Topic Modeling

Ankit Bhadu
2018CSB1073

Indian Institute of Technology
Ropar

Abstract

We will analyze the State of the Union speeches corpus, and try to visualize how main topics of the speeches have shifted over time in relation to historical events using Topic Modeling.

1 Introduction

We start-off by first preprocessing the data and understanding it and then move on to topic modelling. Then we will visualize the main topics of the speeches and give them a title wherever possible. We will then use a method to study how topics changed each decade during the 20th and 21st century.

1.1 Pre-processing and generating TF-IDF scores

Firstly we check out any empty rows in the file and observe that some years have empty rows for eg the years 1829,1830,etc. We remove such empty rows and also some years have more than one state of the union addresses, for such years we combine all the addresses for that particular year.

Next we need to process the text of the speech delivered. We remove any special characters present in the text and punctuatuions as those are not of use to us and then tokenise the words. This is done mainly using preprocessing module of nltk. Now each speech is a series of words.

We use the stopwords from both scikit learn and nltk libraries as they both have a different set of stopwords and thus combining both we can remove almost all the non-useful words. Now we have a series of words for each year and all those words add some meaning to the speeches.

We generate the TF-IDF(Term Frequency- Inverse Data Frequency) scores for these speeches using gensim. Formula for non-normalized weight of term i in document j in a corpus of D documents is given by

$$weight_{i,j} = wlocal(frequency_{i,j}) * wglobal(document_frequency_{i},D)$$

1.2 LSI Topic Modeling

We now have TF-ID scores and now will do topic modeling using LSI. LSI or Latent Semantic Indexing is a method where SVD is first performed on the term document matrix as the spare matrix is very large and it learns latent topics through this process. LSI is faster compared to LDA.

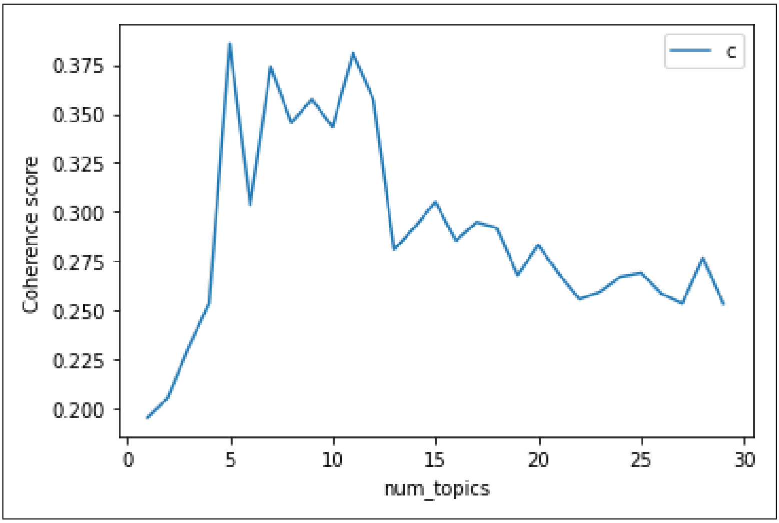


Figure 1: Coherence value vs no of topics

Deciding the number of topics:
This is a very crucial step while performing LSI topic modeling. We use the cohernece metric to determine what number of topics lead to the best human interpretability. Coherence metric measures how semantically similar are the high scoring topics and therefore helps us discard topic which received higher score just because of more occurences. We measure coherence and visualize it using a line plot(Figure 1) with number of topics ranging from 1 to 30. Maximum coherence is found for number of topic equal to 5 and then almost similar for number of topics equal to 12. We choose 12 as assignment required us to study atleast 10 topics.

Annotations:
Now we sort the keywords in each topic by their scores and plot top 12 words and their score using a bar histogram plot. Topic name has been provided manually(figure 2). Topic name is clearly visible in some of the topics but for some it was harder to decide and the decision for name wasn't always perfect. It is just indicative. We can observe that tonight and program are the two words occuring everywhere and contributing heavily. It is because they weren't in the stopwords list and are present across every speech and also multiple times in a single speech. Therefore there TF-IDF score was high leading to this event. If we talk about similarity among the topics then we can summarise all these in two words War and Economy. These are the two things spoken about in almost every State of the Union address.

