Assignment 4: Topic Modeling

Shubham Pradhan
2017EEB1168

Indian Institute of Technology
Ropar, Punjab

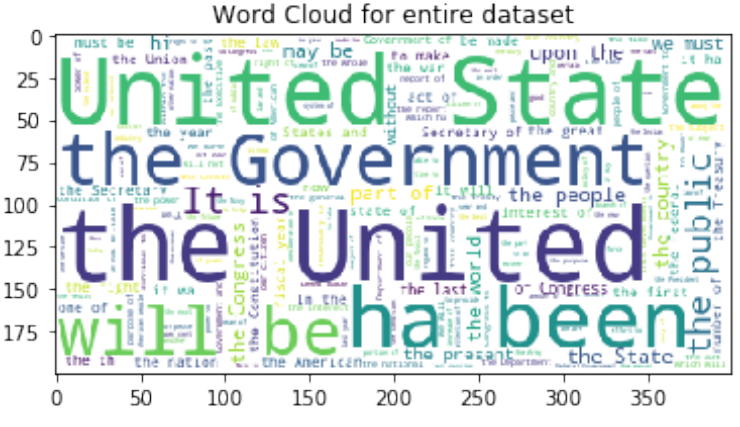
1 Introduction

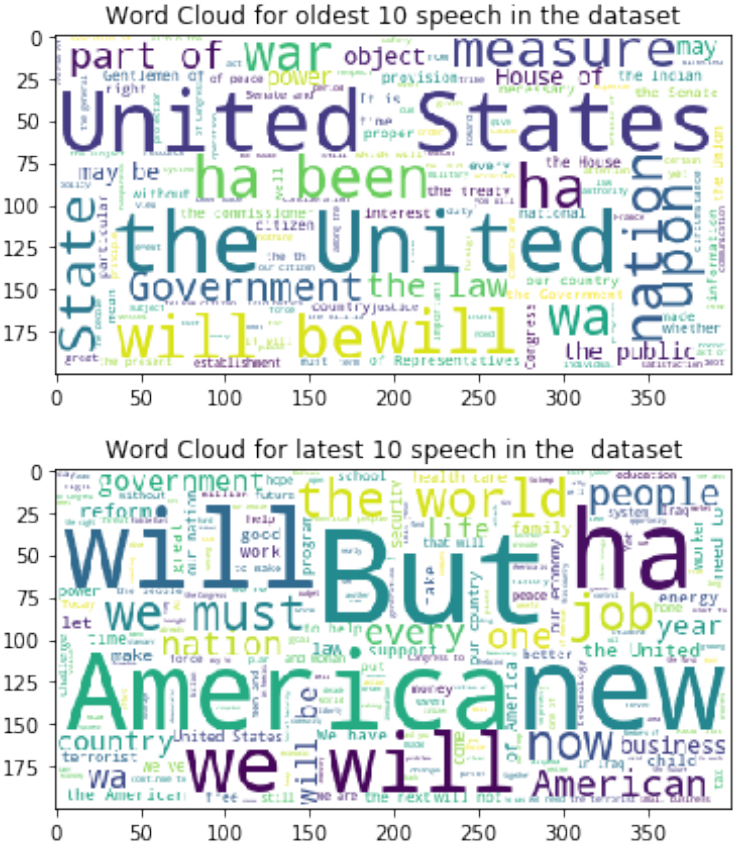
Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. This is known as 'unsupervised' machine learning because it doesn't require a predefined list of tags or training data that's been previously classified by humans.

Topic modeling involves counting words and grouping similar word patterns to infer topics within unstructured data. By detecting patterns such as word frequency and distance between words, a topic model clusters feedback that is similar, and words and expressions that appear most often. With this information, you can quickly deduce what each set of texts are talking about.

2 DataSet

The State of the Union dataset contaions the the State of the Union address from year 1790 to 2012. The top words in the data set are shown below.





