Assignment 4: Topic Modeling

Gurpreet Singh
2018csb1092@iitrpr.ac.in

Indian Institute of Technology
Ropar, Punjab, India

Abstract

The main aim of this assignment is to give initial insight to topic modeling. It is a technique used in Natural Language Processing. This text mining tool is used to extract deeper information from a collection of texts. For this assignment, we have used the GENSIM library to apply topic modeling.

1 Dataset 1: Tasks 1, 2, 3 and 4

For the first three tasks, the dataset given is the State of Union addresses by the US presidents from 1790 to 2012. A few years contain more than one sppech. Also, speeches for many years are missing from the .csv file. In total there are 155 speeches.

1.1 Data Preprocessing

Firstly, all the addresses are converted to lower case. After that, a list of special characters is prepared which contain new line character, full stop, comma, semi colon, question mark, exclamation mark and all types of parenthesis. All the occurences of these characters in the dataset are replaced with spaces. Then, the text is tokenized by splitting the sentences about spaces. Those tokens which do not occur in the list 'stoplist' are kept for further analysis. 'Stoplist' is prepared by using suggestions given in the assignment itself.
Then a list which stores frequency of all the words is obtained. All those words which occur just once are then discarded. From this, we get the 'bag of words' representation using 'doc2bow' function.

1.2 Generation of tf-idf weighted vectors

Tf-idf stands for term frequency- inverse document frequency. It is used to give weights to words which are proportional to their frequency in a document, but inversely proportional to the number of documents which contain that word. Gensim library is used to do convert 'Bag of Words' representation to tf-idf weighted vectors.We obtain a list of 155 sub-lists which contain this vectorized form of all the words in 155 speeches.

1.3 LSI Topic Modeling

LSI is also known as LSA (Latent Semantic Analysis). The maximum number of topics that we can obtain is 155. Beyond that point, the number of topics generated remains 155. If we check coherence scores vs number of topics plot for LSI model, we find that the peak **is in**

the range 130 to 140, slightly less than 135. Also, the corresponding mark for 155 is also not very less. Both plots clearly depict peak in the same region.

