Topic Modelling

Harshit Sakhuja
2018eeb1044

Indian Institute Of Technology - Ropar

Abstract

Implementing LSI and LDA topic modelling using gensim package on two datasets. One is state of the union dataset and other one is collection of AP wire stories.Various inferences are gathered from the topics obtained and an attempt is made to correctly summarize the state of the union over decades in 20th and 21st century using a decade summarization algorithm.
Task 1-4 is related to state of the union dataset.
Task 5 is related to comparing the topics of both datasets.

1 Generating tf-idf weighted document vectors

First of all data was preprocessed, two preprocessing techniques were employed namely tokenization and lemmatization.(Libraries used Gensim and NLTK).
Whole corpus was traversed twice , once to generate the dictionary (consisting of unique tokens) then converting each preprocessed document into a bag-of-words representation.
After preprocessing , whole corpus is converted into document-term matrix where each column represents a document and each entry in the column represents the un-normalized term frequency. Thus each document is converted into a mathematically convenient form(vector) , gensim calls this a bag-of-words representation.
Then finally the tf-idf model(transformation) is applied to convert each bag of words vector into tf-idf vector where tf refers to term frequency and idf refers to inverse document frequency.

2 Latent Semantic Indexing (LSI) to tf-idf vectors

Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.
LSI is based on the principle that words that are used in the same contexts tend to have similar meanings.
LSI converts each tf-idf vector into another vector of dimensions equal to number of topics which consists of weights corresponding to each topic.

Task 2.1 - Choosing appropriate number of topics

Initial Observation:-
Taking number of topics to be less than 10 will cover all the broad topics but since we are provided with a good amount of data, we can also find subtopics . For instance, instead of looking for a topic related to wars we can hunt for topics related to specific types of wars (border wars,gulf wars,civil war, terror war etc.)

Using coherence scores to find optimum number of topics

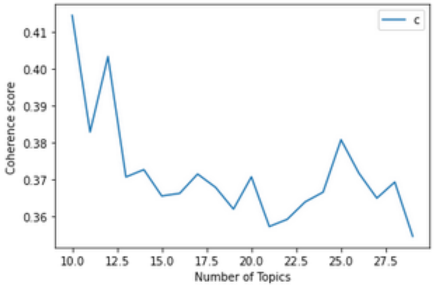
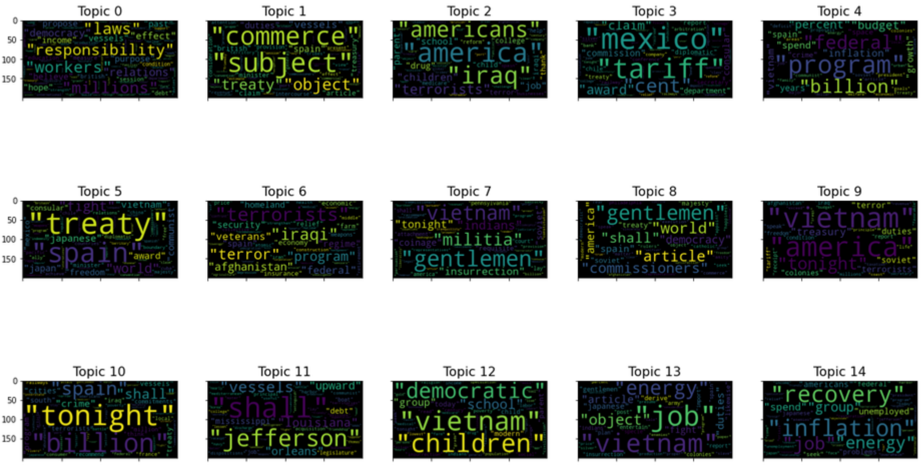


Fig-1:-Coherence scores vs Number of Topics

Going above 30 do not yield any special topics rather it makes the analysis part very complex. It can be inferred from the above plot that after 10 , 25 seems to be a good number of topics since it has the highest coherence score in the range 12-30.

Task 2.2 Annotating Topics

A total of 25 topics were generated:-



Topic 15
Topic 16
Topic 17
Topic 18
Topic 19
Topic 20
Topic 21
Topic 22
Topic 23
Topic 24
0
50
100
150
"bank"
"vietnam"
"tonight"
"loan"
"silver"
"gold"
"specie"
"circulation"
"currency"
"tribes"
"indians"
"object"
"circulation"
"america"
"specie"
"shall"
"savage"
"mississippi"
"river"
"soviet"
"world"
"tariff"
"structure"
"goal"
"america"
"bank"
"minister"
"arrangement"
"decree"
"storage"
"faith"
"interstate"
"interstate"
"japanese"
"problems"
"help"
"weapons"

Fig-2:-WordCloud for LSI Topics

Annotating Randomly sampled ten topics Only 10 words are shown corresponding to each topic , whereas the word clouds are made with a max limit of 100 words so one can refer the word cloud for the corresponding topic to get more insights about the topic.

