# Assignment 4

Abhishek
2017EEB1121

Indian Institute of Technology
Ropar, Punjab
India

## 1 Introduction

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.In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. Topic models are also referred to as probabilistic topic models, which refers to statistical algorithms for discovering the latent semantic structures of an extensive text body. In the age of information, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks.

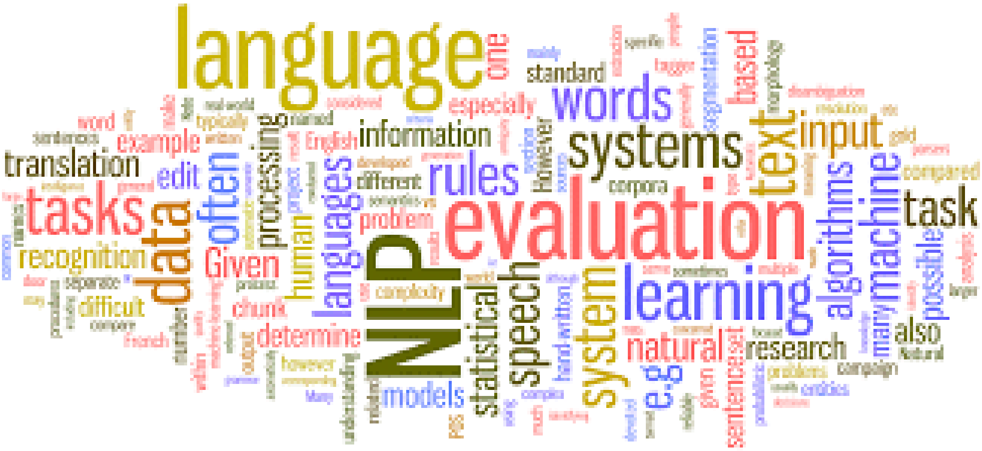


Figure 1: cloud

image copied from : https://www.google.com/imgres?imgurl=https

# 2 About Datasets

## 2.1 state-of-the-union

The State of the Union is an annual address by the President of the United States before a joint session of congress. In it, the President reviews the previous year and lays out his legislative agenda for the coming year.There are two columns: the year of the speech, and the text of the speech.There are 225 rows in the data set

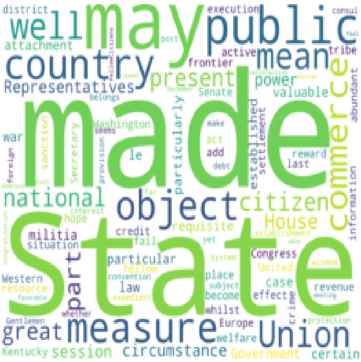


Figure 2: speech-of-union word cloud

## 2.2 AP wire stories

This dats set include 2 column and 2250 rows. Ist column is some id or no. related to AP and 2nd column include text data of stories.

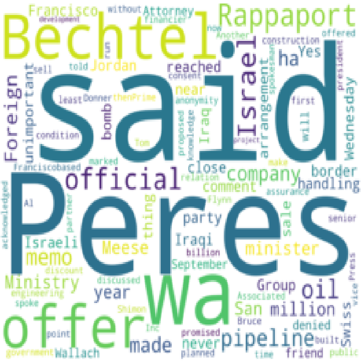


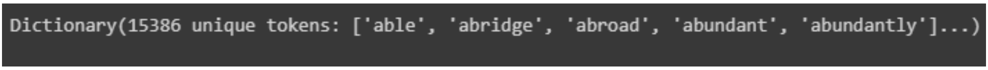
Figure 3: Ap-wire-storie word cloud

# 3 Task 1

Ist i want to talk about library Gensim. Gensim is billed as a Natural Language Processing package that does ‘Topic Modeling for Humans’.Gensim provides algorithms like LDA and LSI and the necessary sophistication to build high-quality topic models. In order to work on text documents, Gensim requires the words (aka tokens) be converted to unique ids. In order to achieve that, Gensim lets you 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.

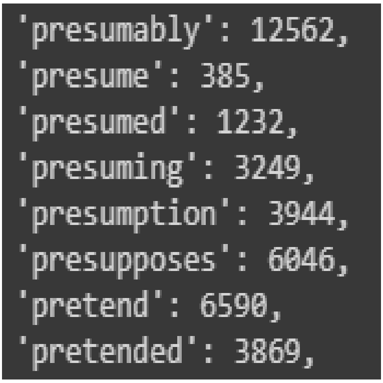
we create a dictionary from a speech of sentences, from a state-of-the-union speeches that contains multiple lines of text and from multiple years.



```
Dictionary(15386 unique tokens: ['able', 'abridge', 'abroad', 'abundant', 'abundantly']...)
```

Figure 4: dictionary

the number with each word is the unique id of each word. we created a dictionary of 15386 words or token.



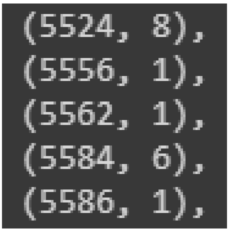
```
'presumably': 12562,
'presume': 385,
'presumed': 1232,
'presuming': 3249,
'presumption': 3944,
'presupposes': 6046,
'pretend': 6590,
'pretended': 3869,
```

Figure 5: dictionary with frequency

The next important object with in order to work in gensim 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.

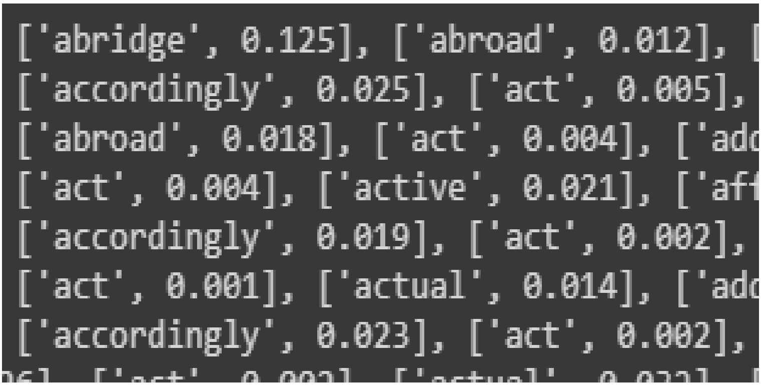
we can see in the figure 6 that 5524 is unique token id and the number(8) with it is his frequency.

