-0.146*"mexico" + -0.132*"texas" + 0.115*"silver" + -0.086*"mexican" + -0.083*"–" + -0.074*"mexico," + -0.073*"kansas" + -0.071*"1859," + 0.068*"per" + 0.065*"gold"
Description: Mexico putting taxes on silver and gold. This topic is very vague and does not capture the true meaning of the speech.

Topic 9

-0.193*"–" + -0.185*"iraqi" + -0.181*"terrorists" + -0.152*"iraq" + -0.149*"al" + -0.133*"terrorist" + -0.103*"qaeda" + -0.091*"iraq," + -0.087*"iraqis" + 0.083*"vietnam"

Description: Reference to the unrest between Iraq and America because of terrorism and some connection to the previous war between vietnam and America.
Topic 10

'-0.299*"–" + -0.117*"soviet" + -0.092*"japanese" + -0.091*"communist" +
-0.086*"fighting" + -0.064*"gold" + 0.062*"bank" + -0.062*"enemy" + -0.060*"hitler" + -0.058*"silver"

Description: Reference to the time of the world war 1 or world war 2. Maybe the reference to the war between Japanese and Germany headed by Adolf Hitler. Plundering of banks and golds can also be inferred.

3 LDA topic modeling

After performing the analysis on the dataset with the lsi model. We will try the LDA topic modeling on it. LDA uses a slightly different algorithm to find the topics, it basically reverse engineers the process of finding it. We can again use the Gensim for it. We will now try analyse the topics generated by the LDA model.

3.1 LDA Topics analysis

Topic: 0

Words: 0.007*"government" + 0.006*"states" + 0.006*"congress" + 0.005*"people" + 0.005*"war" + 0.005*"year" + 0.005*"world" + 0.004*"great" + 0.004*"public" + 0.004*"united"

Description: This is about war.
Topic: 1

Words: 0.010*"government" +
0.006*"states" + 0.006*"united" +
0.006*"congress" + 0.004*"new" + 0.004*"year" + 0.004*"" + 0.004*"people" + 0.004*"years" + 0.004*"time"

Description: From all the subsequent topics that we are going to see. From this we can see that this topic is about US Government congress.

Topic: 2

Words: 0.009*"states" + 0.008*"government" + 0.007*"united" + 0.005*"congress" + 0.005*"great" + 0.005*"year" + 0.004*"people" + 0.004*"public" + 0.004*"new" + 0.003*"war"

Description: Similar to the previous topic, we see that this topic is almost same with the previous one. It is about US state government.

Topic: 3

Words: 0.010*"government" + 0.010*"states" + 0.007*"united" + 0.006*"congress" + 0.005*"people"
+ 0.005*"country" + 0.005*"public" + 0.004*"great" + 0.004*"year" + 0.004*"time"

We can notice a pattern in all the previous and some of the upcoming topics. It is that almost all the topics are capturing almost similar Idea. We can conclude that maybe LDA is trying to capture the overall idea of the corpus.

Topic: 4

Words: 0.011*"government" + 0.007*"states" + 0.006*"" + 0.005*"united" + 0.005*"congress" + 0.004*"great" + 0.004*"year" + 0.004*"country" + 0.004*"time" + 0.004*"people"

Description: Vague description about the topic and it is similar to previous topics.

Topic: 5

Words: 0.012*"states" + 0.010*"government" + 0.007*"united" + 0.007*"congress" + 0.005*"country" + 0.005*"great" + 0.004*"year" + 0.004*"people" + 0.004*"public" + 0.003*"pres

Description: From this topic keywords we can see that this topic is about the united congress government, addressing the public in present.

Topic: 6

Words: 0.009*"united" + 0.009*"states" + 0.006*"war" + 0.006*"public" + 0.006*"congress" + 0.005*"government" + 0.004*"year" + 0.004*"great" + 0.004*"country" + 0.004*"time"

Description: The topic keywords tel that the US is now in peace and that the present time is devoid of war and living in a great time.