===

1:-

Topic 1: -0.157*"america" + -0.144*"tonight" + -0.133*"help" + -0.127*"americans" + -0.116*"program" + -0.108*"world" + 0.099*"subject" + 0.095*"commerce" + -0.093*"job" + -0.091*"budget"

This topic on first look seems to have no clear concept but once I found which year's speech has the most proportion of this topic it started making some sense .
Speech of 1824 had most proportion of this topic, it was the time when The Adams–Onís Treaty was negotiated by John Quincy Adams which was related to spain giving up florida to USA.
Thus in the word cloud we could see [spain,treaty,commerce etc.] but other words doesn't seem to resonate with these words. **Thus only a weak annotation could be given to this topic which is Statecraft**

===

2:-

Topic 14: -0.143*"school" + -0.132*"communist" + 0.116*"inflation" + 0.111*"recovery" + -0.109*"soviet" + -0.107*"interstate" + -0.095*"free" + -0.095*"parent" + 0.093*"job" + 0.092*"energy"

This topic seems to cover multiple concepts , because of the words recovery and inflation it may be representing the theme **economic recovery** or **domestic policies** as per school,job.

Year 1934 had the most proportion of this topic , the second wave of economic depression existed from mid 1931 to 1933 and this very well accounts for the fact that why is president using such terms in the speech of 1934.The problem in the early 1930's was that the rate of inflation was negative; i.e., there was deflation instead of inflation.
Why the words communist and soviet ??
Because the **USSR** was the only **communist state** at the time, it had minimal trade contact with the rest of the world. Because of this the **Soviet economy did not take a hit** during

great depression like that of the capitalist countries who's economies were closely interlinked. Thus this topic was in the highest proportion during the great depression.
Annotation– Economic Reforms and post depression policies

==

3:-

Topic 16: 0.186*"mexico" + -0.157*"america" + -0.137*"vietnam" + 0.101*"circulation" + -0.099*"federal" + 0.097*"specie" + 0.092*"currency" + -0.092*"navy" + 0.091*"silver" + -0.087*"budget"

No clear concept as to what various forms of currencies and these countries have in common.This topic was most prevalent in 1842 during the mexican war . **Thus a very weak annotation of war financing can be given.**

==

4:-

Topic 17: -0.146*"minister" + -0.118*"decree" + -0.115*"british" + 0.106*"indians" + 0.104*"object" + -0.102*"article" + -0.095*"commissioners" + -0.088*"french" + 0.088*"tribes" + 0.088*"savage"

No clear concept even after referring to the document having most proportion of it.

==

5:-

Topic 6: 0.301*"iraq" + 0.239*"terrorists" + 0.172*"iraqi" + 0.142*"terror" + -0.134*"japanese" + -0.114*"vietnam" + 0.111*"program" + 0.103*"afghanistan" + -0.091*"speak" + 0.089*"federal"

This clearly captures the theme of war,Thus it can be annotated as **terror war**.
This topic was most prevalent in 2008 when Iraq war was going on (against the terror groups in Iraq).

==

6:-

Topic 24: -0.126*"tonight" + -0.116*"democratic" + -0.112*"farm" + 0.104*"interstate" + -0.093*"vietnam" + 0.084*"problems" + 0.084*"weapons" + -0.081*"economic" + 0.080*"japane + 0.079*"help"

This was most prevalent in 1922 between the first and second war but it doesn't capture any real human concept.

==

7:-

Topic 0: -0.141*"world" + -0.135*"america" + -0.118*"program" + -0.110*"help" + -0.100*"tonight" + -0.098*"americans" + -0.089*"need" + -0.085*"years" + -0.084*"economic"

+ -0.080*"federal"

This topic was also most prominent during 1932 (The Great Depression) but it doesn't have a central theme .

==

8:-

Topic 19: 0.249*"india" + -0.119*"tariff" + 0.103*"savage" + -0.090*"navigation" + -0.089*"commercial" + 0.089*"mississippi" + -0.074*"duties" + -0.072*"colonies" + -0.071*"can
+ 0.071*"shall"

This topic is mostly concerned with the colonies and was most prevalent in the colonial era thus it can be annotated as **colonial policies**.

==

9:-

Topic 21: -0.125*"budget" + -0.124*"tariff" + -0.115*"billion" + 0.106*"bank" + 0.084*"stru
ture" + 0.082*"goal" + 0.081*"america" + -0.080*"price" + -0.080*"democracy" + -0.076*"perce

The words in this topic are components of monetary policies
Thus this topic can be annotated as **monetary policies**.
This topic was found to be in high proportions in the decade 1930-1940 owing to various reforms in monetary policies post depression.

==

10:-

Topic 22: -0.144*"enemy" + -0.134*"article" + 0.131*"minister" + -0.100*"tariff" + -0.096*"commissioners" + 0.096*"decree" + -0.092*"gentlemen" + -0.088*"treaty" + -0.080*"cru
+ -0.079*"vessels"

This topic also doesn't have a central theme i.e it does not capture any real concept

==

3 LDA Topic Modelling

Using the same technique of plotting a coherence value plot , number of topics = 25 seemed to be the appropriate choice.