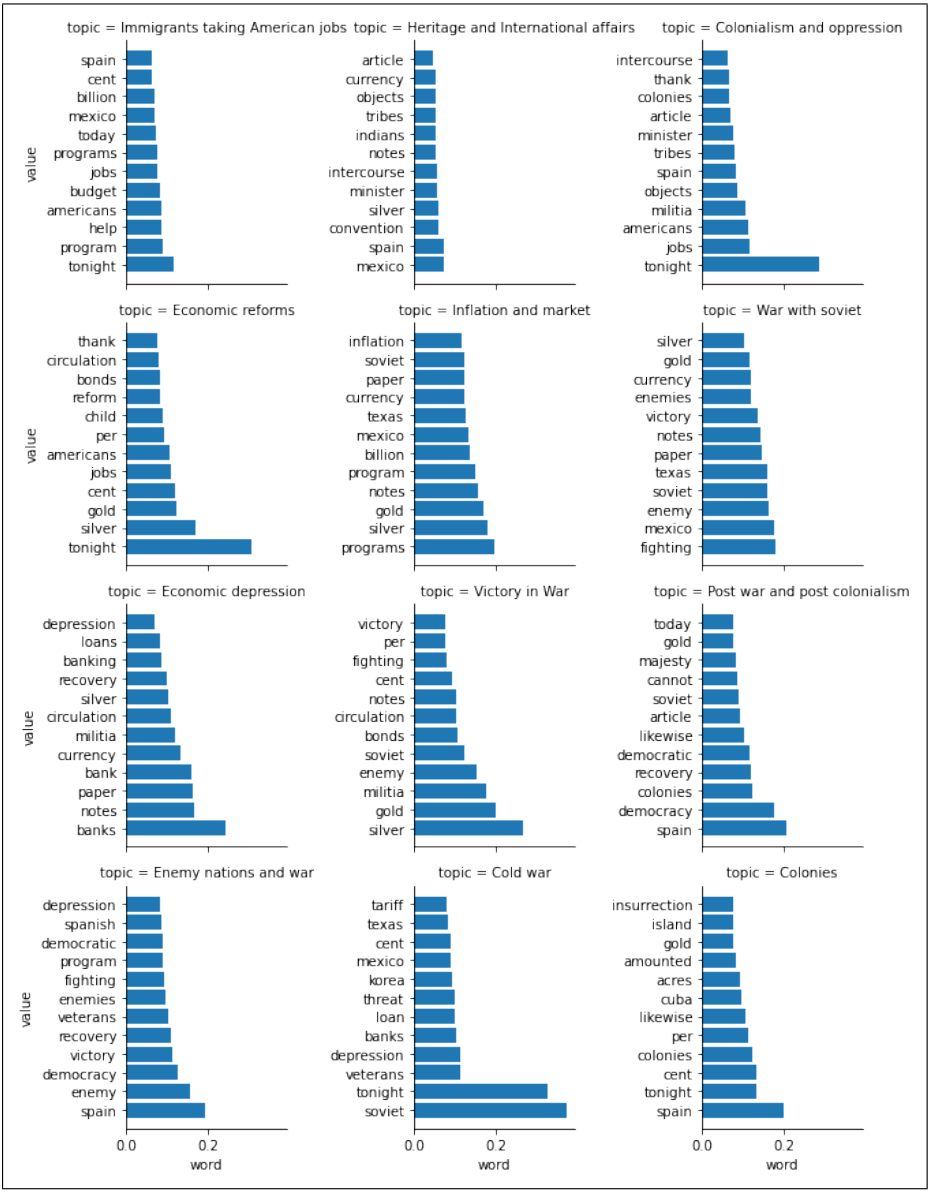


Figure 2: Topics generated using LSI topic modelling with their top contributing keywords. Annotations are done manually and written at the top of each plot.