3 Task 1: Feed the data into the gensim Python package

Gensim is billed as a Natural Language Processing package that does 'Topic Modeling for Humans'. But its practically much more than that. I loaded the documents into Python and fed them into the gensim package to generate tf-idf weighted document vectors. I had to go through the file twice: once to generate the dictionary, and then again to convert each document to what gensim calls the bag-of-words representation, which is un-normalized term frequency.

In order to work on text documents, Gensim requires the words (aka tokens) be converted to unique ids. In order to achieve that, Gensim create a Dictionary object that maps each word to a unique id. The dictionary object is typically used to create a 'bag of words' Corpus. It is this Dictionary and the bag-of-words (Corpus) that are used as inputs to topic modeling and other models that Gensim specializes in. There is implicitly another step here, which is to tokenize the document text into individual word features.

Before creating the dictionary we had to clean the text dataset. The steps involved removing all the unwanted characters such as punctuation, numbers, special characters and unwanted spacing. Next step is the lemmetization, Lemmatization is the process of grouping together

the different inflected forms of a word so they can be analysed as a single item. At last we had to remove all the stopwords from the dataset. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We used the stopwords given in the assignment document as well as those present in the NLTK library. For the given dataset the dictionary contains 14560 unique tokens.

```
Dictionary(14560 unique tokens: ['able', 'abridge', 'abroad', 'abundant', 'abundantly']...)
```

Some of the words from the dictionary with their unique IDs are shown below.

```
'aana': 10947,
'aaron': 3591,
'abandon': 2542,
'abandoned': 1958,
'abandoning': 5084,
'abandonment': 5515,
'abate': 10502,
'abated': 7671,
'abatement': 942,
'abating': 11732,
'abdicated': 12882,
'abdicating': 11892,
'abdication': 6330,
'abet': 7078,
'abettor': 3485,
'abeyance': 9365,
'abhorrence': 7079,
```

The next important object we need to create is the Corpus (a Bag of Words). That is, it is a corpus object that contains the word id and its frequency in each document. We can think of it as gensim’s equivalent of a Document-Term matrix. Once we have the updated dictionary, all we need to do to create a bag of words corpus is to pass the tokenized list of words to the Dictionary.doc2bow().

```
[[(0, 1),
  (1, 1),
  (2, 1),
  (3, 2),
  (4, 1),
  (5, 1),
  (6, 2),
  (7, 2),
  (8, 1),
  (9, 2),
  (10, 1),
  (11, 1),
  (12, 1),
  (13, 1),
  (14, 1),
  (15, 1),
  (16, 1),
  (17, 1),
```

The (0, 1) in line 1 means, the word with id=0 appears once in the 1st document. Likewise, the (6, 2) in the second list item means the word with id 6 appears 2 times in the second document. And so on. The order of the words gets lost. Just the word and it's frequency information is retained.

```
[[('able', 1), ('abridge', 1), ('abroad', 1), ('abundant', 2), ('abundantly', 1), ('accordingly', 1), (
dded', 1), ('address', 1), ('administration', 1), ('advantage', 1), ('affair', 1), ('affectionate', 1),
griculture', 1), ('aid', 1), ('aided', 1), ('alacrity', 1), ('allotted', 1), ('allow', 1), ('along', 1)
```

Notice, the order of the words gets lost. Just the word and it's frequency information is retained.

Next step is to generate TFIDF corpus. The Term Frequency – Inverse Document Frequency(TFIDF) is also a bag-of-words model but unlike the regular corpus, TFIDF down weights tokens (words) that appears frequently across documents. Tf-Idf is computed by multiplying a local component like term frequency (TF) with a global component, that is, inverse document frequency (IDF) and optionally normalizing the result to unit length.

```
[['able', 0.01], ['abroad', 0.02], ['act', 0.0], ['add', 0.04],
', 0.0], ['among', 0.02], ['another', 0.0], ['appears', 0.02],
zed', 0.01], ['bandit', 0.08], ['call', 0.01], ['case', 0.03],
```