Now that we have vectorized our corpus we can begin to transform it using models. We use model as an abstract term referring to a transformation from one document representation to another. In gensim documents are represented as vectors so a model can be thought of as a transformation between two vector spaces. The model learns the details of this transformation during training, when it reads the training Corpus. we use tf-idf model transforms vectors from the bag-of-words representation to a vector space where the frequency counts are weighted according to the relative rarity of each word in the corpus.



```
(5524, 8),
(5556, 1),
(5562, 1),
(5584, 6),
(5586, 1),
```

Figure 6: corpse



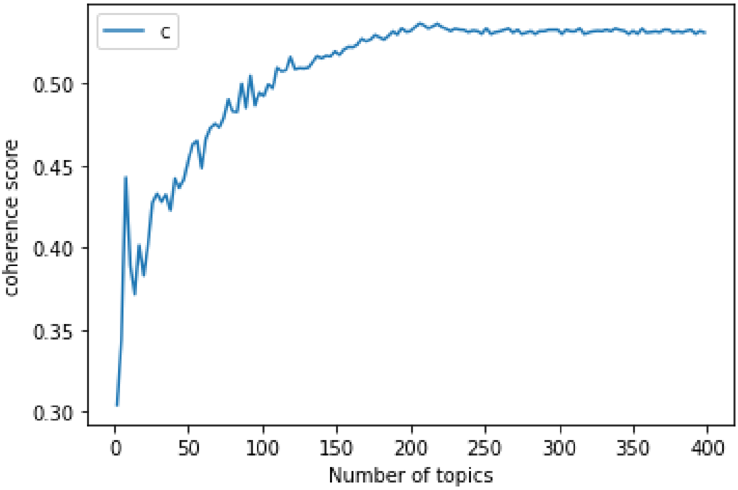
```
['abridge', 0.125], ['abroad', 0.012], [
['accordingly', 0.025], ['act', 0.005],
['abroad', 0.018], ['act', 0.004], ['add
['act', 0.004], ['active', 0.021], ['aff
['accordingly', 0.019], ['act', 0.002],
['act', 0.001], ['actual', 0.014], ['add
['accordingly', 0.023], ['act', 0.002],
```

Figure 7: tf-idf vector

# 4 Task 2

The objective of topic models is to extract the underlying topics from a given collection of text documents. Each document in the text is considered as a combination of topics and each topic is considered as a combination of related words. Topic modeling can be done by algorithms like Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI).In both cases we need to provide the number of topics as input. The topic model, in turn, will provide the topic keywords for each topic and the percentage contribution of topics in each document. The quality of topics is highly dependent on the quality of text processing and the number of topics we provide to the algorithm. To find sufficient number of topics there are two method
1: One way to determine the optimum number of topics is to consider each topic as a cluster and find out the effectiveness of a cluster using the Silhouette coefficient.
2 :Topic coherence measure is a realistic measure for identifying the number of topics.
I used the second one:
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. A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts.I ues Cv measure. Cv measure is based on a sliding window, one-set segmentation of the top words and an indirect

confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.



LSA (Latent Semantic Analysis) also known as LSI (Latent Semantic Index) LSI uses bag of word(BoW) model, which results in a term-document matrix(occurrence of terms in a document). Rows represent terms and columns represent documents. LSI learns latent topics by performing a matrix decomposition on the document-term matrix using Singular value decomposition. LSI is typically used as a dimension reduction or noise reducing technique.

here is the result of 10 random topics by LSI with annotation.the number of topics is 250.we can see in the graph that coherence score is increasing up 250 then it is constant up to 400. that’s why i took number of topics 250

1->(’0.090*"program" + 0.073*"upon" + 0.064*"tonight" + 0.058*"economic" + 0.058*"job" + 0.058*"and" + 0.057*"mexico" + 0.053*"budget" + 0.053*"help" + 0.052*"treaty"’)
topic may be related to economy,budgets and job or may be related with help from mexico and tready.

2->(’0.191*"tonight" + 0.183*"program" + 0.159*"job" + 0.130*"and" + 0.123*"americans" + 0.121*"help" + 0.120*"budget" + 0.098*"billion" + 0.097*"weve" + 0.096*"economic"’)
topic may be related to economy,budgets and job or may be related to help for americans

3->(’-0.182*"democracy" + 0.086*"gold" + 0.075*"acre" + -0.071*"chambers" + 0.071*"tonight + 0.070*"nations" + -0.067*"floating" + -0.058*"goal" + -0.057*"dictator" + 0.057*"group"’)
topic may be related to democracy in nation end of dictator

4->(34, ’0.112*"wool" + -0.112*"california" + 0.083*"chambers" + 0.080*"sheep" + 0.074*"consols" + -0.073*"isthmus" + 0.072*"cable" + 0.072*"fivetwenties" + 0.069*"manufacturer" + -0.069*"panama"’)
topic may be realted to manufactation of wool from sheep

5->(91, ’0.072*"challenge" + -0.071*"saddam" + -0.069*"atomic" + 0.067*"soviet" + -0.062*"hussein" + -0.059*"communist" + -0.051*"california" + -0.047*"delawares" + 0.045*"vietnam" + -0.045*"chance"’)
topic may be related to Saddan Hussain and his atomic weapons

6->(’-0.348*"terrorist" + -0.246*"iraq" + -0.174*"iraqi" + -0.150*"terror" + 0.146*"thats" + -0.126*"al" + -0.117*"regime" + -0.108*"afghanistan" + -0.100*"enemy" + -0.099*"iraqis"’)
topic may be related to terrorist form iraq,afganisthan.