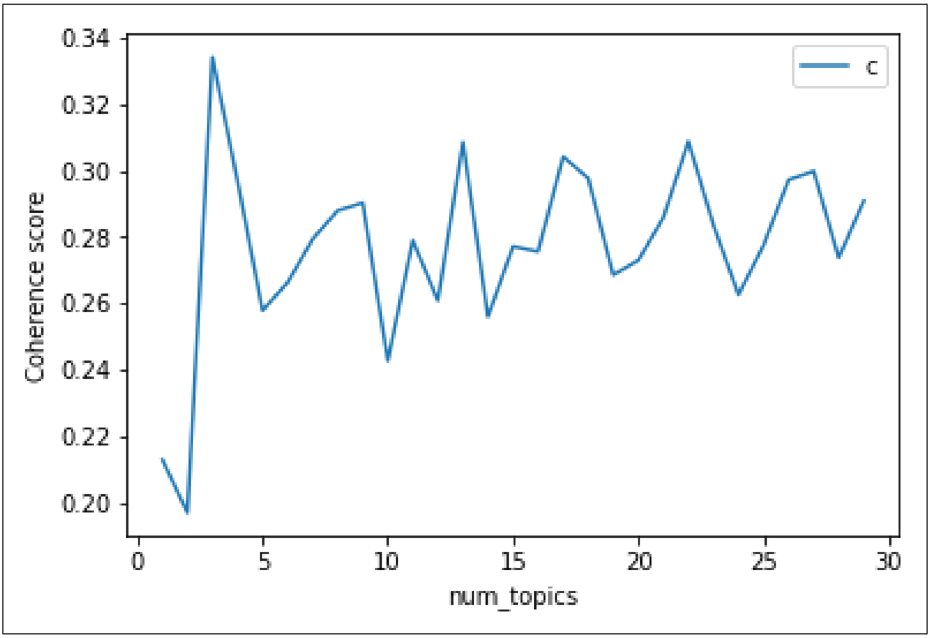


Figure 3: Coherence value vs no of topics for LDA topic modeling

1.3 LDA Topic Modeling :

LDA is an unsupervised topic modelling method and assumes Dirichlet prior over latent topics. LDA is said to have more accuracy than LSI in general but is also slower as compared to LSI. Now we will see if it holds true.

Determining number of topics: As we did in LSI now we again use the coherence metric to pick the no of topics which offers maximum human interpretability. Figure 3 shows the coherence vs no of topics plot. At a first glance we can see the maximum coherence is for no of topic equal to 4 and is less than the max coherence value of LSI topic modeling. However we continue with no of topics greater than 10 and max coherence and thus we get 13 as the optimum number.

Annotations :

Annotating the topics generated by LDA topic modeling was easier as compared to LSI except a few outliers. We employed a similar strategy for plotting as in LSI. The topics seemed more relevant and top keywords conveyed more meaning as compared to LSI although the coherence was lower compared to LSI. LSI is a linear algebraic method as compared to LDA which is a probabilistic model. For looking at topics within a speech and across speeches, according to the results, LDA performs better looking at inter speech topics.