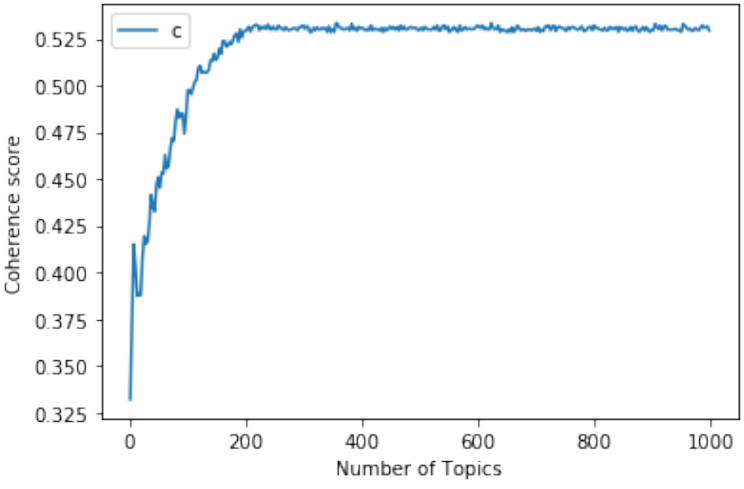
Notice the difference in weights of the words between the original corpus and the tfidf weighted corpus.

4 Task 2: LSI topic modeling

Latent Semantic Indexing, or LSI, is one of the foundational techniques in topic modeling. The core idea is to take a matrix of what we have — documents and terms — and decompose it into a separate document-topic matrix and a topic-term matrix. LSI models typically replace raw counts in the document-term matrix with a tf-idf score.

Topic coherence measure is a realistic measure for identifying the number of topics. Topic Coherence measure is a widely used metric to evaluate topic models. It uses the latent variable models. Each generated topic has a list of words. In topic coherence measure, you will find average/median of pairwise word similarity scores of the words in a topic. The high value of topic coherence score model will be considered as a good topic model.

For LSI model we calculated the coherence measure for number of topics range from 2 to 1000 and the found that after around 250 topics coherence becomes saturated. Thus the optimal number of topics for the given data-set we found out was 250. The graph for the same is sown below. We can see in the plot that the coherence score increases initially with increase in number of topics and then becomes constant after about 250 topics. Thus we can conclude that the 250 is the optimal number of topics.



After running LSI for 250 topics I randomly chose 10 topics from the lot. These 10 topics are explained below:

1. ('0.094*"program" + 0.073*"upon" + 0.066*"tonight" + 0.062*"job" + 0.059*"economic" + 0.058*"mexico" + 0.054*"budget" + 0.054*"help" + 0.052*"americans" + 0.051*"bank"')

   We can conclude from the words present in the topic that this topic refers to economy and the budget of the country (from words like 'economy', 'bank', 'budget') and also to the jobs.

2. ('-0.192*"program" + 0.127*"tonight" + 0.119*"silver" + 0.112*"terrorist" + -0.106*"com + -0.100*"economic" + 0.096*"iraq" + -0.093*"soviet" + 0.089*"cent" + 0.089*"gold"')

   From these topics we can infer that this topic is related to World War II because words like 'soviet', 'communist', 'terrorist' are present.

3. ('-0.208*"silver" + -0.164*"gold" + -0.125*"gentlemen" + -0.119*"coinage" + -0.115*"cir + -0.115*"currency" + 0.109*"spain" + -0.105*"bank" + 0.102*"mexico" + -0.099*"note"'

   Here, the main topics we infer are related to economy and currency since words like 'coinage', 'currency', 'bank' are present.

4. ('-0.202*"vietnam" + -0.173*"gentlemen" + -0.133*"soviet" + 0.120*"nations" + 0.111*"s + 0.089*"hitler" + -0.089*"tonight" + 0.084*"recovery" + 0.081*"job" + 0.080*"japanese"

   these topics also refer to World War II as words like 'Hitler', 'Vietnam', 'Japanese' are present.

5. ('0.194*"program" + 0.194*"tonight" + 0.167*"job" + 0.126*"americans" + 0.125*"help" + 0.124*"budget" + 0.103*"today" + 0.102*"billion" + 0.100*"economic" + 0.097*"percent"')

These topics refer to the economy and the budget of the country due to presence of words like 'budget', 'economy', 'billion'.