Topic: 7

Words: 0.007*"government" + 0.007*"year" + 0.006*"new" + 0.006*"people" + 0.006*"cong + 0.005*"states" + 0.004*"work" + 0.004*"united" + 0.004*"country" + 0.004*"world"

Description: From the given topic keywords it is very difficult to infer and capture the true meaning of the speech. But by looking at the corresponding speeches we can see that the speech is a general address to the public.

Topic: 8

Words: 0.008*"government" + 0.007*"congress" + 0.007*"united" + 0.007*"states" + 0.005*"people" + 0.004*"war" + 0.004*"public" + 0.004*"new" + 0.004*"year" + 0.003*"great"

Description:
The topic is very vague, captures very less information.

Topic: 9

Words: 0.008*"government" + 0.006*"congress" + 0.006*"people" + 0.005*"american" + 0.005*"year" + 0.005*"states" + 0.005*"world" + 0.004*"years" + 0.004*"new" + 0.004*"time"

Description: Very vague, similar to other topics.

4 Understanding the changes made in State of the union in each decade of 20th and 21st century

In this section we will try to summarize how the topics are changing in each decade from 20th century. We are going to have a temporary dataframe that will hold the required data for the particular decade.

4.1 Decade summarisation algorithm

```
lsi_topics=[]
```

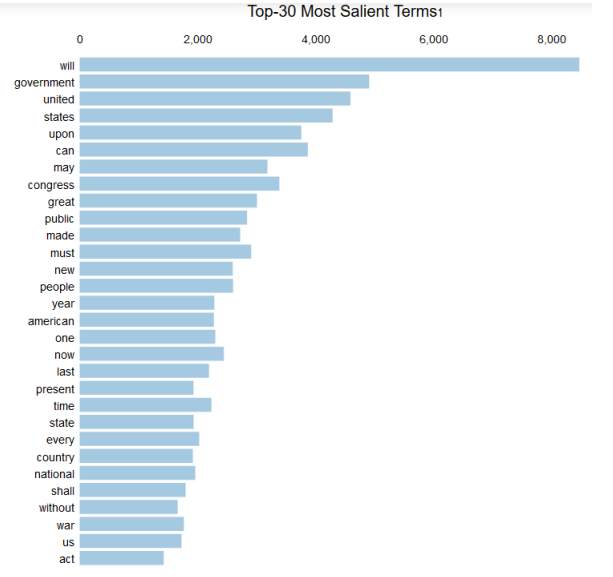

Figure 6: Words in topics 1 LDA

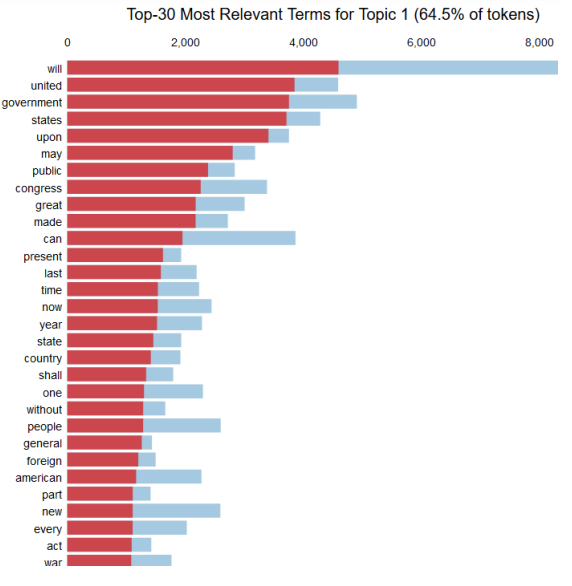

Figure 7: Words in topic 2 LDA

```
for i in range(0,115,10):
    ddf=[]
    corpus=[]
    texts=[]
    processed_corpus=[]
    for j in range(i,i+10):
```

