Assignment-4 Report
Topic Modeling

Samreet Singh Dhami
2018CSB1120
IIT Ropar
Rupnagar

Abstract

The report is designed to develop a feel for the way topic modelling works, the connection to the human meanings of documents, and common ways of handling a time dimension. Analysis includes the State of the Union speeches corpus, and report on how the subjects have shifted over time in relation to historical events. Topic modeling has been done using two of the common methods, Latent Semantic Analysis(LSI) and Latent Dirichlet Allocation(LDA). Visualization on data modeling has been done using different techniques and observations are made on the results. At last, a new document set, collection of AP wire stories was introduced and results of two documents were compared.

1 Pre-processing and generating tf-idf score

Topic Modeling is a technique to extract the hidden topics from large volumes of text.LSI and LDA ser two main algorithm in the gensim package of python to extract good quality topics. To achieve this, pre-processing of the content is necessary. Tokenization of the content and removing the stopwords is also very necessary. Once all these important steps are done, gensim is able to generate tf-idf scores.

2 Latent Semantic Indexing(LSI) on the data

Now its time to apply Latent Semantic Indexing (LSI) to the tf-idf vectors. The major challenge arises is to determine number of topics. Once the number of topics are determined, we can import lsi model and apply it to print topics. Topics are printed in well-defined way. For example Topic 0 is a represented as'0.072*"upon" + 0.062*"tonight" + 0.059*"program" + 0.058*"mexico" + 0.057*"economic" + 0.052*"treaty" + 0.051*"subject" + 0.050*"help" + 0.049*"budget" + 0.049*"americans"')

It means the top 10 keywords that contribute to this topic are: 'upon', 'tonight', 'program'.. and so on and the weight of 'tonight' on topic 0 is 0.062. Let us consider the first ten topics in the lsi model for further analysis.

1. The first topic has keywords program, Mexico, Economic, Treaty. This topic is dominant in initial years. If we deeply read the content, in the first few years after 1790 upto 1816 this topic does not hold good as in these years mostly issues related to Europe and tribes from India and other parts of the world are discussed. But in the years