6. ('-0.192*"program" + 0.127*"tonight" + 0.119*"silver" + 0.112*"terrorist" + -0.106*"commu + -0.100*"economic" + 0.096*"iraq" + -0.093*"soviet" + 0.089*"cent" + 0.089*"gold"')

   These may refer to the slim relation between Iraq and America and the tension between two countries and the terrorist attacks.

7. ('-0.086*"democracy" + -0.086*"circuit" + 0.078*"saddam" + -0.077*"qaeda" + 0.074*"char bers" + 0.066*"program" + -0.065*"vietnam" + 0.065*"hussein" + 0.064*"tariff" + -0.064*"court"')

   These topic may refer to Saddan Hussain and Al-Qaeda and the terrorist attacks.

8. ('0.122*"california" + -0.112*"chambers" + 0.104*"isthmus" + 0.091*"panama" + -0.077*"slave" + -0.071*"iglesias" + -0.069*"consols" + -0.062*"slavery" + 0.058*"canal" + -0.055*"cuba"')

   These topics are related to Isthmus of Panama and the slavery in the Cuba.

9. ('-0.115*"railway" + 0.083*"consols" + 0.077*"autocracy" + -0.064*"insistent" + -0.059*"freight" + 0.059*"iglesias" + -0.057*"tribunal" + -0.057*"motor" + -0.057*"readjustr + -0.055*"texas"')

   These topics refer to the Church, autocracy and tribunal readjustment.

10. ('0.105*"inflation" + -0.094*"billion" + -0.085*"cable" + -0.084*"oil" + 0.083*"gold" + 0.081*"job" + 0.068*"coinage" + -0.065*"barrel" + 0.058*"recovery" + -0.056*"exchequer

    These topics refer to the great depression as words like inflation, recovery, coinage are present.

LSI algorithm is the simplest method which is easy to understand and implement. It also offers better results compared to the vector space model. It is faster compared to other available algorithms because it involves document term matrix decomposition only.

Here, 250 Topics were discovered using Latent Semantic Indexing.Some of the topics were actually captured the real human concepts very well but some of the other topics were totally gibberish. Some of them are overlapping topics. For Capturing multiple meanings with higher accuracy we need to try LDA(latent Dirichlet allocation).

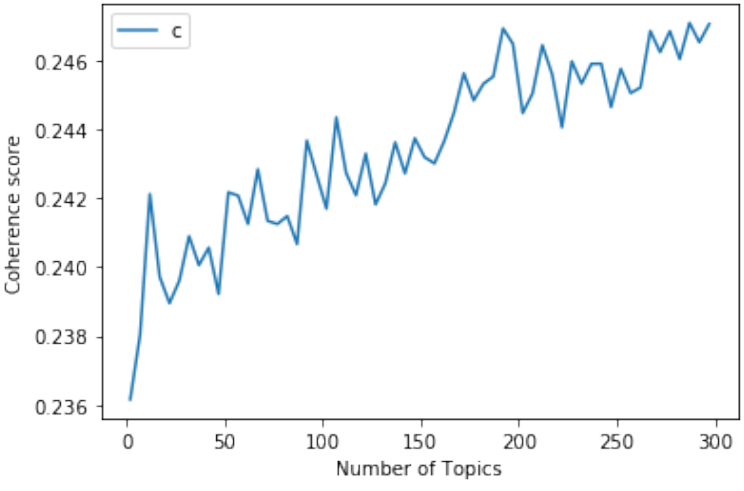
5 Task 3: LDA topic modeling

Latent Dirichlet allocation (LDA) is the most common and popular technique currently in use for topic modeling. LDA is a probabilistic topic modeling technique. In topic modeling we assume that in any collection of interrelated documents, there are some combinations of topics included in each document. The main goal of probabilistic topic modeling is to

discover the hidden topic structure for collection of interrelated documents.