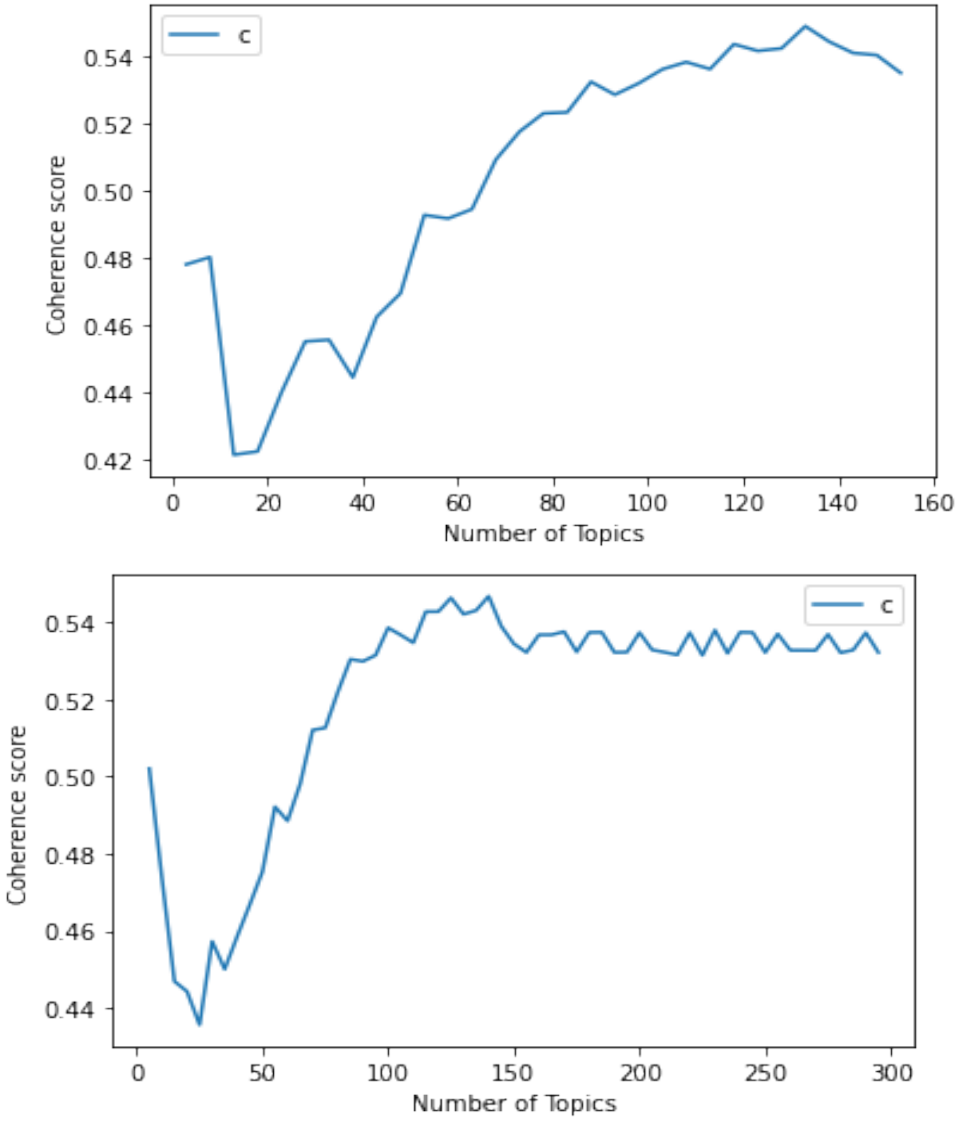

Fig-1.3(a) For LSA

From a rough glance at the topics generated we can get a feel for some events like World War 2, Saddam Hussein's activities in Iraq, external affairs with Japan and Spain, actions against Al-Qaida and Taliban in Afghanistan, sense of togetherness and the idea of a new nation (during the first speech), etc. Let us examine any 10 out of 135 topics generated.

Fig-1.3(b), 1.3(c) and 1.3(d)

In 1.3(b), part 1 discusses about Vietnam and Japan. Years mentioned are 1942 and 1953. Its theme is possibly regarding US-Vietnam war. Part 2 mentions Hussein and Hitler which is a clear indication of its connection with Iraq and its dictator Saddam Hussein. Hitler is probably used here to give analogy. Part 3 discusses about democracy. 1827 and 1878 are mentioned. However, choosing one topic theme is difficult here.
In 1.3(c) part 1, Saddam Hussein is mentioned here again. He is the theme of this topic too. In part 2, hunger and damages are mentioned. The probable theme of this topic is regarding poverty. In part 3, the discussion is probably about economy and external affairs. But the conclusions here are not as concrete.
In 1.3(d), part 1 mentions Vietnam and communism. So ,this seems to be a talk about the region of China and South-East Asia. It is very difficult to make sense out of Part 2.But still there is a mention of provinces and articles. This can be point towards the federal structure, power sharing, local and inter state affairs. Part 3 mentions Al-qaida, Afghanistan and hostages. This is definitely about some terrorist activity carried out by Al-Qaida. Part 4 seems very interesting. It mentions notes, paper, silver, currency, circulation and rebellion. Economy is the protagonist here for sure. But the word 'rebellion' adds a little twist. It might be indicative of the Civil War.
To sum up, analysis using LSI was very informative and useful.