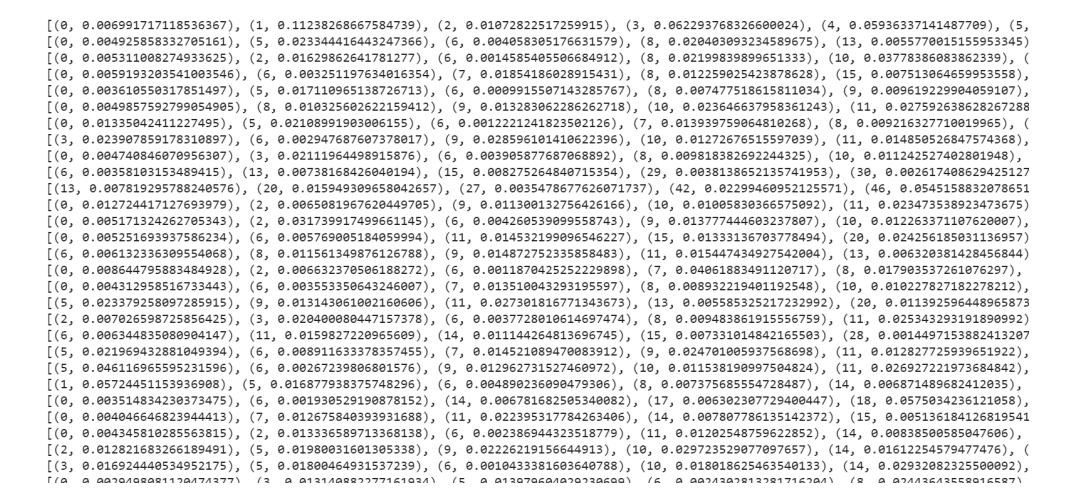


Figure 1: tf-idf scores

[(0,
'0.072*"upon" + 0.062*"tonight" + 0.059*"program" + 0.058*"mexico" + 0.057*"economic" + 0.052*"treaty" + 0.051*"subject" + 0.050*"help" + 0.049*"budget" + 0.049*"americans"'),
(1,
'-0.192*"tonight" + -0.129*"jobs" + -0.125*"americans" + -0.120*"help" + -0.118*"budget" + -0.117*"program" + -0.116*"programs" + -0.105*"billion" + -0.101*"economic" + -0.101*"today"'),
(2,
'0.205*"tonight" + -0.101*"program" + 0.100*"iraq" + 0.097*"jobs" + -0.096*"economic" + 0.089*"americans" + 0.084*"terrorists" + -0.081*"interstate" + -0.080*"veterans" + -0.079*"cent"'),
(3,
'-0.136*"program" + 0.131*"tonight" + 0.116*"silver" + -0.102*"economic" + -0.097*"soviet" + -0.097*"communist" + 0.096*"iraq" + 0.093*"cent" + 0.089*"gold" + -0.083*"programs"'),
(4,
'0.135*"silver" + -0.133*"iraq" + -0.128*"interstate" + 0.128*"gold" + -0.113*"terrorists" + 0.108*"soviet" + 0.102*"programs" + 0.100*"mexico" + 0.094*"notes" + 0.085*"billion"'),
(5,
'0.275*"iraq" + 0.231*"terrorists" + 0.168*"iraqi" + 0.150*"al" + 0.139*"terror" + 0.132*"terrorist" + 0.119*"afghanistan" + 0.102*"saddam" + 0.100*"soviet" + 0.094*"iraqis"'),
(6,
'-0.157*"mexico" + -0.156*"texas" + -0.096*"paper" + -0.085*"kansas" + 0.085*"spain" + -0.084*"constitution" + -0.080*"specie" + -0.077*"slavery" + -0.068*"bank" + 0.068*"consular"'),
(7,
'-0.152*"banks" + -0.131*"iraq" + -0.112*"veterans" + 0.110*"soviet" + 0.109*"vietnam" + -0.094*"terrorists" + 0.089*"japanese" + -0.085*"bank" + -0.081*"notes" + 0.077*"fighting"'),
(8,
'-0.191*"silver" + 0.169*"mexico" + -0.157*"gold" + -0.137*"notes" + 0.125*"texas" + -0.111*"coinage" + -0.103*"gentlemen" + -0.099*"circulation" + -0.093*"currency" + -0.091*"militia"'),
(9,
'-0.187*"vietnam" + -0.115*"gentlemen" + -0.109*"soviet" + 0.095*"enemy" + -0.087*"oil" + 0.086*"spain" + -0.082*"tonight" + 0.080*"japanese" + -0.078*"programs" + 0.077*"hitler"'),
(10,
'0.175*"vietnam" + 0.158*"enemy" + -0.107*"gentlemen" + 0.104*"militia" + 0.100*"whilst" + 0.082*"savages" + -0.081*"democracy" + -0.076*"chambers" + 0.073*"british" + -0.070*"spain"'),
(11,
'-0.240*"gentlemen" + 0.134*"spain" + 0.106*"colonies" + -0.103*"kansas" + -0.093*"slavery" + 0.083*"silver" + 0.082*"gold" + 0.078*"notes" + 0.073*"presumed" + 0.073*"florida"'),
(12,
'0.205*"mexico" + 0.197*"texas" + -0.122*"slavery" + 0.120*"mexican" + -0.104*"kansas" + -0.099*"rebellion" + -0.098*"emancipation" + 0.093*"soviet" + 0.092*"enemy" + 0.083*"annexation"'),
(13,
'0.327*"vietnam" + 0.225*"tonight" + -0.120*"jobs" + 0.077*"democracy" + -0.075*"businesses" + -0.072*"energy" + -0.068*"programs" + -0.068*"oil" + -0.066*"corporations" + -0.062*"forest"'
(14,
'-0.267*"soviet" + 0.211*"vietnam" + -0.167*"oil" + 0.149*"texas" + 0.136*"mexico" + 0.105*"billion" + -0.097*"jobs" + 0.082*"housing" + 0.081*"program" + -0.077*"afghanistan"')]

Figure 2: LSI-Topics

following 1817, Mexico and the Gulf was highly used due to tension which ultimately lead to treaty between the two nation.