LSI worked out great, based on an tf-idf transformation, however, LDA based on the tf-idf transformed corpus has no success, giving non-sense topics. Then I tried again with plain bow, it works. However, there are words that are more widely shared across documents and also have higher within document frequency, it would be ideal to to weigh their frequency/counts using tf-idf before doing LDA. Otherwise, it seems to me that the topic weighting given by LDA has more weights on those topics represented by those more frequent words

To find out the optimal number of topics I again used the same coherence measure. We observed that the coherence keep on increasing with number of topics thus makes it difficult to select the optimal number of topics.



The 10 randomly sampled topics are shown below along with their annotations:

1. (0, '0.001*"program" + 0.001*"tonight" + 0.001*"job" + 0.001*"economic" + 0.001*"upon" + 0.001*"help" + 0.001*"budget" + 0.001*"americans" + 0.001*"today" + 0.001*"billion"')

   These may refers to Economy and the budget of the country.

2. (3, '0.000*"lawfully" + 0.000*"press" + 0.000*"borne" + 0.000*"mexico" + 0.000*"healthiest" + 0.000*"bringing" + 0.000*"infrastructure" + 0.000*"territory" + 0.000*"survey" + 0.000*"ascertained"')

   These topics may refer to the Maxico and there infrastructure.

3. (13, '0.000*"revival" + 0.000*"rocket" + 0.000*"bluntly" + 0.000*"resentment" + 0.000*"sage" + 0.000*"topeka" + 0.000*"argentina" + 0.000*"planned" + 0.000*"draft" + 0.000*"commenting"')

   These topics may refer to the relations between America and Argentina.

4. (8, '0.000*"wrong" + 0.000*"docket" + 0.000*"benefitted" + 0.000*"appraiser" + 0.000*"maysville" + 0.000*"rifle" + 0.000*"assaulted" + 0.000*"insufficiently" + 0.000*"me iting" + 0.000*"fired"')

   These topics may refer to some kind of assault or attack.

5. (20, '0.000*"hurricane" + 0.000*"coordination" + 0.000*"civilized" + 0.000*"aggrieved" + 0.000*"wise" + 0.000*"weak" + 0.000*"reduce" + 0.000*"poorer" + 0.000*"taxed" + 0.000*"following"')

   These topics mau refer to the poors and taxe and hurricane.

6. (1, '0.000*"program" + 0.000*"economic" + 0.000*"tonight" + 0.000*"budget" + 0.000*"soviet" + 0.000*"nations" + 0.000*"vietnam" + 0.000*"help" + 0.000*"need" + 0.000*"employer"')

   These topics may refer to the economy and the World War II.

7. (5, '0.000*"hunger" + 0.000*"boundary" + 0.000*"nazis" + 0.000*"november" + 0.000*"pro + 0.000*"numerical" + 0.000*"maritime" + 0.000*"dollar" + 0.000*"rhine" + 0.000*"salient'

   These topics may also refer to World War II.

8. (9, '0.000*"barracks" + 0.000*"insurance" + 0.000*"permit" + 0.000*"bigotry" + 0.000*"sev + 0.000*"americans" + 0.000*"subsidizing" + 0.000*"glaring" + 0.000*"tenderness" + 0.000*"mortification"')

   These topics may refer to Insurances and the punishments regarding it.

9. (9, '0.000*"government" + 0.000*"must" + 0.000*"year" + 0.000*"congress" + 0.000*"one" + 0.000*"nation" + 0.000*"states" + 0.000*"new" + 0.000*"country" + 0.000*"people"')

   These topic may refer to the new government and the Congress.

10. 198, '0.000*"war" + 0.000*"year" + 0.000*"country" + 0.000*"government" + 0.000*"nation" + 0.000*"u" + 0.000*"great" + 0.000*"united" + 0.000*"right" + 0.000*"upon"')

    These topics may refer to the wars and the rights of the people

Thus after performing both LDA and LSI we can observe that LDA topics typically "look better", more coherent and easier to interpret but LSI is faster by a constant factor. Topics which were result of LDA were much more comprehensive according to what i observed, they were more diverse.