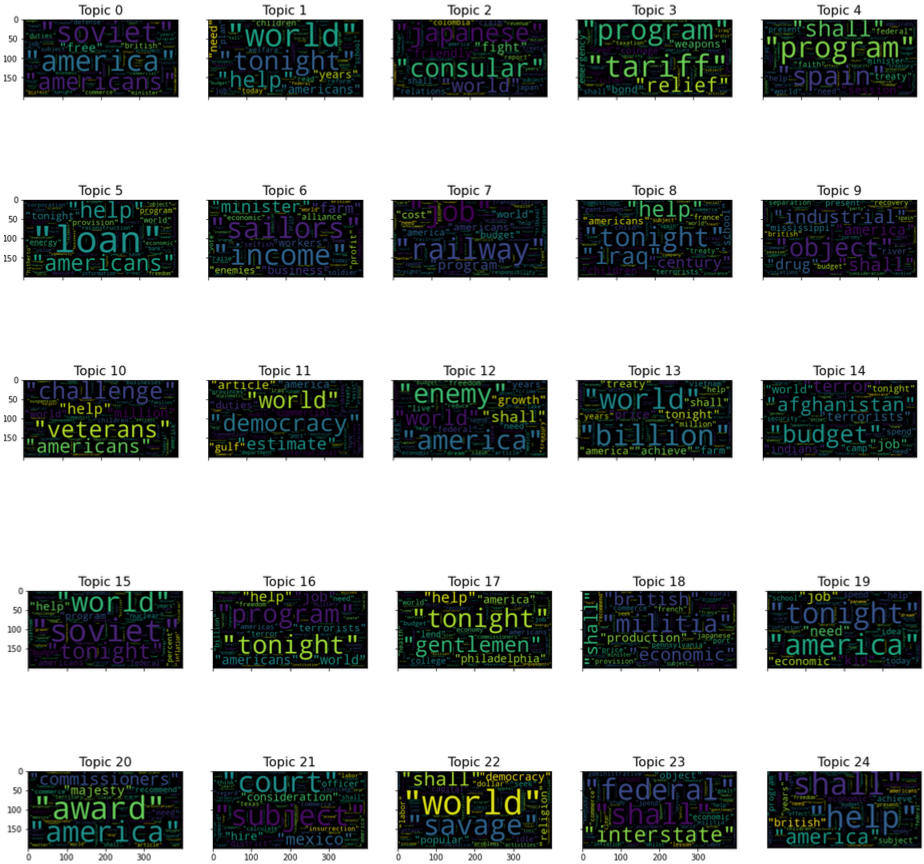


Fig-2:-WordCloud for LDA Topics

Difference between LDA and LSI just by seeing the word clouds

The topics are more clear, the minimum probability parameter helps to make the topics even more centric towards a single theme. So, from first glance they are certainly more appealing and secondly some of the weights per topic or per word were negative in case of LSI whereas since LDA gives positive weights(probability distributuion over topics or words) they are easier to interpret.

3.1 Annotating ten topics

==

1:-

Topic 14: 0.002*"america" + 0.002*"budget" + 0.002*"afghanistan" + 0.002*"terror" + 0.001*"terrorists" + 0.001*"job" + 0.001*"indians" + 0.001*"world" + 0.001*"tonight" + 0.001*"camp"

Annotation:- war against terror (political and economic issues related with it) This topic most frequently occured in 2001-2010 to be more specific 2003 i.e during gulf war(USA vs

Iraq) thus justifying the annotation.

==

2:-

Topic 1: 0.003*"america" + 0.003*"world" + 0.002*"tonight" + 0.002*"help" + 0.002*"americans" + 0.002*"years" + 0.002*"need" + 0.001*"children" + 0.001*"plan" + 0.001*"today" + 0.001*"welfare" + 0.001*"school" + 0.001*"job" + 0.001*"read" + 0.001*"reform"

Annotation:- Social policies Most frequent in 1935 just after second wave of recession thus clearly justifies the annotation.

==

3:-

Topic 21: 0.001*"claim" + 0.001*"subject" + 0.001*"court" + 0.001*"mexico" + 0.001*"consideration" + 0.001*"hire" + 0.001*"officer" + 0.001*"district" + 0.001*"insurrection" + 0.001*"texas" + 0.001*"commerce" + 0.001*"shall" + 0.001*"calculate" + 0.001*"labor" + 0.001*"think"

After seeing the document containing majority proportion of this topic ie in year 1821 it was possible to annotate this .
"In 1821 Mexico won its independence from Spain. The leaders of Mexico and the United States met to talk about American settlers moving into the area of Texas."

Annotation:- Domestic policies after independence

==

4:- Topic 18: 0.002*"world" + 0.002*"militia" + 0.001*"british" + 0.001*"economic" + 0.001*"shall" + 0.001*"production" + 0.001*"pennsylvania" + 0.001*"program" + 0.001*"provision" + 0.001*"repeal" + 0.001*"commerce" + 0.001*"live" + 0.001*"french" + 0.001*"japanes + 0.001*"price"

Annotation:- Industrial production
This topic was most prevalent in the year 1811 which justifies the annotation because "The first major instances of machine breaking took place in 1811 in Nottingham, and the practice soon spread across the English countryside."

==

5:-

Topic 16: 0.002*"america" + 0.002*"program" + 0.002*"tonight" + 0.002*"help" + 0.002*"job" + 0.002*"americans" + 0.002*"world" + 0.002*"terrorists" + 0.001*"terror" + 0.001*"americas" + 0.001*"need" + 0.001*"billion" + 0.001*"federal" + 0.001*"freedom" + 0.001*"military"

Annotation:- War against terror and related economic concerns

==

6:-

Topic 7: 0.002*"federal" + 0.002*"railway" + 0.002*"job" + 0.002*"program" + 0.001*"americans" + 0.001*"world" + 0.001*"cost" + 0.001*"budget" + 0.001*"america" + 0.001*"transportation" + 0.001*"responsibility" + 0.001*"freight" + 0.001*"help" + 0.001*"decisions" +

0.001*"spend"

Annotation:- Infrastructure developmet
Most prevalent topic in 1922 when several policies were being made for infrastructure development.
For example:- 1922 Border Railways act.