1.4 LDA Topic Modeling

LDA stands for Latent Dirichlet Allocation. Unlike LSA, there is not any significant upper limit restriction on the number of topics here (The maximum I tried was around 1000). We again use the same method as LSA to check optimal number of topics in LDA.
From the coherence score plot, we get that the maximum score is occuring at number of topics equal to 35. **So, the optimal number is 35.**

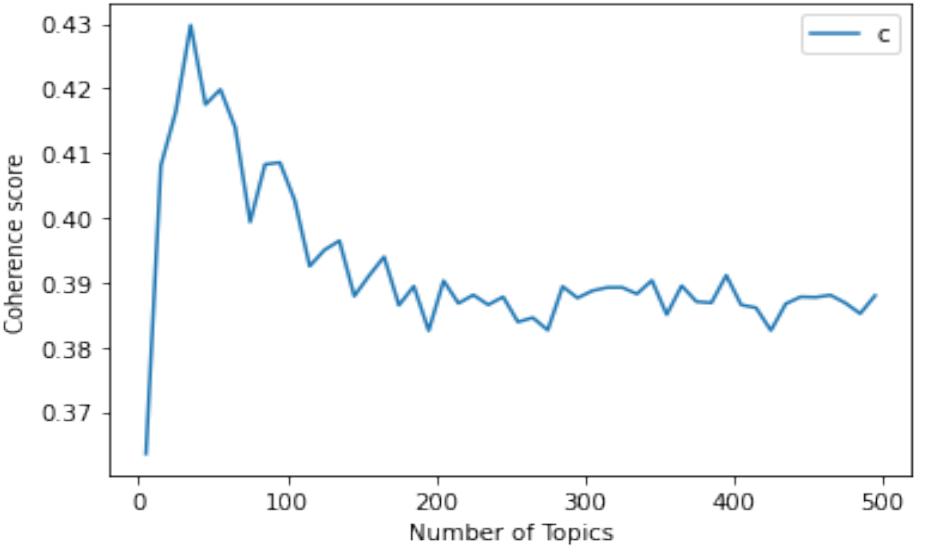

Fig-1.4(a)

All the 35 topics have been printed in the topic. Let us select any 10 out of them. It is very difficult to get something significant out of these 10 topics. All the topics generated seem to be a mix of 'government', 'united', 'people', 'state' and 'congress'. The performance of LDA is very poor as compared to LSI, here. The ten topics which we sampled are nearly impossible to differentiate from each other. The performance of LDA is very much influenced by modals- 'will', 'can' and 'may', pronouns and prepositions. The stoplist was same for LDA and LSI. Even then, LSI produced far better results.

```
[ ] lda_model.show_topic(15)

[('will', 0.016510025),
 ('must', 0.006928436),
 ('can', 0.0061527365),
 ('united', 0.0060042962),
 ('government', 0.0059873736),
 ('people', 0.005278812),
 ('congress', 0.0051245587),
 ('us', 0.0050686435),
 ('year', 0.005042023),
 ('one', 0.0043523307)]
```

```
[ ] lda_model.show_topic(19)

[('will', 0.010163654),
 ('government', 0.0077068997),
 ('must', 0.0071402946),
 ('states', 0.0051086713),
 ('us', 0.0046383967),
 ('world', 0.0046247444),
 ('can', 0.004391574),
 ('congress', 0.004288021),
 ('year', 0.0041529457),
 ('people', 0.004143453)]
```

```
[ ] lda_model.show_topic(17)

[('will', 0.01479972),
 ('must', 0.007606434),
 ('government', 0.0072158566),
 ('us', 0.005724691),
 ('people', 0.005284076),
 ('congress', 0.0052839634),
 ('united', 0.0051261256),
 ('can', 0.0050686053),
 ('world', 0.004887835),
 ('states', 0.0043795537)]
```

```
[ ] lda_model.show_topic(8)

[('will', 0.011218465),
 ('government', 0.008520857),
 ('united', 0.006148245),
 ('states', 0.0050669923),
 ('people', 0.0052705654),
 ('can', 0.005111385),
 ('may', 0.0046788733),
 ('congress', 0.0046211877),
 ('country', 0.004057839),
 ('must', 0.003949522)]
```

```
[ ] lda_model.show_topic(7)

[('will', 0.012592978),
 ('government', 0.0071704597),
 ('states', 0.0068256347),
 ('congress', 0.006195552),
 ('can', 0.005533145),
 ('people', 0.005294831),
 ('new', 0.0048970347),
 ('world', 0.004655728),
 ('must', 0.0045595467),
 ('america', 0.004422013)]
```

```
[ ] lda_model.show_topic(32)

[('will', 0.013705566),
 ('government', 0.006362667),
 ('congress', 0.0063563953),
 ('united', 0.006173442),
 ('can', 0.006073642),
 ('states', 0.005350847),
 ('must', 0.004513758),
 ('people', 0.0042024194),
 ('country', 0.0039204345),
 ('may', 0.003844109)]
```