7->(’0.186*"mexico" + 0.166*"texas" + -0.107*"gentlemen" + 0.101*"kansas" + 0.086*"constitution" + 0.086*"slavery" + 0.084*"california" + -0.081*"silver" + 0.080*"oregon" + -0.076*"program"’)
topic may related to the mexico war and getting texas and new mexica and california from mexico

8->(’-0.150*"terrorist" + -0.139*"bank" + -0.126*"iraq" + 0.113*"soviet" + 0.105*"vietnam" + 0.098*"japanese" + 0.091*"communist" + -0.090*"texas" + 0.089*"fighting" + -0.089*"depression"’)
topic may be related to war with vietnam and soviet

9->(’-0.205*"silver" + -0.165*"gold" + -0.123*"gentlemen" + -0.113*"circulation" + -0.112*"coina
+ 0.111*"spain" + 0.111*"mexico" + -0.109*"currency" + -0.097*"note" + -0.095*"bank"’)
topic may related to bank and currency, gold and silver

10->(’-0.246*"vietnam" + 0.129*"thats" + -0.126*"tonight" + 0.098*"japanese" + 0.095*"fighting" + -0.091*"gentlemen" + 0.091*"enemy" + 0.085*"job" + 0.082*"recovery" + 0.082*"hitler"’)
topic may be related to recovery from war with japan and german

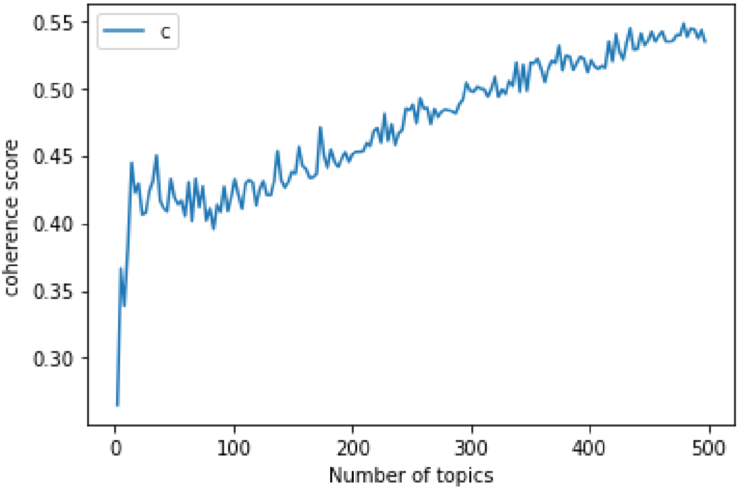
# 5 Task 3

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 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.

so, 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.so I choose 100 as its kind of constant up 100

1->( ’0.000*"iron" + 0.000*"compelled" + 0.000*"soil" + 0.000*"surpassing" + 0.000*"lobbyist" + 0.000*"rounding" + 0.000*"unjust" + 0.000*"underclass" + 0.000*"manned" +



0.000*"strove"’)
topic may be related to minerals

2->(, ’0.000*"rumor" + 0.000*"consistency" + 0.000*"freeman" + 0.000*"dredged" + 0.000*"indissoluble" + 0.000*"explode" + 0.000*"inflicting" + 0.000*"sociological" + 0.000*"worried" + 0.000*"absorbs"’)
topic may related to indissoluble of sociological and rumor

3->( ’0.000*"prima" + 0.000*"profits" + 0.000*"hermitage" + 0.000*"attain" + 0.000*"royalty" + 0.000*"hailed" + 0.000*"collaborated" + 0.000*"averaging" + 0.000*"condolence" + 0.000*"grade"’)
topic may be related to royality in hermitage

4->(’0.000*"government" + 0.000*"the" + 0.000*"states" + 0.000*"united" + 0.000*"country" + 0.000*"may" + 0.000*"year" + 0.000*"great" + 0.000*"law" + 0.000*"i"’)
topic may be related to governments and its law in counrty

5->( ’0.000*"boycotting" + 0.000*"sluggish" + 0.000*"sole" + 0.000*"ascertain" + 0.000*"ideology" + 0.000*"repatriation" + 0.000*"wilful" + 0.000*"encumber" + 0.000*"diffusion" + 0.000*"perturbation"’)
topic may be related to repatriation and diffusion

6->( ’0.000*"costa" + 0.000*"remitted" + 0.000*"diffused" + 0.000*"perpetuation" + 0.000*"trar
form" + 0.000*"aptitude" + 0.000*"excite" + 0.000*"humility" + 0.000*"injunction" + 0.000*"ex
isting"’)
topic may be related to perpetutaion, transfer and humility

7->( ’0.000*"crowded" + 0.000*"imperial" + 0.000*"truth" + 0.000*"cleanest" + 0.000*"per" + 0.000*"spaniards" + 0.000*"universe" + 0.000*"brandon" + 0.000*"vacancy" + 0.000*"already"’) topic may be related to vacancy in brandon

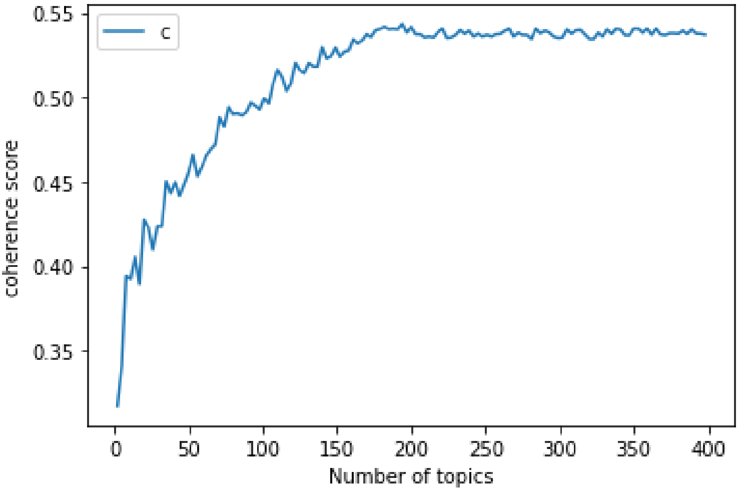
8->( ’0.000*"accrued" + 0.000*"expressing" + 0.000*"corporation" + 0.000*"said" + 0.000*"lung" + 0.000*"noteworthy" + 0.000*"requisition" + 0.000*"peculiar" + 0.000*"navy" + 0.000*"sanguine"’)
tpoic may be related to peculiar requisition and its noteworthy