==

7:- Topic 22: 0.002*"spain" + 0.002*"world" + 0.002*"savage" + 0.001*"shall" + 0.001*"democracy" + 0.001*"popular" + 0.001*"religion" + 0.001*"capital" + 0.001*"idle" + 0.001*"seek" + 0.001*"dollar" + 0.001*"labor" + 0.001*"problems" + 0.001*"activities" + 0.001*"attack"

no clear concept

==

8:-

Topic 5: 0.002*"america" + 0.002*"loan" + 0.002*"help" + 0.002*"americans" + 0.002*"tonigh + 0.001*"program" + 0.001*"provision" + 0.001*"world" + 0.001*"energy" + 0.001*"corporation" + 0.001*"reconstruction" + 0.001*"economic" + 0.001*"information" + 0.001*"hitherto" + 0.001*"value"

Annotation:- Reforms needed to come out of the viscious cycle of recession.
Most prevalent in 1932 during second wave on the great economic depression.

==

9:-

Topic 9: 0.002*"tonight" + 0.002*"object" + 0.001*"industrial" + 0.001*"shall" + 0.001*"america" + 0.001*"drug" + 0.001*"mississippi" + 0.001*"present" + 0.001*"river" + 0.001*"recovery" + 0.001*"separation" + 0.001*"budget" + 0.001*"british" + 0.001*"violations" + 0.001*"debt"

Annotation:- Colonial policies
Prevalent in colonial era

==

10:-

Topic 10: 0.002*"program" + 0.002*"challenge" + 0.002*"veterans" + 0.002*"americans" + 0.001*"help" + 0.001*"million" + 0.001*"world" + 0.001*"children" + 0.001*"businesses" + 0.001*"america" + 0.001*"suspension" + 0.001*"families" + 0.001*"spain" + 0.001*"federal" + 0.001*"communist"

no clear concept

==

Difference between LDA and LSI

1:- Topics were much clearer for example if we compare the topics of both lda and lsi which tried to capture the theme of terror war. Lsi included words which also belonged to different concept whereas LDA was much clearer , it even included Saddam Hussain for that topic.

2:-In LSI, I had to refer to the original documents to figure out the theme of topic but in most cases of LDA it was straight forward and later original documents justified the annotation.

3:-LSI (also known as Latent Semantic Analysis, LSA) learns latent topics by performing a matrix decomposition (SVD) on the term-document matrix. LDA is a generative probabilistic model, that assumes a Dirichlet prior over the latent topics.

4:- The challenge of SVD is that we are hard to determine the optimal number of dimension. In general, low dimension consume less resource but we may not able to distinguish opposite meaning words while high dimension overcome it but consuming more resource.

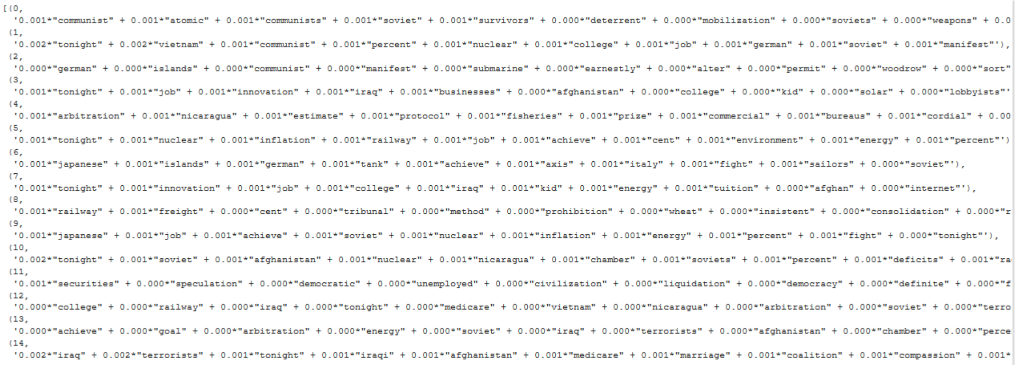
5:- LDA takes a little more time than LSI.

6:- LDA is probabilistic while LSI is deterministic.

4 How topics of speeches have changed over time ?

Goal is to summarize changes in the State of the Union speech in each decade of the 20th and 21st century.
Decade Summarization algorithm :-
Grouped speeches before topic modelling
Formed a dictionary(size=12) where keys represent the decade (1901-1910 to 2011-2020) and corresponding values refer to the speeches of that decade concatenated together.
Apply topic modelling using base algorithm LDA to the tf-idf weighted document vectors generated from the decade dictionary.
The lda model trained on the whole text corpus(1790-2012) cannot be used for extracting topics from the decade dictionary as the tf-idf vectors generated in both cases have different dimensionality owing to different number of tokens in the two dictionaries .
Thus it becomes necessary to train lda model again on this corpus and use that trained model to extract topics from various decades and then analyze how the content of the speeches have changed over decades with respect to those topics.
Note:- In some cases all the 10 speeches are not available thus the remaining speeches represent that decade