6 Task 4: Figure out how topics of speeches have changed over time

To analyze the change in topics of the speeches over each decade I grouped the speeches of every decade and then run the LDA on each of them individually. We have 23 such decades. We chose LDA over LSI since it give more accurate and less overlapping topics, so we can have more insight on how topics change over time.

For each decade I found optimal number of topics to give to LDA as latent variable were 200 and thus I ran LDA for 200 topics. For each decade i have mentioned below the 5 random topics which can summarize the speeches for that decade and from that we can observe how the topics change over time.

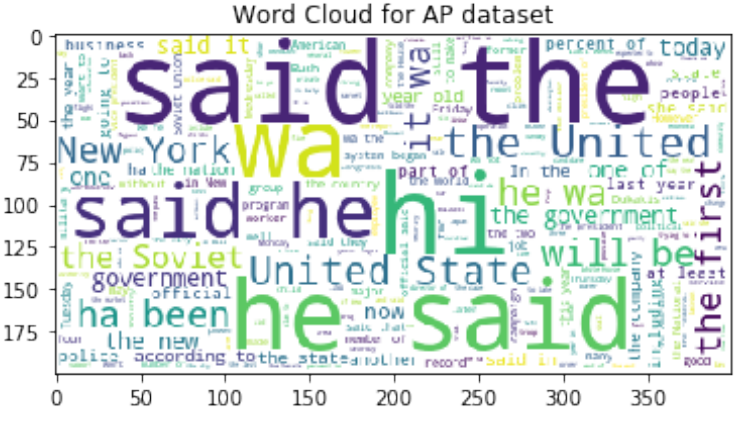
1. 1790-1799: Major topics were related to Indian public situation, military laws, public war, treaty of Paris and France commerce.
2. 1800-1809: Major topics were related to Debt of citizens, employment, British Harbors, Commerce and rights.
3. 1810-1819: Major topics were related to US trading to France and Britain, Florida Purchase Treaty.
4. 1820-1829: Major topics were related to public act 1827, War, public improvement.
5. 1830-1839: Major topics were related to treaty, Finance, Commerce and banks.
6. 1840-1849: Major topics were related to American-Mexican war, public treasury, and treaty of California.
7. 1850-1859: Major topics were related to Indian affairs, Britain and fiscal year ending.
8. 1860-1869: Major topics were related to labour capital, Constitution, American civil war.
9. 1870-1879: Major topics were related to treaty between US and Britain, Coinage, Currency and public relations.
10. 1880-1889: Major topics were related to foreign countries, tariff and taxes and some kind of treaty.
11. 1890-1899: Major topics were related to treasury notes, war with Spain and treaty with Spain.
12. 1900-1909: Major topics were related to Isthmus of Panama, Copyright Act of 1909, business.
13. 1910-1919: Major topics were related to labor war, Business and World War I.
14. 1920-1929: Major topics were related to Railway, national wealth and cultural Civil War.
15. 1930-1939: Major topics were related to banks, Industries, national Income and labors.

16. 1940-1949: Major topics were related to World War II, labors, housing and peace.
17. 1950-1959: Major topics were related to Economy, Soviets and military.
18. 1960-1969: Major topics were related to Economy, taxes, Vietnam, budget.
19. 1970-1979: Major topics were related to energy, oil, inflation, nuclear power.
20. 1980-1989: Major topics were related to tax, economy, budget and spending.
21. 1990-1999: Major topics were related to Children, freedom and job plans.
22. 2000-2009: Major topics were related to budget, tax, Iraq, mortgage.
23. 2010-2012: Major topics were related to jobs, health and economy.