9->(’0.000*"must" + 0.000*"help" + 0.000*"commended" + 0.000*"expounded" + 0.000*"ships" + 0.000*"consent" + 0.000*"received" + 0.000*"talented" + 0.000*"battle" + 0.000*"right-minded"’)
topic may be related to help in battle, ship

10->( ’0.000*"decreasing" + 0.000*"dropped" + 0.000*"program" + 0.000*"approve" + 0.000*"bud
get" + 0.000*"grievous" + 0.000*"disconnected" + 0.000*"alarming" + 0.000*"job" + 0.000*"farme
topic may be related to job, budgets and framers

# 6 Task 4

to analyze the change in topic of union of speech over the decade first i divide the data into decade speech like 1790-1800 is one speech then i used lsi to find the topic. for lsi used number of topic 200 and we can measure it from the figure below. then i use 5 random topic out of 200 and print then write possible topic for each decade. I again used the same coherence measure. We observed that the coherence become constant after 200 topics so i used 200 for each decade.



Note: i use that only on speech of decade 1790-1800
decade 1
(0, ’0.095*"commissioner" + 0.073*"philadelphia" + 0.073*"article" + 0.073*"th" + 0.073*"france"

+ 0.064*"within" + 0.063*"appointed" + 0.060*"vessel" + 0.058*"great" + 0.058*"boundary"’)

(1, ’0.125*"philadelphia" + 0.102*"commissioner" + 0.095*"france" + 0.088*"th" + 0.085*"article" + 0.085*"determination" + 0.084*"british" + 0.080*"st" + 0.079*"appointed" + 0.079*"minister"’)

(2, ’-0.126*"pennsylvania" + -0.096*"insurrection" + 0.091*"belongs" + 0.091*"convention" + 0.091*"secretary" + 0.091*"whilst" + -0.089*"inspector" + 0.079*"consul" + 0.073*"particularly" + -0.072*"arm"’)

(3, ’-0.120*"review" + 0.094*"want" + -0.090*"insurrection" + -0.080*"emperor" + -0.080*"retal
+ -0.080*"mine" + -0.079*"army" + -0.066*"external" + 0.063*"liberate" + 0.063*"installment"’)

(4, ’0.118*"resolution" + 0.118*"subscription" + -0.104*"secretary" + -0.104*"belongs" + -0.104*"whilst" + -0.104*"convention" + -0.099*"pennsylvania" + -0.093*"let" + -0.073*"consul"
+ -0.070*"crime"’)

topic may be related to commissioners, treaty of paris and france commerce, army and boundary

decade 2

(0, ’0.073*"minister" + 0.061*"river" + 0.059*"decree" + 0.059*"western" + 0.058*"instruction" + 0.057*"title" + 0.056*"million" + 0.055*"number" + 0.055*"acquisition" + 0.055*"among"’)

(1, ’-0.152*"minister" + -0.115*"proposal" + -0.103*"decree" + -0.100*"instruction" + 0.090*"title" + -0.090*"step" + 0.087*"acquisition" + -0.084*"manifested" + -0.084*"embargo" + 0.073*"tax"’)

(2, ’0.296*"gentlemen" + 0.178*"gentleman" + 0.119*"affords" + 0.119*"confidently" + 0.119*"dangerous" + 0.090*"residence" + 0.083*"protection" + 0.077*"forever" + 0.064*"suggest" + 0.064*"true"’)

(3, ’-0.154*"proposal" + 0.116*"gentlemen" + -0.100*"minister" + -0.085*"step" + -0.084*"enga
+ -0.077*"universal" + -0.077*"transfer" + -0.077*"unworthy" + -0.077*"disavowal" + -0.077*"followed"’)

(4, ’0.131*"acquisition" + 0.102*"practicable" + 0.102*"sovereignty" + -0.097*"tax" + -0.086*"large" + 0.078*"gentlemen" + -0.071*"save" + -0.071*"countervailing" + -0.071*"smalle
+ -0.071*"expend"’)

topic may be related to tax, ministers, sovereignty and residensce,proposal of instruction and transfer

decade 3

(0, ’0.206*"spain" + 0.120*"treaty" + 0.096*"enemy" + 0.070*"savage" + 0.070*"lake" + 0.068*"prisoner" + 0.065*"ratification" + 0.064*"party" + 0.063*"strong" + 0.063*"respecting"’)

(1, ’0.323*"spain" + -0.149*"enemy" + 0.148*"treaty" + 0.120*"ratification" + -0.111*"prisoner"
+ 0.103*"ratified" + 0.093*"respecting" + 0.093*"strong" + 0.088*"adventurer" + 0.077*"formed

(2, ’-0.148*"prisoner" + -0.146*"enemy" + -0.116*"spain" + -0.095*"lake" + 0.087*"trade"
+ 0.080*"decree" + -0.074*"movement" + 0.072*"neutral" + -0.068*"victory" + -0.066*"gained"

(3, ’-0.104*"decree" + -0.098*"spain" + -0.087*"neutral" + -0.087*"french" + 0.084*"liberal" + 0.082*"aggregate" + 0.077*"find" + 0.073*"constant" + 0.073*"contracted" + 0.072*"shal

(4, ’-0.119*"blockade" + 0.113*"trade" + -0.100*"academy" + -0.098*"statement" + -0.094*"bea
+ -0.094*"combined" + -0.094*"science" + -0.094*"furnished" + -0.094*"comprehended" +

-0.094*"accomplishment"’)
topic may be related to spain treaty and their enemy, victory, trade acadmey and science

decade 4
(0, ’0.073*"upon" + 0.061*"provinces" + 0.060*"likewise" + 0.060*"enumeration" + 0.060*"presumed" + 0.057*"limit" + 0.055*"increase" + 0.054*"greater" + 0.053*"respective" + 0.052*"officer"’)
(1, ’-0.113*"provinces" + 0.098*"enumeration" + -0.092*"presumed" + -0.089*"likewise" + 0.077*"upon" + -0.076*"raised" + -0.075*"sustaining" + -0.069*"ceded" + -0.064*"greater" + -0.062*"entered"’)
(2, ’-0.114*"whence" + -0.102*"raised" + -0.100*"cortes" + -0.100*"depends" + -0.100*"administe
+ -0.096*"entered" + 0.091*"sustaining" + -0.082*"suffer" + -0.074*"pressure" + -0.073*"game"’)
(3, ’-0.167*"provinces" + 0.100*"fort" + -0.098*"ceded" + -0.096*"allowed" + -0.077*"production
+ 0.076*"contractor" + 0.076*"postage" + 0.076*"attacked" + 0.076*"accompanies" + -0.076*"strict"’)
(4, ’-0.147*"fraud" + -0.128*"election" + -0.092*"circuit" + -0.092*"choice" + -0.092*"task" + 0.089*"provinces" + -0.087*"injurious" + -0.073*"prosecution" + -0.073*"consistently" + -0.069*"evil"’)
topic may be related to enumeration of production and sustaining after war and attack and fraud and election.