Fig-1.4(b),(c) For LDA

1.5 A General Analysis:Task 4

For visualising data obtained from LDA, we have used the library 'pyLDAvis'. It produces marvellous visualisations and graphics. They give a very good insight into the significance of all the topics. The html page which is generated using this library has been embedded in the folder submitted. On that page, one can vary relevance metric for any topic and can see top 30 words that are contained in that topic. Also, its image has been attached in this report.

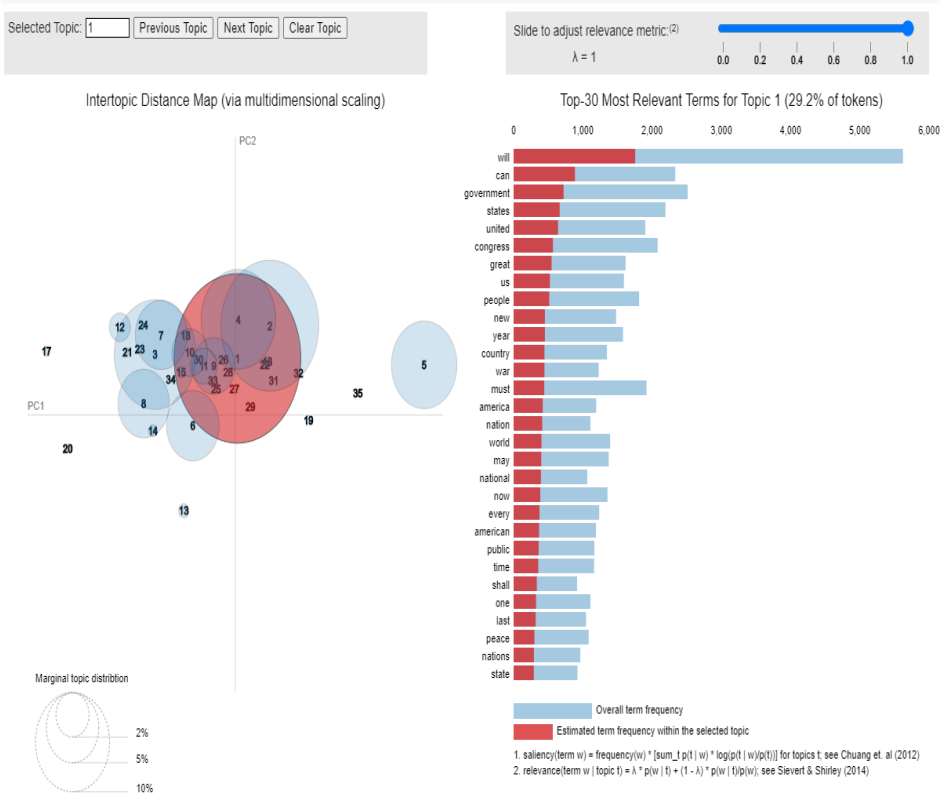

Fig-1.5(a) Number of topics=35

2 Dataset 2: Task 5

Dataset chosen for this task is the collection of AP wire stories. Data preprocessing is done in same way as dataset 1. After that, tf-idf weighted vectors are generated. Then, the optimal number of topics is chosen for LDA by following the same procedure as task 3. From the plot, it is clear that the peak is observed at number of topics equal to 15.