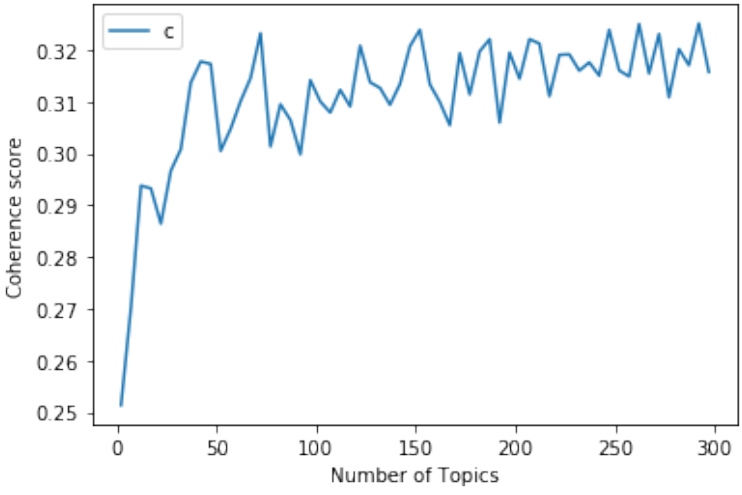
Key point of every decade is shown above. The topics vary diversely over the time. Topics related to World war I, World War II, Civil rights, Great depression, recession and many more historical events were observed.

7 Analysis of collection of AP wire stories

For this data-set also I did the similar approach. I first cleaned the data-set by removing the unwanted characters like punctuation, digits, special characters and stop-words. Then I created the dictionary for the data-set and then created the bag of words corpus. Since LDA works well with bag of words corpus I didn't create TFIDF corpus.



To find the optimal number of topics I again plotted coherence values with number of topics and found that the graph was saturated after 100 topics. Thus I ran the model with number of topics = 100.



Here are the 10 randomly selected annotated topics from the results:

1. ('0.006*"cent" + 0.005*"dollar" + 0.005*"yen" + 0.004*"gold" + 0.004*"franc" + 0.004*"troy" + 0.003*"ounce" + 0.003*"bid" + 0.003*"late" + 0.003*"london" + 0.002*"w + 0.002*"said" + 0.002*"lower" + 0.002*"pound" + 0.002*"higher" + 0.002*"lira" + 0.002*"dealer" + 0.002*"silver" + 0.001*"mark" + 0.001*"future"'),

   This topics is related to money and England.

2. ('0.000*"said" + 0.000*"party" + 0.000*"wa" + 0.000*"would" + 0.000*"sajudis" + 0.000*"new" + 0.000*"ha" + 0.000*"usda" + 0.000*"year" + 0.000*"program" + 0.000*"department" + 0.000*"million" + 0.000*"people" + 0.000*"official" + 0.000*"government" + 0.000*"bush" + 0.000*"soviet" + 0.000*"york" + 0.000*"state" + 0.000*"loan"

   This topic is related to American government.

3. ('0.016*"said" + 0.016*"percent" + 0.012*"wa" + 0.008*"price" + 0.008*"year" + 0.008*"market" + 0.007*"stock" + 0.007*"million" + 0.006*"rate" + 0.006*"new" + 0.005*"billion" + 0.004*"cent" + 0.004*"share" + 0.004*"dollar" + 0.004*"rose" + 0.004*"oil" + 0.003*"sale" + 0.003*"point" + 0.003*"month" + 0.003*"exchange"'),

   This topic is related to the stock market.

4. ('0.001*"transistor" + 0.001*"wa" + 0.000*"function" + 0.000*"temperature" + 0.000*"supercomputer" + 0.000*"sandia" + 0.000*"superconductivity" + 0.000*"superconductive" + 0.000*"chilled" + 0.000*"nordman" + 0.000*"said" + 0.000*"year" + 0.000*"superconducting" + 0.000*"voltage" + 0.000*"circuitry" + 0.000*"fahrenheit" + 0.000*"electronics" + 0.000*"first" + 0.000*"defense" + 0.000*"u"'),

   This topic is related to electronics and technology.

5. ('0.002*"said" + 0.001*"wa" + 0.001*"would" + 0.001*"ha" + 0.000*"soviet" + 0.000*"people" + 0.000*"year" + 0.000*"one" + 0.000*"two" + 0.000*"today" + 0.000*"president" + 0.000*"last" + 0.000*"u" + 0.000*"defense" + 0.000*"union" + 0.000*"also" + 0.000*"official" + 0.000*"new" + 0.000*"bush" + 0.000*"pacs"'),

This topic is related to Union ministry of America and and Soviets.