decade 5
(0, ’0.156*"chambers" + 0.092*"suspension" + 0.083*"price" + 0.083*"french" + 0.082*"paris" + 0.068*"ministry" + 0.063*"surplus" + 0.058*"note" + 0.058*"ration" + 0.055*"speculation"’)
(1, ’0.311*"chambers" + 0.176*"paris" + 0.151*"ministry" + 0.142*"french" + -0.132*"suspension
+ 0.123*"spanish" + 0.109*"payable" + 0.097*"menace" + -0.088*"price" + 0.077*"explanation"’)
(2, ’-0.126*"suspension" + 0.093*"majesty" + 0.087*"indemnity" + 0.082*"discontent" + 0.078*"neglected" + 0.073*"evinced" + -0.069*"chambers" + 0.069*"thousand" + -0.066*"specula
+ 0.065*"fellow"’)
(3, ’0.140*"suspension" + -0.126*"ration" + 0.115*"writ" + -0.109*"collect" + 0.100*"circuit" + 0.099*"mandamus" + -0.098*"price" + -0.077*"governor" + 0.072*"provinces" + -0.071*"surplus"’)
(4, ’-0.126*"payable" + -0.123*"spanish" + -0.114*"discriminating" + -0.100*"price" + -0.090*"royal" + -0.081*"communicate" + -0.081*"electioneering" + -0.081*"record" + -0.074*"domain" + -0.074*"minimum"’)
topic may be related to spain and france and thie fellow or may be realted to price of goods and other things

decade 6
(0, ’-0.151*"california" + -0.115*"mexico" + -0.085*"annexation" + -0.080*"paredes" + -0.077*"bill" + -0.075*"exchequer" + -0.074*"contribution" + -0.068*"mexican" + -0.062*"grande"
+ -0.061*"enemy"’)
(1, ’0.193*"exchequer" + -0.120*"california" + -0.113*"paredes" + 0.089*"urge" + 0.089*"medium
+ -0.083*"grande" + -0.080*"mexico" + 0.070*"indispensably" + -0.067*"mexican" + -0.065*"contribution"’)
(2, ’-0.344*"california" + 0.128*"paredes" + -0.109*"nicaragua" + 0.094*"annexation" + -

0.091*"isthmus" + -0.082*"study" + -0.082*"frankfort" + -0.082*"denmark" + -0.082*"mining"
+ -0.080*"deficit"')
(3, '-0.218*"exchequer" + 0.072*"indemnification" + 0.072*"fire" + -0.071*"certificate" + -
0.064*"canada" + -0.063*"inquiry" + -0.062*"paredes" + -0.059*"remission" + -0.058*"bill"
+ -0.058*"medium"')
(4, '-0.146*"seldom" + -0.121*"federal" + -0.109*"jesup" + -0.109*"endeavored" + 0.108*"california"
+ -0.077*"prejudice" + -0.075*"independently" + -0.073*"seminoles" + -0.073*"treacher
+ -0.073*"prostitution"')
major topic amy be related to california, canada and mexico or semimoles tribe and denmark
mining

decade 7
(0, '-0.123*"kansas" + -0.116*"bank" + -0.110*"slave" + -0.069*"circulation" + -0.066*"granada
+ -0.060*"election" + -0.059*"acre" + -0.056*"african" + -0.055*"constitution" + -0.054*"estima
(1, '-0.258*"bank" + -0.201*"slave" + -0.163*"circulation" + -0.151*"kansas" + -0.100*"african"
+ -0.088*"had" + -0.084*"loan" + -0.078*"honduras" + -0.077*"note" + -0.076*"silver"')
(2, '0.126*"municipal" + 0.100*"repeal" + 0.099*"pretension" + 0.097*"denmark" + 0.094*"balize"
+ 0.093*"book" + 0.081*"compact" + 0.081*"enjoyment" + 0.080*"paris" + 0.077*"conference"')
(3, '-0.119*"smyrna" + -0.100*"greytown" + -0.096*"koszta" + -0.094*"acre" + -0.088*"magnitu
+ -0.075*"austria" + 0.074*"repeal" + -0.072*"restricted" + -0.072*"inseparable" + -0.072*"punt
(4, '-0.181*"bank" + -0.120*"circulation" + 0.116*"slave" + -0.113*"smyrna" + 0.091*"greytown"
+ -0.091*"koszta" + 0.086*"african" + -0.084*"austria" + -0.079*"kansas" + -0.068*"restri
topic may be related to banks, election and kankas state and african people slavery isue with
austria,france and denmark

decade 8
(0, '-0.098*"emancipation" + -0.090*"currency" + -0.064*"here" + -0.061*"inclusive" + -
0.059*"constitution" + -0.059*"specie" + -0.054*"depreciated" + -0.054*"negroes" + -0.053*"cir
+ -0.051*"seat"')
(1, '-0.185*"emancipation" + 0.103*"currency" + 0.102*"inclusive" + -0.098*"consul" +
0.094*"depreciated" + 0.090*"negroes" + -0.083*"hired" + -0.083*"expediency" + 0.079*"negro"
+ 0.079*"caput"')
(2, '0.143*"emancipation" + -0.119*"here" + -0.094*"seat" + -0.080*"hired" + 0.078*"currency"
+ -0.077*"evacuation" + -0.077*"representation" + 0.074*"depreciated" + 0.073*"caput"
+ 0.069*"circulating"')
(3, '-0.150*"here" + -0.135*"emancipation" + -0.097*"is" + 0.093*"hired" + -0.092*"separation"
+ -0.092*"can" + 0.091*"seat" + -0.079*"forever" + -0.077*"square" + -0.077*"easier"')
(4, '-0.165*"cable" + -0.132*"passenger" + -0.103*"cuba" + 0.101*"hired" + -0.099*"s" + -
0.087*"tariff" + -0.085*"company" + -0.078*"suggest" + -0.070*"endeavor" + -0.070*"funding"'
major topic may be related to currency, bussiness, jobs, companies and their funding