2. The second topic has keywords Jobs, budget,economic, program etc. clearly depicts the economic reforms, programs to increase production and Job in the state. This topic was dominant around 1850s.

3. The second topic and third topic does not differ too much.

4. The fourth topic has varied keywords, therefore, cannot be annotated. It does not give details about any significant topic except economy like cents, gold.farms etc.

5. Fifth topic has keywords has interstate, silver, iraq, terrorist, soviet etc. As we know, United States of America has almost 50 states, In beginning of its structural development, there were many disputes between states and many reforms and rules were made to make peace and for the betterment of all the states.

6. Sixth topic had keywords iraq,al,terrorist,saddam etc. This obviously is the major era of the United States history. This is around 2003 when US invaded iraq. It had also involvement of Afghanistan.

7. Seventh topic comprises of keywords texas, kansas, constitution ,spain ,slavery. This topic is very much generalised which talk about basic achievements and progress states and the cities have made. Constitution word is used frequently because any amendment in the constitution has a say in union address.

8. Topic Number 8 has keywords bank, vietnam ,veterans ,banks etc. This is basically around 1974 when the vietnam war happened. If we go through csv file vietnam came into addresses in the year 1966 and continued upto 1990s. In one of the year whole decription of the war is given.

9. Ninth topic includes notes, coinage, gold,silver,currency,notes. These words are very much frequent because it comprises of the basic financial system of the country.

10. Tenth topic has keyword enemy,japanese ,oil etc. This topic does not get focussed to any specific point but gives the hint of tension between Japan and USA before world war 2.

3 LDA topic modeling

Above exercise is repeated using Latent Dirichlet Allocation on the data. Overall it is found that LDA is better for visualization than LSI. Basically it covers more important topics than lsi model has covered. All the observations are mentioned in the following points. Let us consider first ten topics in the LDA model also.

1. First topic has keywords iraq,enemy,terrorists etc. This is in resemblance with the sixth topic of lsi model.

```
[(0,
 '0.000*"iraq" + 0.000*"tonight" + 0.000*"terrorists" + 0.000*"jobs" + 0.000*"americans" + 0.000*"program" + 0.000*"america" + 0.000*"workers" + 0.000*"terror" + 0.000*"help"'),
(1,
 '0.000*"tonight" + 0.000*"jobs" + 0.000*"americans" + 0.000*"inflation" + 0.000*"help" + 0.000*"child" + 0.000*"college" + 0.000*"economic" + 0.000*"children" + 0.000*"programs"'),
(2,
 '0.000*"mexico" + 0.000*"award" + 0.000*"treaty" + 0.000*"majesty" + 0.000*"mexican" + 0.000*"texas" + 0.000*"june" + 0.000*"upon" + 0.000*"freedmen" + 0.000*"cent"'),
(3,
 '0.000*"soviet" + 0.000*"economic" + 0.000*"jobs" + 0.000*"texas" + 0.000*"today" + 0.000*"upon" + 0.000*"subject" + 0.000*"americans" + 0.000*"cable" + 0.000*"programs"'),
(4,
 '0.000*"banks" + 0.000*"tonight" + 0.000*"subject" + 0.000*"silver" + 0.000*"upon" + 0.000*"treasury" + 0.000*"challenge" + 0.000*"militia" + 0.000*"americans" + 0.000*"bank"'),
(5,
 '0.000*"vietnam" + 0.000*"spain" + 0.000*"billion" + 0.000*"article" + 0.000*"programs" + 0.000*"gentlemen" + 0.000*"vessels" + 0.000*"percent" + 0.000*"program" + 0.000*"economic"'),
(6,
 '0.000*"soviet" + 0.000*"oil" + 0.000*"program" + 0.000*"budget" + 0.000*"cent" + 0.000*"afghanistan" + 0.000*"help" + 0.000*"emancipation" + 0.000*"tons" + 0.000*"programs"'),
(7,
 '0.000*"gentlemen" + 0.000*"budget" + 0.000*"upon" + 0.000*"today" + 0.000*"program" + 0.000*"help" + 0.000*"let" + 0.000*"tonight" + 0.000*"economic" + 0.000*"oil"'),
(8,
 '0.000*"economic" + 0.000*"kansas" + 0.000*"banks" + 0.000*"democracy" + 0.000*"cent" + 0.000*"veterans" + 0.000*"upon" + 0.000*"atomic" + 0.000*"communist" + 0.000*"per"'),
(9,
 '0.000*"program" + 0.000*"programs" + 0.000*"billion" + 0.000*"gold" + 0.000*"economic" + 0.000*"tonight" + 0.000*"budget" + 0.000*"help" + 0.000*"federal" + 0.000*"silver"'),
(10,
 '0.000*"tonight" + 0.000*"economic" + 0.000*"americans" + 0.000*"budget" + 0.000*"jobs" + 0.000*"help" + 0.000*"upon" + 0.000*"recovery" + 0.000*"attentions" + 0.000*"militia"'),
(11,
 '0.000*"tonight" + 0.000*"americans" + 0.000*"democracy" + 0.000*"help" + 0.000*"jobs" + 0.000*"programs" + 0.000*"billion" + 0.000*"program" + 0.000*"cannot" + 0.000*"today"'),
(12,
 '0.001*"tonight" + 0.000*"vietnam" + 0.000*"mexico" + 0.000*"billion" + 0.000*"budget" + 0.000*"help" + 0.000*"gold" + 0.000*"today" + 0.000*"program" + 0.000*"upon"'),
(13,
 '0.000*"spain" + 0.000*"gold" + 0.000*"kansas" + 0.000*"treaty" + 0.000*"currency" + 0.000*"paper" + 0.000*"upon" + 0.000*"constitution" + 0.000*"silver" + 0.000*"minister"'),
(14,
 '0.000*"economic" + 0.000*"hitler" + 0.000*"program" + 0.000*"gentlemen" + 0.000*"upon" + 0.000*"texas" + 0.000*"mexico" + 0.000*"interstate" + 0.000*"statute" + 0.000*"industrial"')]
```