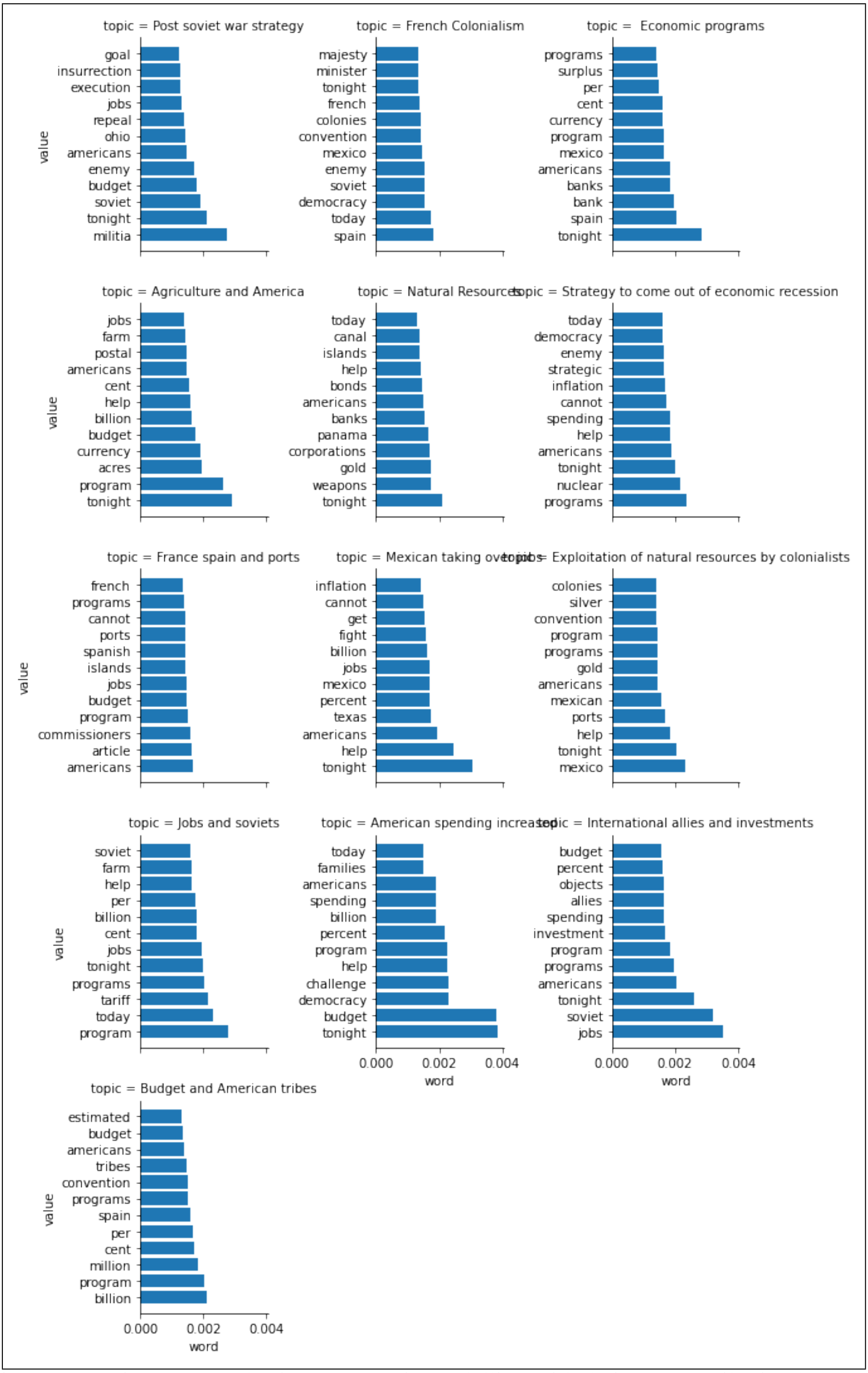


Figure 4: Topics generated using LDA topic modelling with their top contributing keywords. Annotations are done manually and written at the top of each plot.

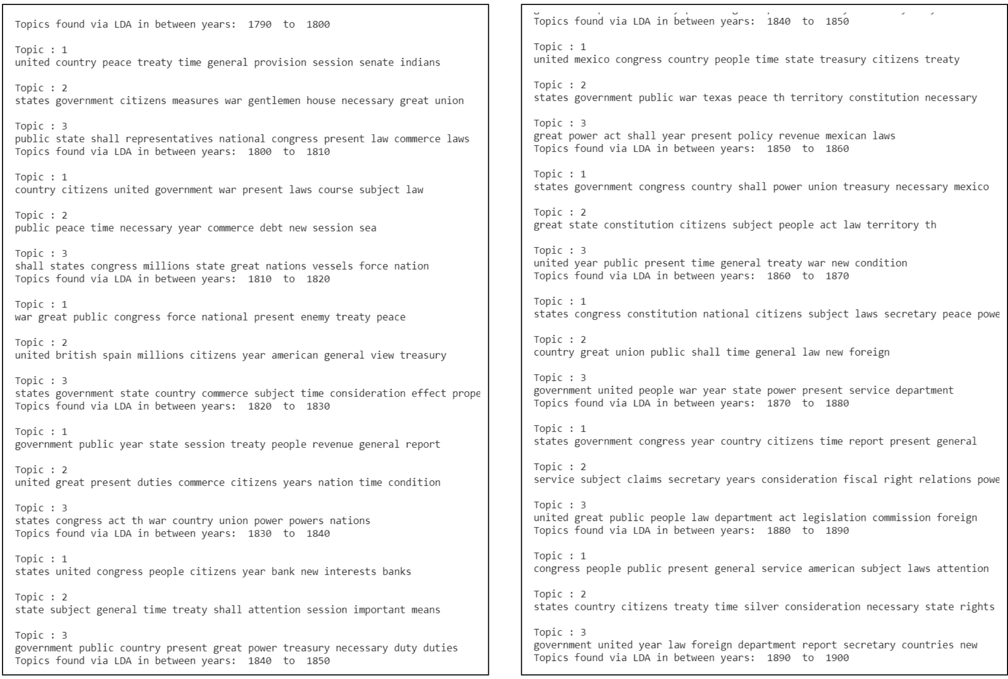
1.4 How speech topics differ over time :