decade 9
(0, '-0.101*"coinage" + -0.097*"acre" + -0.089*"cable" + -0.070*"majesty" + -0.070*"silver"
+ -0.065*"her" + -0.062*"partisan" + -0.060*"currency" + -0.059*"appointment" + -0.058*"total'
(1, '-0.218*"coinage" + 0.147*"majesty" + 0.130*"her" + -0.126*"silver" + 0.106*"acre" + -
0.090*"partisan" + -0.080*"subordinate" + 0.077*"san" + -0.064*"tender" + 0.063*"domingo"')
(2, '0.187*"majesty" + 0.132*"her" + 0.095*"coinage" + -0.086*"acre" + 0.084*"emanci-

pation" + 0.070*"interpretation" + 0.070*"haro" + 0.070*"boston" + 0.070*"counsel" + -0.069*"cable"’)

(3, ’0.149*"acre" + 0.109*"canada" + 0.092*"san" + 0.092*"voyage" + 0.085*"coinage" + 0.079*"bay" + -0.074*"county" + -0.073*"progressing" + -0.073*"namely" + -0.073*"rival"’)

(4, ’-0.120*"domingo" + 0.115*"prostration" + 0.083*"added" + 0.083*"currency" + 0.080*"mediu + -0.077*"rival" + -0.077*"namely" + -0.077*"progressing" + -0.072*"hills" + -0.072*"black"’)

topic may be related to currency and coinage and silver and canada treaty

decade 10

(0, ’0.104*"silver" + 0.087*"price" + 0.066*"loan" + 0.064*"office" + 0.064*"indian" + 0.061*"cost" + 0.060*"day" + 0.060*"increased" + 0.057*"gun" + 0.057*"statute"’)

(1, ’0.219*"loan" + 0.132*"consols" + 0.132*"fivetwenties" + 0.131*"iglesias" + 0.124*"from" + -0.119*"price" + 0.104*"of" + -0.089*"wool" + -0.085*"sheep" + 0.081*"funded"’)

(2, ’0.172*"price" + 0.161*"wool" + 0.157*"sheep" + 0.129*"loan" + -0.105*"appointment" + -0.095*"polygamy" + 0.091*"iglesias" + -0.087*"patronage" + -0.075*"colombia" + -0.074*"sect"’)

(3, ’0.143*"colombia" + 0.123*"ross" + 0.120*"sheep" + 0.120*"price" + 0.116*"wool" + 0.087*"yorktown" + 0.082*"institute" + 0.082*"outward" + 0.082*"contagious" + 0.082*"conforming"’)

(4, ’-0.175*"wool" + -0.168*"sheep" + -0.153*"price" + 0.107*"specification" + 0.089*"expended" + -0.076*"appointment" + -0.075*"patronage" + -0.074*"polygamy" + -0.072*"manufactu + 0.071*"discover"’)

major topic related to sheep, wool and their demmand in market

decade 11 (0, ’0.074*"inch" + 0.069*"election" + 0.064*"volunteer" + 0.064*"manila" + 0.060*"gun" + 0.059*"combination" + 0.059*"nations" + 0.058*"pound" + 0.055*"seed" + 0.054*"belligerency"’)

(1, ’-0.141*"election" + -0.107*"elector" + -0.100*"gerrymander" + -0.093*"think" + -0.088*"wag + -0.088*"mill" + -0.088*"choice" + 0.078*"nations" + -0.075*"canadian" + -0.074*"protective"’)

(2, ’0.133*"nations" + 0.132*"manila" + 0.123*"volunteer" + 0.122*"belligerency" + 0.115*"combination" + -0.113*"inch" + 0.088*"santiago" + -0.085*"seed" + 0.083*"philippines" + 0.081*"protocol"’)

(3, ’0.253*"nations" + 0.232*"belligerency" + -0.143*"combination" + -0.125*"volunteer" + 0.113*"tribes" + -0.108*"manila" + -0.089*"philippines" + 0.086*"cubans" + -0.071*"doubtless" + -0.071*"widow"’)

(4, ’0.135*"cancellation" + 0.119*"meridian" + 0.113*"withdrawal" + -0.113*"nations" + 0.104*"payable" + -0.103*"belligerency" + 0.102*"retirement" + 0.096*"fortyfirst" + -0.086*"inch + -0.077*"seed"’)

major topic may be related to election in nation, seed , philippines

decade 12

(0, ’-0.096*"corporation" + -0.086*"domingo" + -0.086*"santo" + -0.080*"conference" + -0.077*"naturalization" + -0.065*"colored" + -0.061*"man" + -0.060*"legation" + -0.059*"tax" + -0.055*"abuse"’)

(1, ’0.158*"legation" + 0.138*"colombian" + 0.118*"colombia" + 0.107*"hon" + 0.106*"observatory" + 0.096*"granada" + 0.096*"riot" + 0.086*"isthmus" + 0.085*"yamen" + 0.080*"peking

(2, ’0.157*"sounding" + 0.125*"absorption" + -0.104*"colored" + 0.102*"cable" + 0.097*"colombian" + 0.093*"colombia" + -0.083*"conference" + 0.083*"freedelivery" + -0.073*"seal" +

0.073*"reciprocity"’)

(3, ’-0.188*"colombian" + -0.155*"colombia" + -0.143*"isthmus" + -0.137*"naturalization" + -0.134*"domingo" + -0.134*"santo" + -0.130*"granada" + -0.130*"riot" + 0.089*"library" + -0.087*"revolution"’)