Topics= [0,war(border wars)],[1,Foreign Policy and statecraft],[2,Statecraft] ,[3,Foreign policy] ,[4,Navy] ,[5,economic reforms],[6,world war], [7,innovation and democratization of technology],[8,production and transportation], [9,Foreign Policy] , [10,war(gulf)], [11,recession], [12 and 13,diplomatic foreign policy] , [14,terror war]



```
[(0,
  '0.001*"communist" + 0.001*"atomic" + 0.001*"communists" + 0.001*"soviet" + 0.001*"survivors" + 0.000*"deterrent" + 0.000*"mobilization" + 0.000*"soviets" + 0.000*"weapons" + 0.0
 (1,
  '0.002*"tonight" + 0.002*"vietnam" + 0.001*"communist" + 0.001*"percent" + 0.001*"nuclear" + 0.001*"college" + 0.001*"job" + 0.001*"german" + 0.001*"soviet" + 0.001*"manifest"'),
 (2,
  '0.000*"german" + 0.000*"islands" + 0.000*"communist" + 0.000*"manifest" + 0.000*"submarine" + 0.000*"earnestly" + 0.000*"alter" + 0.000*"permit" + 0.000*"woodrow" + 0.000*"sort"
 (3,
  '0.001*"tonight" + 0.001*"job" + 0.001*"innovation" + 0.001*"iraq" + 0.001*"businesses" + 0.000*"afghanistan" + 0.000*"college" + 0.000*"kid" + 0.000*"solar" + 0.000*"lobbyists"'
 (4,
  '0.001*"arbitration" + 0.001*"nicaragua" + 0.001*"estimate" + 0.001*"protocol" + 0.001*"fisheries" + 0.001*"prize" + 0.001*"commercial" + 0.001*"bureaus" + 0.001*"cordial" + 0.00
 (5,
  '0.001*"tonight" + 0.001*"nuclear" + 0.001*"inflation" + 0.001*"railway" + 0.001*"job" + 0.001*"achieve" + 0.001*"cent" + 0.001*"environment" + 0.001*"energy" + 0.001*"percent"')
 (6,
  '0.001*"japanese" + 0.001*"islands" + 0.001*"german" + 0.001*"tank" + 0.001*"achieve" + 0.001*"axis" + 0.001*"italy" + 0.001*"fight" + 0.001*"sailors" + 0.000*"soviet"'),
 (7,
  '0.001*"tonight" + 0.001*"innovation" + 0.001*"job" + 0.001*"college" + 0.001*"iraq" + 0.001*"kid" + 0.001*"energy" + 0.001*"tuition" + 0.000*"afghan" + 0.000*"internet"'),
 (8,
  '0.001*"railway" + 0.001*"freight" + 0.000*"cent" + 0.000*"tribunal" + 0.000*"method" + 0.000*"prohibition" + 0.000*"wheat" + 0.000*"insistent" + 0.000*"consolidation" + 0.000*"r
 (9,
  '0.001*"japanese" + 0.001*"job" + 0.001*"achieve" + 0.001*"soviet" + 0.001*"nuclear" + 0.001*"inflation" + 0.001*"energy" + 0.001*"percent" + 0.001*"fight" + 0.000*"tonight"'),
 (10,
  '0.002*"tonight" + 0.001*"soviet" + 0.001*"afghanistan" + 0.001*"nuclear" + 0.001*"nicaragua" + 0.001*"chamber" + 0.001*"soviets" + 0.001*"percent" + 0.001*"deficits" + 0.001*"ra
 (11,
  '0.001*"securities" + 0.000*"speculation" + 0.000*"democratic" + 0.000*"unemployed" + 0.000*"civilization" + 0.000*"liquidation" + 0.000*"democracy" + 0.000*"definite" + 0.000*"f
 (12,
  '0.000*"college" + 0.000*"railway" + 0.000*"iraq" + 0.000*"tonight" + 0.000*"medicare" + 0.000*"vietnam" + 0.000*"nicaragua" + 0.000*"arbitration" + 0.000*"soviet" + 0.000*"terro
 (13,
  '0.000*"achieve" + 0.000*"goal" + 0.000*"arbitration" + 0.000*"energy" + 0.000*"soviet" + 0.000*"iraq" + 0.000*"terrorists" + 0.000*"afghanistan" + 0.000*"chamber" + 0.000*"perce
 (14,
  '0.002*"iraq" + 0.002*"terrorists" + 0.001*"tonight" + 0.001*"iraqi" + 0.001*"afghanistan" + 0.001*"medicare" + 0.001*"marriage" + 0.001*"coalition" + 0.001*"compassion" + 0.001*
```

Fig-3:-Lda topics on summarrized speeches

4.1 Content of the speech in each decade and their relation with historical events

==

1:- 1901-1910

1901-1910 [(0, 0.0038882014), (1, 0.003888221), (2, 0.0038881984), (3, 0.003888198), **(4, 0.94556516)**, (5, 0.003888207), (6, 0.0038882045), (7, 0.0038881984), (8, 0.0038881989), (9, 0.003888199), (10, 0.0038881984), (11, 0.003888199), (12, 0.003888198), (13, 0.003888198), (14, 0.003888201)]