Figure 3: LDA-Topics

2. Second topic has keywords Jobs, Americans, inflation, economic ,college. This topic has resemblance with second topic of lsi model but it can be clearly observed that words are well interconnected and covers many of the topics in which reforms have been made.

3. Third topic consist of mexico, treaty ,award, mexican etc. This directly gives reference to the Treaty of Guadalupe Hidalgo, signed on February 2, 1848, ended the Mexican-American War in favor of the United States.

4. Fourth topic has varied words which do not depict a major event in the history.

5. The fifth topic has keywords bank, treasure,bank ,militia etc. This topic hints about the financial system.

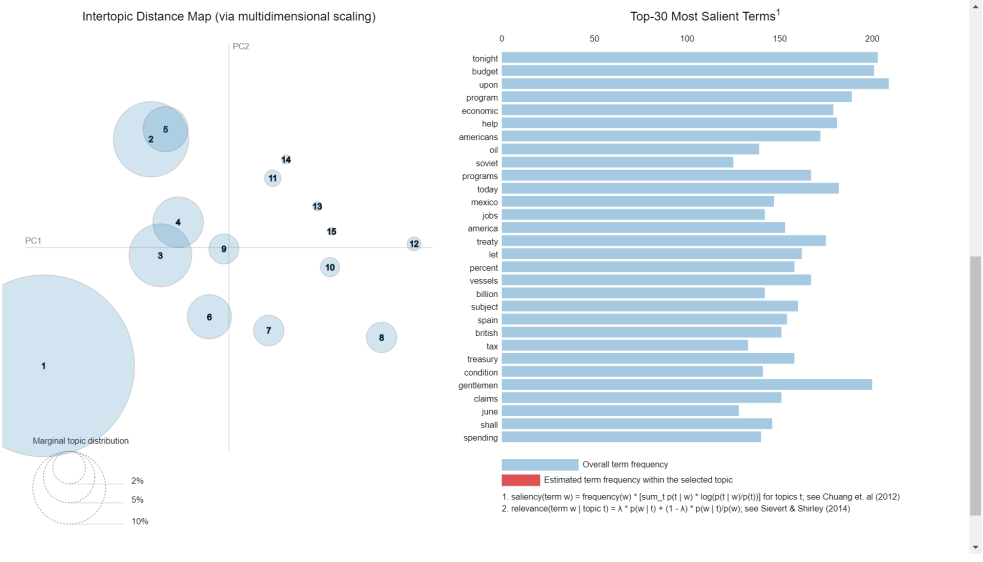
6. The sixth topic has keywords vietnam, spain, billions,articles ,programs. This also different words which means the speech did not focus on a particular event.