(4, ’0.172*"colombian" + -0.150*"santo" + -0.150*"domingo" + 0.147*"colombia" + 0.128*"deforestation" + 0.119*"riot" + 0.119*"granada" + 0.117*"isthmus" + -0.106*"naturalization" + 0.099*"photograph"’)

major topic related to confrences, colombia and tax and santo

decade 13

(0, ’0.096*"wool" + 0.096*"china" + 0.094*"year" + 0.094*"per" + 0.093*"tariff" + 0.092*"foreign" + 0.090*"court" + 0.088*"canal" + 0.086*"american" + 0.085*"cent"’)

(1, ’0.136*"wrong" + -0.126*"china" + -0.096*"wool" + -0.090*"diplomatic" + -0.090*"tariff" + -0.088*"court" + 0.083*"billion" + -0.075*"canal" + -0.073*"commercial" + 0.069*"cable"’)

(2, ’-0.195*"hesitate" + -0.195*"interstate" + -0.141*"election" + -0.139*"eighthour" + -0.136*"train" + -0.125*"administrative" + 0.122*"wrong" + -0.108*"concerted" + -0.105*"legal" + -0.105*"conciliation"’)

(3, ’-0.190*"wrong" + 0.125*"nominee" + 0.094*"presidency" + -0.093*"unrest" + 0.093*"debatable" + -0.092*"interstate" + -0.089*"storage" + -0.079*"austriahungary" + 0.076*"season" + -0.071*"chemical"’)

(4, ’0.223*"wrong" + 0.120*"nominee" + 0.110*"austriahungary" + 0.100*"germany" + -0.094*"cruiser" + 0.090*"presidency" + -0.087*"billion" + -0.083*"hundred" + 0.082*"autocracy" + 0.082*"prussian"’)

major topic related to chian and wool and austriahungary and chemicals and train

decade 14

(0, ’0.093*"railway" + 0.090*"court" + 0.068*"we" + 0.065*"department" + 0.064*"mile" + 0.062*"always" + 0.061*"mail" + 0.059*"whenever" + 0.058*"application" + 0.058*"agreement"’)

(1, ’-0.248*"railway" + -0.124*"republic" + -0.100*"insistent" + -0.086*"paralysis" + -0.084*"readjustment" + -0.080*"disordered" + -0.076*"manager" + -0.074*"strike" + 0.072*"indian" + -0.071*"today"’)

(2, ’0.535*"democracy" + 0.243*"floating" + 0.146*"serviceable" + 0.146*"respectfully" + 0.127*"victory" + 0.119*"faith" + 0.113*"receipt" + 0.102*"sentence" + 0.097*"notes" + 0.097*"fulfilled"’)

(3, ’-0.113*"court" + -0.088*"whenever" + 0.086*"indian" + -0.085*"reform" + -0.084*"propose + -0.083*"nitrogen" + 0.076*"mail" + 0.076*"mile" + -0.070*"cheap" + 0.066*"disordered"’)

(4, ’0.239*"railway" + 0.095*"nitrogen" + 0.088*"democracy" + -0.086*"disordered" + -0.082*"whenever" + -0.080*"russia" + -0.069*"bushel" + -0.065*"popular" + -0.065*"starvation" + -0.065*"nontaxable"’)

major topic may related to railway, courts, decocracy and mail

decade 15

(0, ’0.115*"construction" + 0.104*"democracy" + 0.079*"loan" + 0.077*"depression" + 0.070*"emergency" + 0.066*"recommend" + 0.065*"purchasing" + 0.064*"banks" + 0.064*"attack" + 0.063*"let"’)

(1, ’-0.182*"construction" + -0.171*"loan" + -0.137*"banks" + -0.110*"depression" + -

0.106*"distress" + -0.105*"railway" + -0.102*"cent" + -0.099*"department" + -0.099*"recommend + 0.096*"democracy"’)

(2, ’0.127*"learned" + 0.113*"religion" + -0.104*"exploitation" + -0.091*"project" + -0.088*"desti + -0.086*"restoration" + 0.085*"eighty" + 0.083*"democracy" + -0.077*"broad" + 0.075*"ahead"’)

(3, ’0.181*"democracy" + -0.109*"autocracy" + 0.100*"tenant" + 0.100*"continuously" + 0.097*"interpretation" + -0.090*"exploitation" + -0.085*"save" + -0.082*"popular" + -0.076*"resto + -0.076*"say"’)

(4, ’0.194*"loan" + -0.138*"construction" + 0.128*"banks" + 0.106*"reconstruction" + 0.095*"mobilized" + 0.094*"democracy" + -0.093*"cent" + 0.092*"july" + 0.080*"constitutional" + 0.079*"corporation"’)

topic may be related to corporation, constitutinal and decomcracy and banks

decade 16

(0, ’-0.109*"veteran" + -0.101*"war" + -0.101*"dollar" + -0.098*"price" + -0.097*"housing" + -0.093*"expenditure" + -0.077*"fiscal" + -0.077*"japanese" + -0.074*"welfare" + -0.071*"nurse"

(1, ’-0.132*"japanese" + -0.128*"axis" + -0.117*"hitler" + 0.095*"veteran" + 0.093*"housing" + -0.092*"battle" + -0.088*"fighting" + -0.086*"pacific" + -0.085*"enemy" + -0.085*"conques

(2, ’-0.168*"schedule" + -0.126*"impressive" + -0.107*"went" + -0.104*"dictator" + -0.099*"revol + 0.094*"hitler" + 0.089*"axis" + -0.088*"friendly" + 0.085*"japanese" + -0.084*"putting"’)

(3, ’0.110*"schedule" + -0.109*"usual" + -0.109*"hull" + -0.109*"lawwhich" + -0.102*"soldier" + -0.087*"ally" + -0.085*"profit" + 0.082*"impressive" + -0.082*"stabilization" + 0.079*"dictator"’)

(4, ’-0.165*"dispute" + -0.153*"veteran" + -0.137*"dollar" + -0.100*"commission" + 0.084*"challenge" + -0.081*"reconversion" + -0.080*"employer" + 0.078*"session" + -0.074*"emergency" + 0.071*"benefit"’)

topic may be related to japan and german and war and army and soldier and hitler.