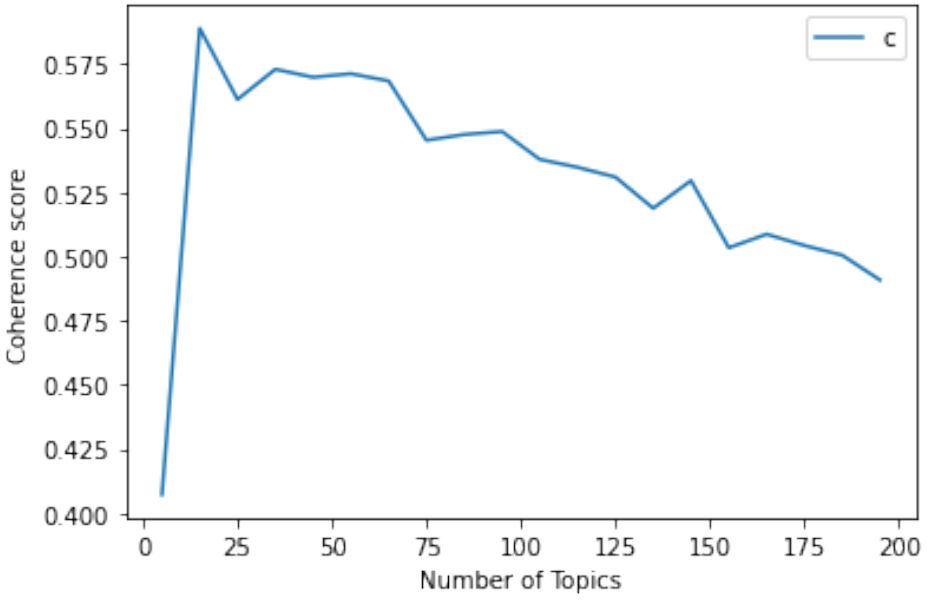

Fig-2(a)


```
[(0,
  '0.015*"said" + 0.008*"\'\'" + 0.005*"stock" + 0.004*"corroon" + 0.004*"cdy" + 0.004*"percent" + 0.004*"market" + 0.004*"black" + 0.004*"clr" + 0.004*"2"'),
 (1,
  '0.015*"said" + 0.011*"\'\'" + 0.005*"poll" + 0.005*"dukakis" + 0.005*"new" + 0.005*"percent" + 0.004*"will" + 0.003*"party" + 0.003*"also" + 0.003*"last"'),
 (2,
  '0.024*"said" + 0.012*"\'\'" + 0.007*"year" + 0.006*"percent" + 0.006*"s" + 0.006*"u" + 0.004*"will" + 0.004*"last" + 0.004*"owen" + 0.003*"united"'),
 (3,
  '0.027*"said" + 0.011*"percent" + 0.010*"\'\'" + 0.005*"million" + 0.004*"1" + 0.004*"u" + 0.003*"will" + 0.003*"soviet" + 0.003*"two" + 0.003*"president"'),
 (4,
  '0.021*"said" + 0.010*"\'\'" + 0.005*"president" + 0.005*"bush" + 0.005*"dresses" + 0.005*"will" + 0.004*"at&t" + 0.004*"s" + 0.004*"macmillan" + 0.003*"u"'),
 (5,
  '0.019*"said" + 0.011*"\'\'" + 0.006*"will" + 0.006*"000" + 0.005*"u" + 0.005*"s" + 0.004*"pacs" + 0.004*"new" + 0.004*"court" + 0.003*"united"'),
 (6,
  '0.014*"said" + 0.010*"\'\'" + 0.009*"1" + 0.007*"late" + 0.006*"dollar" + 0.005*"yen" + 0.005*"000" + 0.005*"million" + 0.004*"hyundai" + 0.004*"police"'),
 (7,
  '0.023*"\'\'" + 0.021*"said" + 0.006*"south" + 0.004*"people" + 0.004*"new" + 0.004*"primary" + 0.004*"state" + 0.003*"``i" + 0.003*"president" + 0.003*"bush"'),
 (8,
  '0.023*"said" + 0.020*"\'\'" + 0.006*"u" + 0.006*"government" + 0.005*"trade" + 0.005*"will" + 0.005*"s" + 0.005*"states" + 0.005*"united" + 0.003*"new"'),
 (9,
  '0.010*"said" + 0.009*"\'\'" + 0.005*"will" + 0.004*"000" + 0.004*"president" + 0.003*"people" + 0.003*"new" + 0.003*"million" + 0.003*"also" + 0.003*"two"'),
 (10,
  '0.015*"said" + 0.012*"\'\'" + 0.005*"new" + 0.005*"million" + 0.004*"government" + 0.004*"1" + 0.003*"u" + 0.003*"today" + 0.003*"market" + 0.003*"stock"'),
 (11,
  '0.020*"said" + 0.015*"\'\'" + 0.012*"percent" + 0.007*"billion" + 0.005*"year" + 0.004*"million" + 0.004*"000" + 0.004*"last" + 0.003*"1" + 0.003*"will"'),
 (12,
  '0.020*"said" + 0.020*"\'\'" + 0.007*"will" + 0.005*"one" + 0.004*"new" + 0.003*"_" + 0.003*"s" + 0.003*"000" + 0.003*"years" + 0.003*"can"'),
 (13,
  '0.031*"said" + 0.013*"\'\'" + 0.004*"one" + 0.003*"new" + 0.003*"will" + 0.003*"also" + 0.003*"1" + 0.003*"two" + 0.003*"tuesday" + 0.002*"percent"'),
 (14,
  '0.027*"said" + 0.015*"\'\'" + 0.004*"police" + 0.004*"-" + 0.004*"people" + 0.004*"two" + 0.003*"000" + 0.003*"united" + 0.003*"government" + 0.003*"will"')]
```