This decade had many missing speeches due to incomplete dataset but still it gives us an idea of that decade. As we can see all the topics had almost equal proportion or probability of occuring in the speech except one which is **navy (it accounts for 94% of the whole speech corpus).**
This is in resonance with the fact that **US expanded it's naval power in the early 1900's** because By the 1890s, the American economy was increasingly dependent on foreign trade. so to become a major naval power, the United States began to replace its wooden sailing ships with steel vessels powered by coal or oil.
==

2:- 1911-1920

1911-1920 [(0, 0.0022860232), **(1, 0.96799564)**, (2, 0.002286021), (3, 0.0022860204), (4, 0.0022861024), (5, 0.0022860335), (6, 0.0022860223), (7, 0.002286021), (8, 0.0022860216), (9, 0.0022860214), (10, 0.0022860214), (11, 0.002286024), (12, 0.0022860202), (13, 0.002286020
(14, 0.0022860246)]

As we can see all the topics had almost equal proportion or probability of occuring in the speech except one which is **foreign policy** (it accounts for 96% of the whole speech corpus). This was the decade of world war 1 ,to avoid war mongering nature of germany or maybe to make strategic advances , reforms were made in foreign policies but since WW1 happened 1917-1918 ie the later half of the decade the influence of topic war is seen less in the decade speech corpus. ==

3:- 1921-1930

1921-1930 [(0, 0.0023241625), **(1, 0.18449105)**, (2, 0.0023241602), (3, 0.0023241602), (4, 0.047673903), **(5, 0.7399451)**, (6, 0.0023241637), (7, 0.0023241602), (8, 0.002324161), (9, 0.002324161), (10, 0.002324162), (11, 0.002324166), (12, 0.00232416), (13, 0.00232416), (14, 0.0023241665)]

Topics **foreign policy and economic reforms** are the most dominant among all other topics// This was the decade just after World War 1 so after they acquired all the territories it sought, the United States pursued a **policy of isolationism**, refusing to get involved in the wars of Europe.
Also during the later half of this decade the world faced the first part of **The great economic depression** from 1929-1931 owing to which leaders started discussing various economic reforms.

==

4:- 1931-1940

1931-1940 [(0, 0.0020273414), **(1, 0.36459532)**, (2, 0.0020273367), (3, 0.002027337), (4, 0.018936021), (5, 0.033460356), (6, 0.002027354), (7, 0.002027337), (8, 0.0020273372), (9, 0.0020273412), (10, 0.002027339), **(11, 0.56070757)**, (12, 0.0020273365), (13, 0.002027336
(14, 0.0020273423)]

Topics like **Foreign policy and recession** had the most proportion in this decade.
During this decade they faced the **second wave of great economic depression** mid 1931-1933 so now words like **stagflation,unemployment and deflation** became more frequent thus changing the **major topic from economic measures to recession**.

==

5:- 1941-1950

1941-1950 [(0, 0.0021808443), **(1, 0.41927302)**, (2, 0.0021808348), (3, 0.002180835), (4, 0.0021808895), (5, 0.0021808816), **(6, 0.552376)**, (7, 0.0021808352), (8, 0.0021808355), (9, 0.0021808383), (10, 0.0021808394), (11, 0.0021808418), (12, 0.0021808343), (13, 0.002180
(14, 0.0021808497)]

Foreign policy,Statecraft and world war were the major topics.
During this decade (1939–1945) USA was involved in world war which would obviously increase discussions on the topic foreign policy and statecraft .
Here the model's capability to not raise the proportions of other topics like war(border wars),gulf war should be appreciated. Maybe for less number of topics all those words might have resided in same topic.

==

6:- 1951-1960

1951-1960 [**(0, 0.45535958)**, **(1, 0.38492873)**, (2, 0.0018862848), (3, 0.0018862848), (4, 0.0018863792), (5, 0.13707615), (6, 0.0018863463), (7, 0.0018862849), (8, 0.0018862848), (9, 0.0018862926), (10, 0.0018862885), (11, 0.0018862879), (12, 0.0018862843), (13, 0.001886
(14, 0.0018862919)]

Border wars and (foreign policy, statecraft) were the major topics.

During these decade USA was involved in Korean War(North Korea,Soviet Union) and Vietnam War(North Vietnam,Soviet Union) and as we have seen previously and which is kind of obvious too that during a war period topics like (foreign policy and statecraft have high proportions).

==

7:- 1961-1970

1961-1970 [(0, 0.0022712902), **(1, 0.968202)**, (2, 0.0022712809), (3, 0.002271281), (4, 0.0022713202), (5, 0.0022712934), (6, 0.0022712876), (7, 0.0022712813), (8, 0.002271281), (9, 0.0022712834), (10, 0.0022712876), (11, 0.0022712834), (12, 0.0022712806), (13, 0.00227128
(14, 0.0022712864)]

During this decade Vietnam war was going on(Part of the Cold War and Indochina Wars) and the reason why the topic statecraft and foreign policy is almost present in every decade is that **state of the union is a very consistent data , president don't deliver random speeches they have a particular format and sometimes a gradual change is observed owing to circumstances:-war,depression,inflation,assassination,impeachment etc.**

===

8:- 1971-1980

1971-1980 [(0, 0.0022579052), **(1, 0.26155856)**, (2, 0.002257903), (3, 0.002257904), (4, 0.0022580724), **(5, 0.70908844)**, (6, 0.0022579809), (7, 0.0022579047), (8, 0.0022579029), (9, 0.0022579136), (10, 0.002257916), (11, 0.0022579052), (12, 0.0022579029), (13, 0.002257902
(14, 0.0022579124)]

Apart from foreign policy , economic reforms were again present in great proportion this time it was not because of great depression rather it was because of inflation.
The 1970s saw some of the highest rates of inflation in the United States in recent history, with interest rates rising in turn to nearly 20%. Central bank policy, the abandonment of the gold window, Keynesian economic policy, and market psychology all contributed to this decade of high inflation.
Thus there was an urgent need to make some reforms in the current monetary policies.