7. Seventh topic consist of keywords soviet ,oil ,budget, cent ,aghanistan. Basically soviet union was formed in 1922 and it is one of the major events in world history and therefore it has much frequency in the Union addresses.

8. Topic number 8 has common keywords. This topic is dominant in those years in which focuses on general topics like in initial years.

9. Ninth topic of lda model resembles with the seventh topic. It has keywords economic, banks ,democracy,cent etc. Kansas city is an important city in the USA due to its aviation manufacturing and revenue generation, Moreover, there is a high amount of mention of Kansas Pacific in 1900s due to some financial debts in the area.

10. The 14th topic has keywords hitler,economic,texas,mexico. This is the time of World War 1 when Hitler and its Nazi germans ruled in some parts of Europe. This topic is not mentioned with significant amount of weight in lsi. Therefore, lda model proves to be better in the data of state of the union.csv



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	4.0	0.7355	banks, tonight, subject, silver, upon, treasur...	[state, union, address, george, washington, de...
1	1	5.0	0.5350	vietnam, spain, billion, article, programs, ge...	[state, union, address, george, washington, oc...
2	2	7.0	0.5582	gentlemen, budget, upon, today, program, help,...	[state, union, address, george, washington, no...
3	3	14.0	0.4922	economic, hitler, program, gentlemen, upon, te...	[state, union, address, george, washington, de...
4	4	7.0	0.5091	gentlemen, budget, upon, today, program, help,...	[state, union, address, george, washington, no...
5	5	7.0	0.5747	gentlemen, budget, upon, today, program, help,...	[state, union, address, george, washington, de...
6	6	5.0	0.5832	vietnam, spain, billion, article, programs, ge...	[state, union, address, george, washington, de...
7	7	5.0	0.5561	vietnam, spain, billion, article, programs, ge...	[state, union, address, john, adams, november,...
8	8	7.0	0.9991	gentlemen, budget, upon, today, program, help,...	[state, union, address, john, adams, december,...
9	9	7.0	0.9987	gentlemen, budget, upon, today, program, help,...	[state, union, address, john, adams, december,...

Figure 4: Some important visualizations

4 Change in Speeches overtime

There are many ways to summarize the changes overtime. this. Possibilities include: visualizations, grouping speeches by decade after topic modeling, and grouping speeches by decade before topic modeling.LDA model has been taken as a base for the algorithm. pyLDAvis ,an inbuilt pyton package, which is commonly used in topic modeling visualization proves to be very helpful in this case. This is the best tool according to many of the data scientists. Left-hand side contains bubbles which represent topics. A good topic will have big, non-overlapping bubble scattered throughout the chart. If we move the mouse pointer over one of the bubbles, the words order will change accordingly and bars will be updated and some part of it will be red showing percentage of word in the content . These words are the salient keywords that form the selected topic. The corresponding visualization is shown in **Figure:4**. Using pyLDAvis notebook and topics of LDA following observations can be recognised-

- US got independent and 1770s and in years following it, speeches include more about reforms and programs etc.