Here we use LDA and generate top 3 topics during each decade and see how they change over time due to various events throughout history. For this purpose topic names are a combination of top 10 contributing keywords for that topic. For generating decade wise data we first combined Speeches over 10 year periods and then used LDA over them. Results are shown in Figure 5. There are two ways to do this, first to directly combine the speeches in one text and use lda or keep them distinct and use LDA so that cross speech frequency can help get better results.

First we make broad inferences and observe that before the 1950s the speeches were comprising more legislative and fiscal terms and it changed into more simpler terms that would make sense to the general public immediately. It is majorly because 1950-60s was the time of boom of Television in America and therefore the population watching the SOTU had changed. Later we see a shift to more war related terms majorly because of a rising spirit of nationalism and violence as the modern era begins. Modern era speeches also comprised more of healthcare and college policies as they made direct impact on the public which has now widened due to various technical platforms.

Another way to infer is how in decades which were near economic recession such as 1920's, 1940-50's, 1970s and the late 2000s we see more economic reforms and jobs related terms/topics as these were the topics of interest over the short run. Then we can observe how during times of war such as 1910-1920 and 1940-1950 and the cold war in 70s, the topics revolved around military, appraisal of the military, economy(post war years), allies, enemy nations, etc.

We can also observe how topics like Islam, Afghanistan, Jihad, etc are used more frequently during decades that USA was attacked by Terrorists.



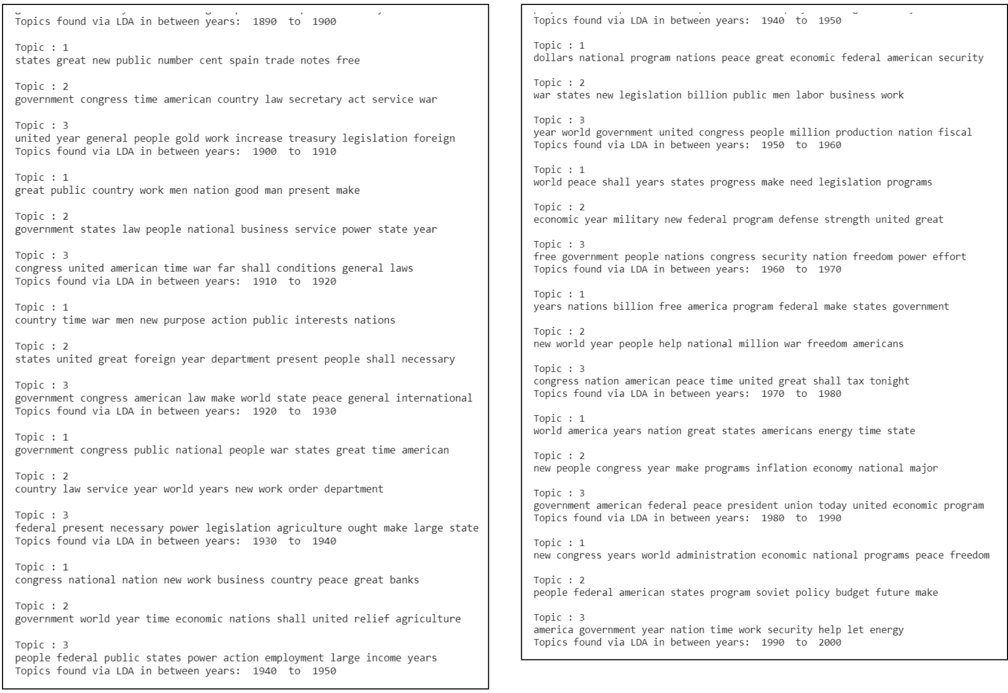
```
Topics found via LDA in between years:  1790  to  1800

Topic : 1
united country peace treaty time general provision session senate indians

Topic : 2
states government citizens measures war gentlemen house necessary great union

Topic : 3
public state shall representatives national congress present law commerce laws
Topics found via LDA in between years:  1800  to  1810

Topic : 1
country citizens united government war present laws course subject law

Topic : 2
public peace time necessary year commerce debt new session sea

Topic : 3
shall states congress millions state great nations vessels force nation
Topics found via LDA in between years:  1810  to  1820

Topic : 1
war great public congress force national present enemy treaty peace

Topic : 2
united british spain millions citizens year american general view treasury

Topic : 3
states government state country commerce subject time consideration effect prope
Topics found via LDA in between years:  1820  to  1830

Topic : 1
government public year state session treaty people revenue general report

Topic : 2
united great present duties commerce citizens years nation time condition

Topic : 3
states congress act th war country union power powers nations
Topics found via LDA in between years:  1830  to  1840

Topic : 1
states united congress people citizens year bank new interests banks

Topic : 2
state subject general time treaty shall attention session important means

Topic : 3
government public country present great power treasury necessary duty duties
Topics found via LDA in between years:  1840  to  1850
```