6. ('0.000*"said" + 0.000*"wa" + 0.000*"year" + 0.000*"eastern" + 0.000*"late" + 0.000*"monday" + 0.000*"would" + 0.000*"first" + 0.000*"security" + 0.000*"dollar" + 0.000*"computer" + 0.000*"company" + 0.000*"three" + 0.000*"federal" + 0.000*"air" + 0.000*"two" + 0.000*"ha" + 0.000*"rate" + 0.000*"yen" + 0.000*"nation"'),

   This topic is related to security.

7. ('0.002*"cdy" + 0.002*"clr" + 0.001*"rn" + 0.000*"said" + 0.000*"new" + 0.000*"aires" + 0.000*"mexico" + 0.000*"paris" + 0.000*"nassau" + 0.000*"city" + 0.000*"london" + 0.000*"rome" + 0.000*"lima" + 0.000*"caracas" + 0.000*"tokyo" + 0.000*"oslo" + 0.000*"rio" + 0.000*"singapore" + 0.000*"delhi" + 0.000*"harare"'),

   This topic is related to different cities of the world.

8. ('0.000*"said" + 0.000*"wa" + 0.000*"aid" + 0.000*"would" + 0.000*"government" + 0.000*"new" + 0.000*"year" + 0.000*"u" + 0.000*"stock" + 0.000*"ha" + 0.000*"baker" + 0.000*"also" + 0.000*"support" + 0.000*"political" + 0.000*"million" + 0.000*"case" + 0.000*"united" + 0.000*"one" + 0.000*"could" + 0.000*"coast"'),

   This topic is related to politics and government.

9. ('0.001*"musical" + 0.001*"said" + 0.001*"wa" + 0.000*"year" + 0.000*"ha" + 0.000*"featured" + 0.000*"bjornson" + 0.000*"new" + 0.000*"webber" + 0.000*"drama" + 0.000*"wou + 0.000*"butterfly" + 0.000*"actress" + 0.000*"one" + 0.000*"day" + 0.000*"best" + 0.000*"two" + 0.000*"department" + 0.000*"actor" + 0.000*"also"'),

   This is topic is related to Entertainment.

10. ('0.001*"tyson" + 0.000*"givens" + 0.000*"divorce" + 0.000*"weitzman" + 0.000*"drug" + 0.000*"said" + 0.000*"wa" + 0.000*"champ" + 0.000*"united" + 0.000*"annulment" + 0.000*"new" + 0.000*"dissolving" + 0.000*"stock" + 0.000*"year" + 0.000*"ha" + 0.000*"would" + 0.000*"manipulated" + 0.000*"quicker" + 0.000*"marriage" + 0.000*"heavyweight"'),

    This topic is related to sports, specifically wrestling.

Topics is this set of documents after applying LDA were much more clearer. These topics were diverse and easy to comprehend compared to State of the Union data-set. Each topics contains words specifically related to a particular genre and was easy to understand. Although the number of unique tokens in the data-set were almost equal but still the result of the LDA on this data-set was much better.

After comparing to the previous LDA result on this document I found that the most of the words match and the results were almost similar. We got diverse set of words from the results of LDA.

8 Conclusion

In this assignment I applied different natural language processing techniques and Topic modelling techniques which are LDA and LSI. I implemented these topics on the two different dataset and observed the difference in functionality of both the algorithms. I found out the optimal number of topics required in each case. I managed to extract the topic from the document and tried to annotate them.

After implementation of both LSI and LDA i can conclude that If you have a large corpus of news you know you are interested in, maybe LDA will offer better results because it can detect a model of the topics in that specific news. Otherwise, if you only have a few news to start from, it would be better to find similar ones using LSI. LDA tries to determine an appropriate prior for the corpus/language rather than using just the one fits all simple 2nd order correlation that underlies LSI. Of course there is the question of what assumptions are made in estimating the priors of LDA, and how valid these are. But generally these assumptions are more specific and informative even when not entirely correct or complete, and so LDA will give better performing models than LSI.