- On February 2, 1848, the Treaty of Guadalupe Hidalgo was signed between Mexico and USA which brought to a close the Mexican-American War in 1850s when Mr. James Polk was the president. It can be easily observed mexico,mexicans and treaty were important words in 1850s.
- In 1900s President Theodore Roosevelt had many progressive goals in terms of economy,financial and political power. Therefore, the topics include interstate, states, bank ,cent,notes etc.
- In 1910s, World War-1 raged in Europe, but the President avoided any involvement at first but then were part of allies. Topic which includes war,enemy,hitler, europe were dominant in this era.
- In 1920s came the year of great depression which lead to stock market crash. In the speeches of these years, topic mainly consisted of budget, economic,reforms , jobs, programs,gold and words related to finance.
- In the era between great depression and world-war-II, many reforms were made in fields related to agriculture, industry, finance etc. by the president Franklin D. Roosevelt.
- 1940s was the era of world-war-II. In these years speeches had too much content related to japanese, enemy,assassination,killing,war, atomic("Atomic bomb"). It was clear from the content, it was all war in these times.
- In the year 1950s and 1960s, topic which had keywords militia,tactics,vietnam,war were dominant because of Vietnam war which also lead to civil rights movement.
- The watergate scandal cannot be determined in lda model. Topics related to it cannot get recognised through this algorithm.
- In 2000s, there were two major events one was 9/11 attack on the world trade centre. Topics related to terror, killing,terrorists were common in these years.
- In 2003, US invaded iraq and therefore there are many topics which had iraq, iraqi ,saddam with significant amount of weight. The President at that time was George W. Bush. Therefore, many historical events in US can easily be recognised through this model.

5 Different Document Set

LDA is applied on different document set which is a collection of AP wire stories. Same process is repeated as for the state of the union dataset and results of both are compared. If we go through the csv file, we can easily observe that they are news articles.Now the news has widespread topics, therefore,hundred number of topics were chosen. Lets here also consider first 10 topics of lda model.

1. Now, the first topic has main keywords sales, taxes,billions,india, delhi. Sales and taxes are obviously will have more weights they are a core part of any nation. India has come into consideration as after going through the data, it is clearly visible that there is too much mention of indians slavery and indians residing in the nation. Moreover war between India and Pakistan is discussed.

```
[(41,
 '0.014*"sales" + 0.009*"tax" + 0.009*"agreement" + 0.007*"united" + 0.006*"abuse" + 0.006*"billion" + 0.005*"delhi" + 0.005*"news" + 0.005*"india" + 0.005*"joints"'),
(82,
 '0.014*"primary" + 0.010*"cars" + 0.008*"plant" + 0.008*"rocks" + 0.007*"domestic" + 0.007*"jefferson" + 0.007*"williams" + 0.006*"techniques" + 0.005*"bush" + 0.005*"pont"'),
(38,
 '0.013*"sasser" + 0.013*"climbed" + 0.011*"barr" + 0.010*"cosby" + 0.008*"november" + 0.007*"adopt" + 0.006*"strains" + 0.006*"afghanistan" + 0.006*"welcome" + 0.006*"realistic"'),
(63,
 '0.029*"presley" + 0.017*"disc" + 0.016*"happens" + 0.013*"elvis" + 0.011*"le" + 0.010*"french" + 0.007*"love" + 0.007*"extreme" + 0.007*"article" + 0.006*"sank"'),
(33,
 '0.013*"interstate" + 0.009*"epa" + 0.008*"miller" + 0.007*"hong" + 0.007*"kong" + 0.007*"tariffs" + 0.007*"chinese" + 0.006*"connection" + 0.006*"arrested" + 0.006*"earned"'),
(49,
 '0.009*"economy" + 0.008*"study" + 0.006*"hoffman" + 0.006*"percent" + 0.005*"houston" + 0.005*"groups" + 0.005*"agencies" + 0.005*"part" + 0.005*"toronto" + 0.005*"usually"'),
(94,
 '0.011*"energy" + 0.007*"oh" + 0.007*"dress" + 0.006*"women" + 0.006*"committee" + 0.006*"educated" + 0.005*"ourselves" + 0.005*"salinas" + 0.005*"chernobyl" + 0.005*"she"'),
(10,
 '0.016*"thompson" + 0.007*"wilson" + 0.007*"delegates" + 0.006*"now" + 0.006*"ala" + 0.006*"air" + 0.005*"iran" + 0.005*"kashmir" + 0.005*"speed" + 0.004*"literary"'),
(90,
 '0.007*"below" + 0.006*"she" + 0.006*"deaths" + 0.006*"tickets" + 0.005*"texas" + 0.005*"ms" + 0.005*"there" + 0.004*"state" + 0.004*"night" + 0.004*"mexico"'),
(83,
 '0.018*"video" + 0.015*"speaker" + 0.012*"indictment" + 0.010*"ohio" + 0.010*"cocaine" + 0.008*"sing" + 0.007*"smuggler" + 0.007*"ethics" + 0.007*"committee" + 0.006*"wage"'),
(14,
 '0.006*"panama" + 0.006*"lafayette" + 0.006*"late" + 0.006*"gold" + 0.006*"down" + 0.006*"together" + 0.006*"london" + 0.005*"noriega" + 0.005*"technology" + 0.005*"square"'),
(69,
 '0.047*"ec" + 0.018*"bendjedid" + 0.015*"hamilton" + 0.012*"buses" + 0.010*"sweden" + 0.008*"blast" + 0.007*"deductions" + 0.007*"nelson" + 0.006*"legislature" + 0.005*"drivers"'),
(48,
 '0.012*"awards" + 0.011*"kohl" + 0.008*"flood" + 0.008*"deer" + 0.007*"retire" + 0.007*"white" + 0.007*"chancellor" + 0.006*"waters" + 0.006*"intervention" + 0.005*"random"'),
(39,
 '0.017*"trash" + 0.014*"mecham" + 0.013*"cent" + 0.013*"roberts" + 0.010*"bushel" + 0.008*"lower" + 0.007*"victims" + 0.007*"higher" + 0.006*"name" + 0.006*"futures"'),
(98,
 '0.014*"navy" + 0.012*"ships" + 0.011*"sea" + 0.009*"computer" + 0.009*"detained" + 0.007*"sony" + 0.007*"expenses" + 0.006*"tools" + 0.006*"rehabilitation" + 0.006*"straight"'),
(6,
 '0.012*"shearson" + 0.008*"draws" + 0.007*"sweden" + 0.007*"starring" + 0.007*"allen" + 0.006*"winners" + 0.006*"producer" + 0.006*"undecided" + 0.005*"jonathan" + 0.005*"gun"'),
```