(a)

```
Topics found via LDA in between years:  1840  to  1850

Topic : 1
united mexico congress country people time state treasury citizens treaty

Topic : 2
states government public war texas peace th territory constitution necessary

Topic : 3
great power act shall year present policy revenue mexican laws
Topics found via LDA in between years:  1850  to  1860

Topic : 1
states government congress country shall power union treasury necessary mexico

Topic : 2
great state constitution citizens subject people act law territory th

Topic : 3
united year public present time general treaty war new condition
Topics found via LDA in between years:  1860  to  1870

Topic : 1
states congress constitution national citizens subject laws secretary peace powe

Topic : 2
country great union public shall time general law new foreign

Topic : 3
government united people war year state power present service department
Topics found via LDA in between years:  1870  to  1880

Topic : 1
states government congress year country citizens time report present general

Topic : 2
service subject claims secretary years consideration fiscal right relations powe

Topic : 3
united great public people law department act legislation commission foreign
Topics found via LDA in between years:  1880  to  1890

Topic : 1
congress people public present general service american subject laws attention

Topic : 2
states country citizens treaty time silver consideration necessary state rights

Topic : 3
government united year law foreign department report secretary countries new
Topics found via LDA in between years:  1890  to  1900
```

(b)

Figure 5: Speech topics over time



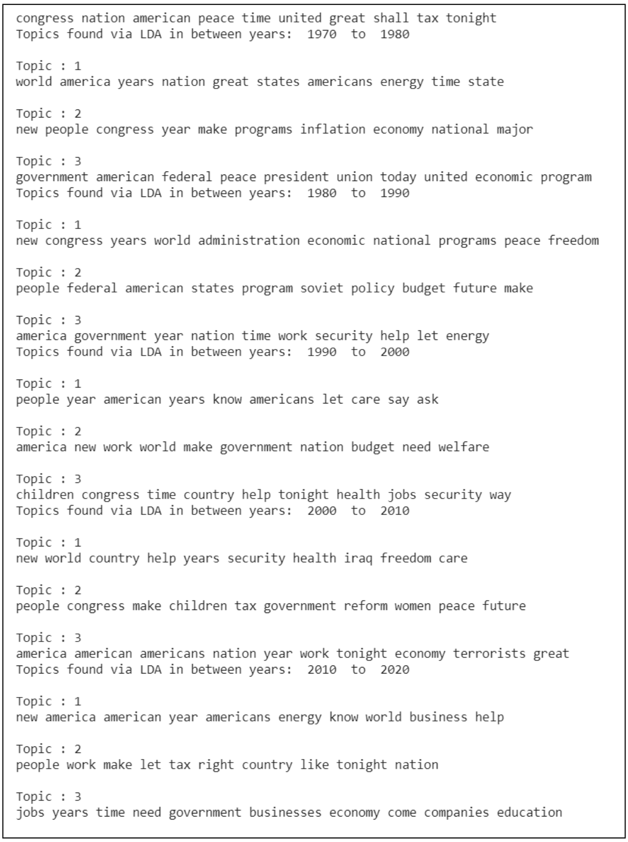
```
Topics found via LDA in between years:  1890  to  1900

Topic : 1
states great new public number cent spain trade notes free

Topic : 2
government congress time american country law secretary act service war

Topic : 3
united year general people gold work increase treasury legislation foreign
Topics found via LDA in between years:  1900  to  1910

Topic : 1
great public country work men nation good man present make

Topic : 2
government states law people national business service power state year

Topic : 3
congress united american time war far shall conditions general laws
Topics found via LDA in between years:  1910  to  1920

Topic : 1
country time war men new purpose action public interests nations

Topic : 2
states united great foreign year department present people shall necessary

Topic : 3
government congress american law make world state peace general international
Topics found via LDA in between years:  1920  to  1930

Topic : 1
government congress public national people war states great time american

Topic : 2
country law service year world years new work order department

Topic : 3
federal present necessary power legislation agriculture ought make large state
Topics found via LDA in between years:  1930  to  1940

Topic : 1
congress national nation new work business country peace great banks

Topic : 2
government world year time economic nations shall united relief agriculture

Topic : 3
people federal public states power action employment large income years
Topics found via LDA in between years:  1940  to  1950
```