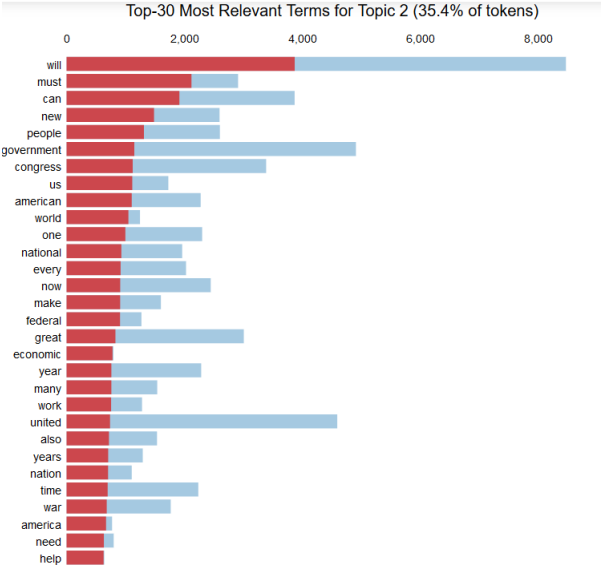

Figure 8: Words in topics 3 LDA

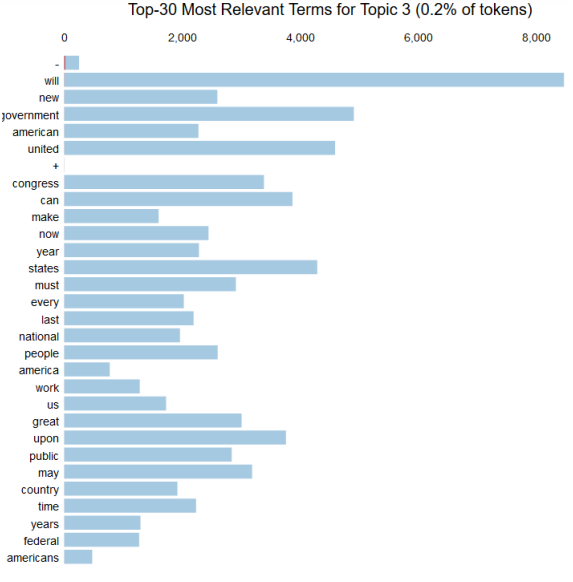

Figure 9: Words in topic 4 LDA

```
    if(j<115):
        ddf.append(cent_data_df.iloc[j])
        ddf_df=pd.DataFrame(ddf,columns=['year','speech'])
print(ddf_df)
for k in ddf_df.speech:
```

```
        corpus.append(k)
    texts = [[word for word in document.lower().split()
    if word not in stoplist]
    for document in corpus]
    frequency = defaultdict(int)
    for text in texts:
        for token in text:
            frequency[token] += 1
    processed_corpus = [[token for token in text if
    frequency[token] > 1] for text in texts]
    dictionary = corpora.Dictionary(processed_corpus)
    bow_corpus = [dictionary.doc2bow(text) for text in
    processed_corpus]
    tfidf_score=[]
    for l in range(len(texts)):
        tfidf_score.append(tfidf[dictionary.doc2bow(texts[l])])

    lsi_model = models.LsiModel(tfidf_score, id2word=dictionary, n
    lsi_topics.append(lsi_model.print_topics(3))
```

4.2 Decade wise analysis

4.2.1 1900-1910

Topic:
'-0.348*"states" + 0.307*"men" + 0.265*"must" + -0.135*"will" + '
'0.134*"power" +
-0.130*"government" + -0.123*"secretary" + -0.121*"new" + '
'0.111*"come" +
-0.105*"settlement"

Description: This topic is simple in terms of events. No major events occured during this time. And the government in power is electing some secretary and other people.

4.2.2 1910-1920

Topic:
'-0.377*"-" + 0.258*"interstate" + -0.231*"peace" + 0.209*"bill" + '
'0.166*"industrial" + 0.150*"administrative" + 0.148*"public" + '
'-0.145*"war" + 0.133*"may" + -0.130*"world"'

Description: This is the time when the world war took place. From the topics we can see that it is clearly indicated in the keywords of the topic. The decade speeches are about

achieving peace after the war and the peace between the states. After observing the decade speech we can see that this is the major prominent event that occurred in that decade.