Fig-2(b) All 15 topics

The results of LDA are a lot better here. A wide variety of words is captured here. The effect of modals is very diminished as compared to the previous dataset.
The visualisations obtained can be viewed much better in the web page embedded in the folder submitted. Nevertheless, an image of the page is attached here.
The term 'better' means that a variety of words are captured with proper significance on the necessary words.

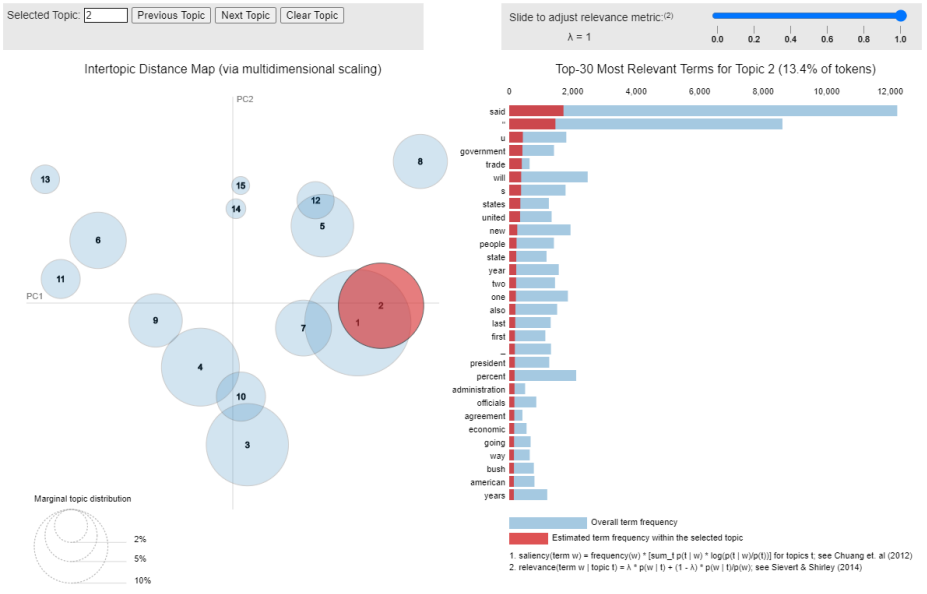
Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (13.4% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
said
"
u
government
trade
will
s
states
united
new
people
state
year
two
one
also
last
first
--
president
percent
administration
officials
agreement
economic
going
way
bush
american
years
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig-2(c)