===

9:- 1981-1990

1981-1990 [(0, 0.0024575714), **(1, 0.4088244)**, (2, 0.0024575633), (3, 0.002457565), (4, 0.0024576245), (5, 0.0024575808), (6, 0.002457567), (7, 0.002457567), (8, 0.0024575635), (9, 0.002457567), **(10, 0.5592271)**, (11, 0.0024575677), (12, 0.0024575626), (13, 0.0024575626), (14, 0.0024576245)]

During this decade **USA was involved in tanker war** which explains why foreign policy and statecraft was present in high proportions.
And again we can see it is able to differentiate between border wars, world wars and gulf wars.**As a result topic capturing the theme gulf war was present in highest proportion.**

==

10:-1991-2000

1991-2000 [(0, 0.0026210668), **(1, 0.96330446)**, (2, 0.0026210637), (3, 0.0026210689), (4, 0.0026211021), (5, 0.0026210828), (6, 0.002621165), (7, 0.0026210817), (8, 0.0026210642), (9, 0.002621076), (10, 0.0026210852), (11, 0.002621068), (12, 0.0026210635), (13, 0.00262106 (14, 0.0026214323)]

During this decade the ongoing gulf war against Iraq and terror threats from terror groups based in Afghanistan forced USA to display political statecraft by imposing various sanctions on trade with them .

==

11:-2001-2010

2001-2010 [(0, 0.00343612), **(1, 0.38918182)**, (2, 0.003436112), (3, 0.0034361181), (4, 0.0034362203), (5, 0.0034361293), (6, 0.0034361258), (7, 0.0034361323), (8, 0.0034361123), (9, 0.0034361163), (10, 0.0034361342), (11, 0.0034361132), (12, 0.0034361116), (13, 0.003436 **(14, 0.5661485)**]

Terror War was the most dominant topic during this decade.
During this decade USA was engaged in terror war against terror groups in Afghanistan ,during this time they invaded afghanistan to take revenge of the attacks on their homeland,Saddam Hussein was killed whose name also appears in the dominant topic's word distribution.

==

4.2 Summarizing changes over all decades

It can be inferred from the content of the speeches of each decade that proportions of topics like:- **Statecraft,world wars,recession,border wars,Naval Force** have declined .
Whereas proportions of topics like **Domestic policy,national security,Foreign policy,terror wars,technology and innovation and production.** have increased.
Whereas some topics like economic reforms have more or less remained same , there were spikes in their proportion during economic depression and inflation in 1970's.

5 LDA on collection of AP wire stories

Datset from associated press news.