(c)

```
Topics found via LDA in between years:  1940  to  1950

Topic : 1
dollars national program nations peace great economic federal american security

Topic : 2
war states new legislation billion public men labor business work

Topic : 3
year world government united congress people million production nation fiscal
Topics found via LDA in between years:  1950  to  1960

Topic : 1
world peace shall years states progress make need legislation programs

Topic : 2
economic year military new federal program defense strength united great

Topic : 3
free government people nations congress security nation freedom power effort
Topics found via LDA in between years:  1960  to  1970

Topic : 1
years nations billion free america program federal make states government

Topic : 2
new world year people help national million war freedom americans

Topic : 3
congress nation american peace time united great shall tax tonight
Topics found via LDA in between years:  1970  to  1980

Topic : 1
world america years nation great states americans energy time state

Topic : 2
new people congress year make programs inflation economy national major

Topic : 3
government american federal peace president union today united economic program
Topics found via LDA in between years:  1980  to  1990

Topic : 1
new congress years world administration economic national programs peace freedom

Topic : 2
people federal american states program soviet policy budget future make

Topic : 3
america government year nation time work security help let energy
Topics found via LDA in between years:  1990  to  2000
```

(d)

Figure 6: Speech topics over time



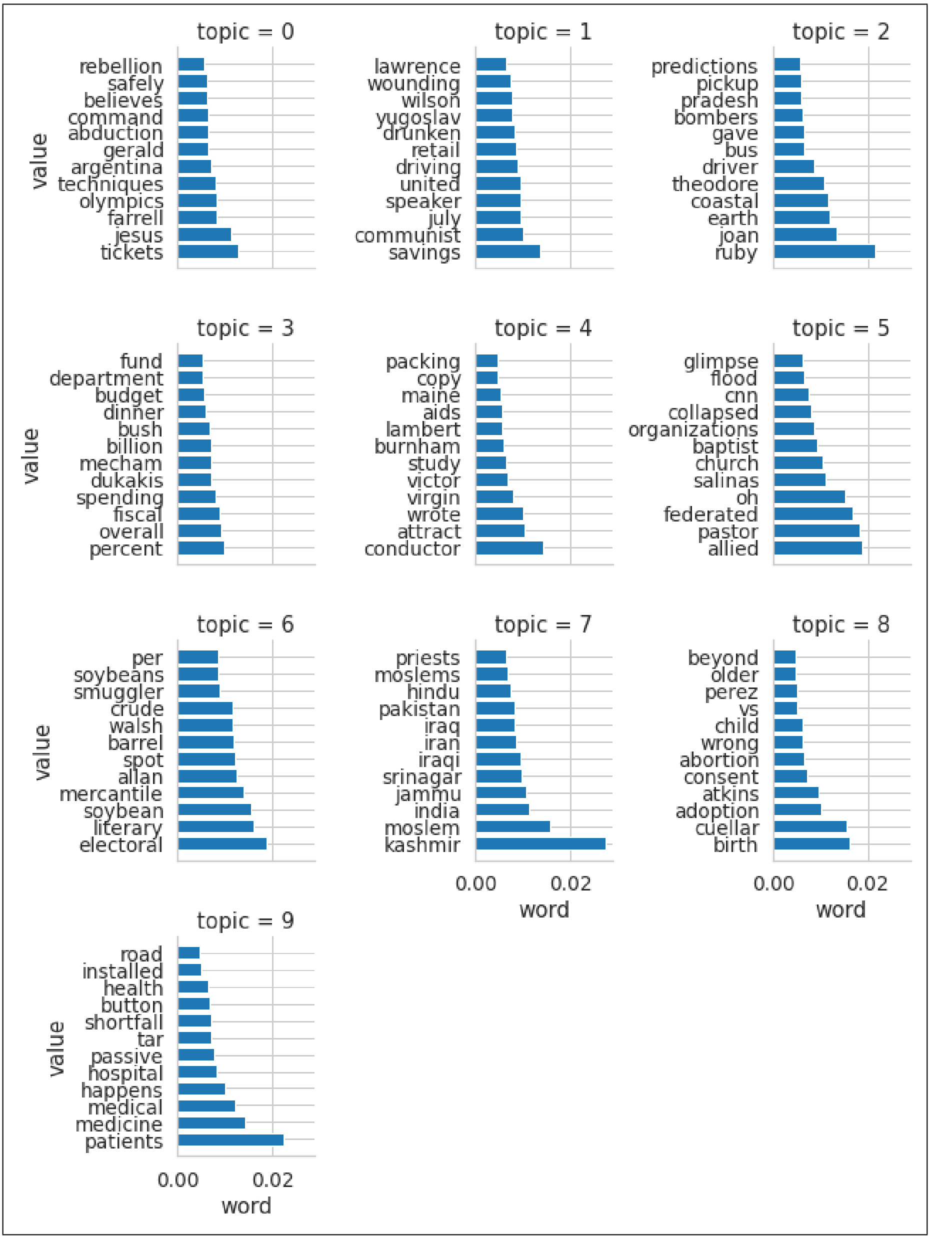
```
congress nation american peace time united great shall tax tonight
Topics found via LDA in between years:  1970  to  1980

Topic : 1
world america years nation great states americans energy time state

Topic : 2
new people congress year make programs inflation economy national major

Topic : 3
government american federal peace president union today united economic program
Topics found via LDA in between years:  1980  to  1990

Topic : 1
new congress years world administration economic national programs peace freedom

Topic : 2
people federal american states program soviet policy budget future make

Topic : 3
america government year nation time work security help let energy
Topics found via LDA in between years:  1990  to  2000

Topic : 1
people year american years know americans let care say ask

Topic : 2
america new work world make government nation budget need welfare