Figure 5: Topics of the second document

2. The second topic has keywords plant, cars and some names. After going through the data, there many news which involve accidents, car chase of police and train accidents. These events are described thoroughly and therefore, has a high weight. Names are basically of presidents of USA. News always mention the name of the President when major event happens.

3. The third topic also consists of a name sasser. He was United States Senator. But he other words are common english words which are commonly used. Hence, this topic was not useful.

4. Fourth topic consists of keywords presley, disc, happens,elvis, french, love. This topic has varied words but some can be connected like Elvis Presley is an American singer whose life is too many discussed. The words like love and french are frequently used.

5. Fifth topic can be annotated as interstate, miller, hong kong, tariffs, chinese etc. Miller is again a famous Personality. Hong Kong and China have been in news for many of the reasons.

6. Sixth topic can annotated as economy study houston groups agencies. This topic mainly covers the news which are related economy and its future.

7. Seventh Topic has keywords women,dress,energy, chernobyl ,she. Whenever topic related to women dresses and education are discussed this topic is dominant. Moreover, it has word chernobyl which is related to the disaster in chernobyl and its consequences around that region.

8. Eighth topic is a bit wider one as it comprises of Thomson, Wilson ,delegates, kashmir, iran. First two are again names. Iran and US ahs a long history which are frequently described in the news. Basically, Iranian fighter planes, killings, islamic revolution were the main reasons due to which a few news are based on Iran.

9. Ninth topic again does not depict any significant news or event. It consists of common english words.

10. Tenth Topic has important keywords like video, smuggler, Ohio, cocaine, indictment, speaker. This topic is clearly about drug smuggling and serious punishment which was given for the crime. Videos were part of evidence in only a few of the cases. The speaker word is mainly due to speaker of the house which is an important person in the country.

Now, to compare the two lda model. Its obvious one news article focuses on a single topic while union speech covers all important reforms and events occurred. Therefore, in the second data, number of topics has to be increased and if we look at the dominant topics of each row, easily we can two or more words that can be the topic. But in union addresses some topics, just does not represent the content of the speech. Some topics even had no clear concept. Increasing the number of topics also did not help because of varied content in the speech. Overall, due to reasons mentioned, **output of AP wire collection document proves to be the better one.**