4.2.3 1920-1930

Topic:
-0.242*"government" + -0.175*"debt," + -0.173*"fiscal" + 0.167*"must" + ’
’-0.158*"goods" + -0.155*"order" + -0.131*"floating" + -0.128*"faith" +
’ ’-0.126*"right" + -0.117*"part"

Description: From the keywords of the topics we can see that the words like debt and fiscal deficit are used. From this and after observing the decade speech we can see that it was the time of great depression which is clearly evident from the keywords that are highlighted in here.

4.2.4 1930-1940

Topic:
0.444*"-" + 0.193*"upon" + 0.181*"congress" + -0.173*"national" + ’
’0.164*"construction" + 0.151*"temporary" + 0.132*"cent" + ’
’0.114*"unemployment" + 0.104*"employment" + -0.103*"70"

Description: From the keywords in the topic we can see that the terms like employment and unemployment is used. We can infer from this that the goverment is trying to construct industries and create employment oppurtunities thus reducing unemployment.

4.2.5 1940-1950

Topic:
.241*"make" + -0.200*"defense" + 0.199*"will" + 0.184*"united" + ’
’-0.165*"every" + 0.146*"federal" + -0.136*"resources" + -0.130*"social" + ’
’-0.119*"immediate" + -0.114*"world.

Description: This topic refers to the time of the world war 2 from the keywords like defence and resources we can make it out.

4.2.6 1950-1960

Topics:
-0.240*"must" + 0.198*"world" + 0.176*"vital" + 0.163*"field" + ’
’0.159*"congress" + 0.154*"atomic" + -0.150*"defense" + -0.149*"part" + ’
’0.146*"free" + 0.141*"security"

Description: Recovering from the great war.

4.2.7 1960-1970

0.360*"will" + -0.185*"development" + -0.176*"world" + -0.173*"today" + ’
’-0.169*"eight" + -0.147*"united" + -0.139*"federal" + -0.132*"economic" + ’

’-0.130*"since" + 0.128*"recommend

Description: Development after the war struck the countries.

4.2.8 1970-1980

’0.459*"can" + 0.450*"will" + 0.297*"government" + 0.176*"federal" + ’
’0.148*"great" + 0.120*"americans" + 0.118*"new" + 0.104*"nations" + ’
’0.102*"many" + 0.093*"programs"

Description: The topic is about the new relations developed by the US with other countries.

4.2.9 1990-2000

’0.327*"american" + 0.267*"people" + 0.172*"year" + 0.162*"new" + ’
’0.162*"last" + 0.158*"years" + 0.152*"let" + 0.151*"government" + ’
’0.142*"americaś" + 0.137*"america"

Description: The development for country took place.

4.2.10 2000-2010

’-0.378*"saddam" + 0.194*"want" + -0.138*"seniors" + 0.130*"national" + ’
’-0.129*"weapons" + -0.121*"good" + -0.116*"seek" + 0.111*"thank" + ’
’-0.1 07*"urge" + 0.107*"states"’

Description: The killing of saddam hussain.

4.2.11 2010-2012

0.454*"–" + 0.205*"care" + -0.181*"-" + 0.172*"will" + -0.161*"one" + ’
’-0.159*"right" + -0.137*"american" + -0.120*"like" + -0.114*"million" + ’
’-0.113*"gas
Description: Deals regarding selling of oil and gas.

5 Topic modeling on AP dataset

Now we will try to model LSI and LDA topic modeling on AP dataset.

5.1 LDA topic modeling number of optimal topics

We first try to find the number of optimal topics for the LDA topic modeling performed on the AP wired dataset. From the figure 10 we can see that the optimal number of topics in this turned out to be 7.

Topic 1

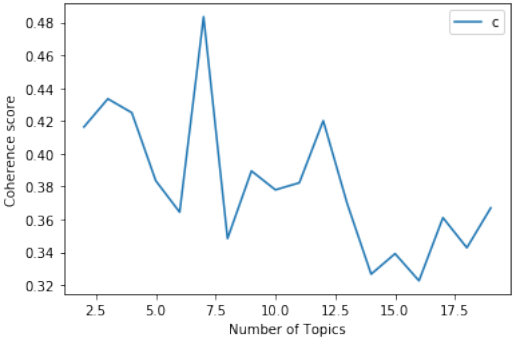

Figure 10: Optimal number of topics in LDA modeling

’0.015*"said" + 0.007*"said." + 0.006*"new" + 0.004*" " +
0.004*"president" + 0.003*"will" + 0.003*"one" + 0.003*"“i" +
0.003*"“the" + 0.003*"bush"’

Description: From the words of the topic we can see that they are related to the presidential elections. It is related to some meeting held there.