Topic : 3
children congress time country help tonight health jobs security way
Topics found via LDA in between years:  2000  to  2010

Topic : 1
new world country help years security health iraq freedom care

Topic : 2
people congress make children tax government reform women peace future

Topic : 3
america american americans nation year work tonight economy terrorists great
Topics found via LDA in between years:  2010  to  2020

Topic : 1
new america american year americans energy know world business help

Topic : 2
people work make let tax right country like tonight nation

Topic : 3
jobs years time need government businesses economy come companies education
```

(e)

Figure 7: Speech topics over time

1.5 Topic Modeling on AP wire stories dataset

The data in this dataset is much more diverse as the stories are not related to each other directly and are very diverse. There are a lot of names of people places, objects, etc. As compared to SOTU dataset which mainly comprised of war, economy, culture, etc, we here observe a diverse array of topics and the coherence increases as we increase number of topics upto some extent. LDA has been used for topic modelling for this AP wire stories dataset as it offers more accuracy. Results are shown in figure 8. The same process was followed and while annotating it was observed that naming the topics felt easier in the sense that this time we had a lot of options to choose from whereas in the SOTU case there were limited choices due to the content being almost the same. The topics could be for example topic 7 in figure 8 is religions, topic 5 could be christianity and disasters, topic 3 could be related to economy.



(a)

Figure 8: Top 12 contributing words for 10 selected topics after LDA topic modelling on AP Wire stories dataset

1.6 Conclusion

To summarise Topic modelling worked better on AP wire stories due to diversity in topics. If we limit the number of topics in SOTU data then there will be no problem seen in annotating as the coherence is also maximum and that is actually the no of topics seen if topics are considered to be broad. Also in comparison to LSI, LDA gave better results although it took more time.

1.7 References

1. https://en.wikipedia.org
2. https://tm4ss.github.io/docs/Tutorial_6_Topic_Models.html
3. https://medium.com/@jonathan_hui/machine-learning-latent-dirichlet-allocation-lda-1d9d
4. https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd?gi=c98
5. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b0