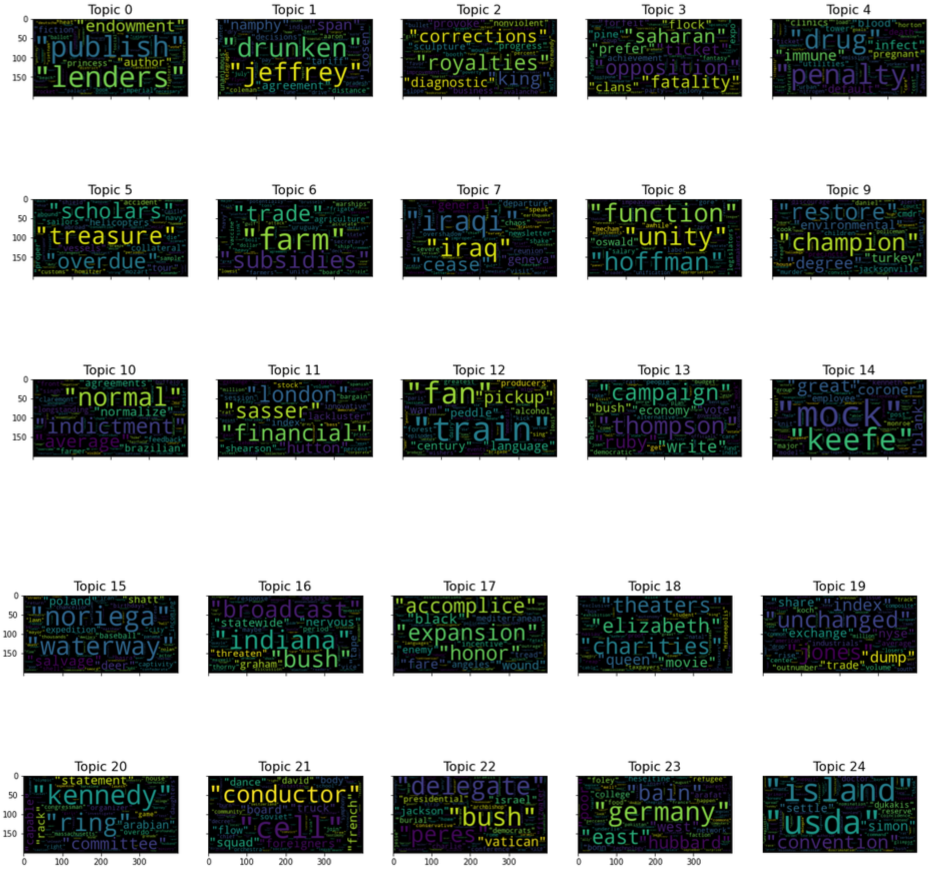


Fig-2:-WordCloud for LDA Topics on Ap stories dataset

1:-Choosing number of topics was easy here since this dataset consist of press stories, this can be mined into many topics in contrast to state of the union where topics are confined and only gradual changes are observed in the topics due to some special circumstances.
2:-Moreover there can be many orthogonal topics here(criminal stories,entertainment stories etc.) in contrast to state of the union where let say a war may lead the country towards economic,social and political change.
3:-In the SOTU dataset there were many tokens which occured in more than 60% of the documents but here there are no such tokens thus this simplified the job of LDA.

5.1 Annotated Topics

==

1:(Topic 54, ’0.014*"bushel" + 0.011*"corn" + 0.011*"cent" + 0.011*"lower" + 0.010*"cents" + 0.009*"subsidies" + 0.008*"soviet" + 0.008*"farmers" + 0.007*"futures" + 0.006*"soybean"’)]

Annotation:- Subsidies over agricultural produce(government policy)

==

2: (Topic 48, ’0.016*"climb" + 0.007*"larceny" + 0.006*"bust" + 0.004*"height" + 0.004*"exit" + 0.004*"taxi" + 0.003*"obligations" + 0.003*"slight" + 0.003*"bearish" + 0.003*"weaker"’)

Annotation:- Description of the thief

==

3: (Topic 88, ’0.021*"bendjedid" + 0.011*"energy" + 0.007*"hearts" + 0.006*"insurgents" + 0.006*"belfast" + 0.006*"wishers" + 0.006*"crowd" + 0.006*"private" + 0.006*"detentions" + 0.005*"episodes"’)

Annotation:-Description of a movie based on insurgent

==

4: (Topic 98, ’0.012*"economic" + 0.008*"percent" + 0.008*"rat" + 0.008*"july" + 0.007*"inflation" + 0.007*"wage" + 0.007*"government" + 0.007*"economy" + 0.006*"german" + 0.006*"price"’),

Annotation:- Catering to inflation(could be someone complaining/ economic reforms read out by government official).

===

5: (51, ’0.019*"hose" + 0.016*"iraqi" + 0.015*"iran" + 0.014*"food" + 0.012*"iraq" + 0.009*"aziz" + 0.009*"cease" + 0.007*"celebration" + 0.007*"talk" + 0.007*"deaths"’)

Annotation:- Something related to detention camps

===

6:

(Topic 7, ’0.019*"firm" + 0.011*"brokerage" + 0.009*"nelson" + 0.009*"contributions" + 0.007*"broker" + 0.007*"sales" + 0.007*"office" + 0.006*"securities" + 0.005*"secret" + 0.005*"paulo"’)

Annotation:- Stock Market

===

7:

(Topic 36, ’0.009*"site" + 0.008*"murder" + 0.008*"carry" + 0.007*"plot" + 0.006*"priests" + 0.006*"weapon" + 0.006*"command" + 0.006*"warmer" + 0.006*"venture" + 0.005*"tougher"

Annotation:- Murder Scene

==

8: (Topic 95, ’0.011*"share" + 0.010*"market" + 0.010*"stock" + 0.009*"index" + 0.008*"party" + 0.007*"point" + 0.007*"exchange" + 0.007*"financial" + 0.006*"germany" + 0.006*"billion"’)

Annotation:- Stock exchanges in Germany(where trading of billions is done)

===

9: (Topic 4, ’0.014*"indictment" + 0.012*"china" + 0.012*"file" + 0.010*"document" + 0.008*"rupture" + 0.007*"houston" + 0.006*"employer" + 0.006*"criminal" + 0.005*"meaningful" + 0.005*"illegally"’)

Annotation:- Legal action

==

10: (Topic 8, ’0.017*"iran" + 0.016*"iraqi" + 0.012*"iraq" + 0.009*"cease" + 0.008*"general" + 0.008*"geneva" + 0.007*"departure" + 0.005*"chaos" + 0.005*"reunion" + 0.005*"newsletter"’)

Annotation:- Political News(related to some global convention)

==

5.2 Differences between the output on these documents vs. the State of the Union documents.

1:- The topics are way more clearer than those generated on state of the union state of the union.
2:- As mentioned above the repeating tokens are less,after preprocessing it was found that there was no token which occured more than 70% in the whole corpus. This level of diversity simplifies tha job of LDA.
3:- The weight distribution for words given a topic is way more better for this dataset as comparet to SOTU, the least weight of the most frequent 20 words in a topic is 0.003 whereas in case of SOTU , 0.003 was almost the highest weight for top 20 words in a given topic.
Thus LDA performs way more confidently on this dataset.
4:- Here if you pick next story distribution over topics for that story may change drastically but in SOTU changes were gradual , a sudden change can be observed there only when change happens from one president to another or if some major national/global event had occured.