Topic 2

’0.017*"said" + 0.007*"said." + 0.004*"will" + 0.004*"two" +
0.003*"police" + 0.003*"one" + 0.003*"also" + 0.003*"last" +
0.003*"percent" + 0.002*"million"’),

Description: We cannot make anything from the topic. The topic doesn’t capture any true meaning at all. By looking at the topic we can see that the original document is about police activties. Since this is a news document we can say so.

Topic 3

’0.014*"said" + 0.008*"said." + 0.003*"will" + 0.003*"also" + 0.003*"two" + 0.003*"one" + 0.003*"new" + 0.003*"million" + 0.003*" " + 0.002*"people"

Description: From the topic we can see that the words like million , people aare used. It could be that the government is trying to form a reform for the people that will benefit millions of them.

Topic 4

’0.017*"said" + 0.005*"said." + 0.005*"new" + 0.004*"u.s." +
0.004*"will" + 0.003*"government" + 0.003*"also" + 0.003*"two"
+ 0.003*"president" + 0.002*"united"’

Description: The topic is little vague. The LSI model gives clear topic results but not the

lda results. LDA results give more generalised results of all the topics.

Topic 5

'0.018*"said" + 0.010*"said." + 0.005*"will" + 0.004*"one" +
0.003*"million" + 0.003*"police" + 0.003*"people" +
0.003*"new" + 0.003*"two" + 0.003*"also"'

Description:This topic is related to something regarding the police.

Topic 6

'0.020*"majority" + 0.005*"will" + 0.005*"said." + 0.004*"government" + 0.004*"new"
+ 0.003*"also" + 0.003*"president" + 0.003*"last" + 0.003*"u.s." + 0.003*"two"

Description: The topic tells about the new president election that is going to be held.

Topic 7

'0.014*"said" + 0.007*"will" + 0.005*"soviet" + 0.004*"u.s."
+ 0.004*"enough." + 0.003*"president" +
0.003*"new" + 0.003*"also" + 0.003*"percent" + 0.003*"united"

Description: From the topic keywords we can see that the keywords are related to the soviet and USA. It could be about the conference held between those countries. From the text we can see that there was some little tension between both the countries but the model failed to capture the idea accurately.

Topic 8

'0.012*"said" + 0.006*"will" + 0.005*"new" + 0.004*"said." + 0.003*"million" + 0.003*"one"
+ 0.003*"first" + 0.003*"last" + 0.003* + 0.003*"cents"

Description:This topic tells about the millions of dollars spent on various reforms of the government.

Topic 9

'0.012*"said" + 0.009*"said." + 0.005*"will" + 0.004*"also" + 0.004*"police" + 0.004*"two"
+ 0.003*"new" + 0.003*"people" + 0.003*"one" + 0.003*"u.s."'

Description: This topic is about the police in the country.

Topic 10

'0.010*"said" + 0.007*"percent" + 0.007* + 0.007*"said." + 0.004*"will" + 0.003*"new"
+ 0.003*"last" + 0.003*"u.s." + 0.003*"united" + 0.003*"people"'

Description: These last two topics are very similar to each other. These topics do not hold any clear meaning in the document.

5.2 Comparison of two datasets

When comparing we can see that both the datasets are have data in the same format. When we apply LSI on the state of the union dataset we can see that the topic formed were coherent and that the topic that are identified by it are very meaningful. When we apply LDA topic modeling on the dataset we can see that the we were not able to capture the true meaning very accurately unlike LSI model. Similarly we see this when we apply LDA topic modeling on the ap wired stories we are not able to capture the true meanings that we want to. We can say that the topic modeling worked better on the state of the union dataset than the ap wired stories. This is because the I found that the meaning captured by the LSI and LDA model on that dataset made more sense and many information can be retrieved from it.