decade 17

(0, ’0.073*"fighting" + 0.071*"missile" + 0.068*"initiative" + 0.056*"federal" + 0.055*"propose" + 0.053*"administration" + 0.053*"concern" + 0.052*"small" + 0.052*"faith" + 0.051*"answered"’)

(1, ’-0.181*"fighting" + -0.168*"b" + -0.146*"plane" + -0.140*"debate" + -0.097*"ruler" + -0.096*"answered" + -0.096*"steel" + -0.084*"notice" + -0.083*"run" + -0.080*"put"’)

(2, ’-0.212*"missile" + -0.160*"ballistic" + -0.103*"extra" + -0.091*"soviets" + 0.083*"fighting" + -0.074*"concentrate" + -0.074*"thrift" + 0.073*"b" + -0.066*"asset" + -0.066*"genuine"’)

(3, ’-0.131*"scene" + -0.096*"regional" + -0.093*"attitude" + -0.088*"link" + -0.088*"fidelity" + -0.088*"document" + -0.088*"should" + -0.088*"deny" + -0.088*"commerce" + -0.088*"roundec

(4, ’-0.229*"answered" + 0.123*"fighting" + -0.122*"ruler" + -0.120*"eight" + -0.107*"seven" + -0.104*"postwar" + -0.093*"sought" + 0.085*"b" + -0.083*"above" + -0.083*"he"’)

major topic related to fight and war and post war situation debate and agrements

decade 18

(0, ’0.209*"tonight" + 0.118*"vietnam" + 0.060*"try" + 0.056*"propose" + 0.055*"session" + 0.055*"eight" + 0.053*"bill" + 0.052*"desire" + 0.051*"lack" + 0.050*"reach"’)

(1, ’0.321*"tonight" + 0.163*"vietnam" + 0.086*"try" + 0.083*"trying" + 0.070*"propose" + -0.068*"lack" + 0.064*"beauty" + -0.062*"instead" + -0.062*"eight" + 0.062*"battle"’)

(2, ’-0.132*"session" + -0.108*"library" + 0.107*"eight" + 0.105*"debt" + 0.098*"management" + 0.088*"collective" + 0.081*"veteran" + -0.077*"cutting" + 0.074*"steel" + -

0.073*"transit"’)

(3, ’-0.130*"session" + -0.089*"library" + 0.083*"hill" + 0.079*"liability" + 0.079*"wind" + 0.076*"beauty" + 0.074*"restless" + 0.069*"recession" + 0.067*"enrich" + -0.064*"bill"’)

(4, ’-0.120*"asset" + -0.094*"sound" + -0.092*"tool" + -0.085*"lack" + 0.072*"wind" + 0.072*"liability" + -0.071*"sufficient" + -0.070*"gold" + -0.070*"domination" + -0.070*"properl

major topic related to vietnam and recession and battle

decade 19

(0, ’0.118*"energy" + 0.101*"weve" + 0.082*"shall" + 0.080*"oil" + 0.077*"door" + 0.066*"propose" + 0.061*"barrel" + 0.060*"look" + 0.060*"investigation" + 0.058*"citizen"’)

(1, ’0.149*"door" + -0.139*"barrel" + -0.137*"energy" + -0.135*"oil" + 0.105*"shall" + -0.090*"supply" + -0.088*"import" + -0.088*"crude" + 0.076*"park" + -0.074*"powerplants"’)

(2, ’0.169*"outline" + 0.166*"lesson" + 0.161*"credibility" + 0.161*"nuclear" + 0.114*"excessive" + 0.113*"february" + 0.108*"discovery" + 0.103*"message" + 0.095*"affair" + 0.095*"higher"’)

(3, ’0.208*"weve" + 0.123*"surtax" + 0.123*"kappel" + 0.102*"poverty" + 0.092*"progressive" + 0.085*"senator" + 0.083*"try" + 0.080*"canal" + 0.077*"trying" + -0.071*"investigation"

(4, ’0.250*"weve" + 0.113*"investigation" + 0.096*"canal" + 0.094*"privacy" + -0.087*"clean" + -0.082*"criminal" + -0.072*"barrel" + -0.071*"big" + -0.071*"gun" + -0.069*"enforcement"’)

topic related to oil,nuclear energy,power plant, discovery of new things and investigation

decade 20

(0, ’0.090*"oil" + 0.088*"space" + 0.088*"administration" + 0.084*"tax" + 0.070*"foundation" + 0.064*"us" + 0.059*"faith" + 0.057*"nations" + 0.057*"salt" + 0.054*"commission"’)

(1, ’-0.185*"oil" + -0.104*"salt" + -0.098*"israel" + -0.096*"nations" + 0.096*"win" + -0.094*"gulf" + -0.092*"persian" + -0.092*"foundation" + -0.083*"import" + 0.079*"constitution"’)

(2, ’-0.111*"oil" + -0.110*"constitution" + -0.086*"sdi" + -0.086*"sandinistas" + -0.076*"pound" + 0.074*"tax" + -0.072*"h" + -0.072*"why" + -0.072*"chair" + -0.072*"delegate"’)

(3, ’-0.125*"space" + -0.086*"oil" + -0.082*"music" + 0.079*"heroism" + 0.075*"local" + 0.074*"constitution" + -0.070*"win" + -0.070*"currency" + 0.068*"foundation" + 0.062*"recession"’)

(4, ’-0.234*"salt" + -0.223*"foundation" + -0.184*"myth" + -0.156*"influence" + -0.138*"antiinf + -0.100*"building" + -0.096*"deterrent" + -0.095*"campaign" + -0.092*"choose" + -0.092*"bel

major topic related to oil and tax and israel and music and salt

decade 21

(0, ’-0.115*"welfare" + -0.098*"covenant" + -0.084*"recommend" + -0.074*"produced" + -0.071*"richard" + -0.069*"insurance" + -0.067*"ought" + -0.066*"standard" + -0.063*"pension" + -0.062*"thing"’)

(1, ’0.162*"where" + 0.135*"panama" + 0.106*"aggression" + 0.082*"iraq" + 0.081*"symbol" + 0.081*"markwell" + 0.081*"anchor" + 0.077*"struggle" + -0.073*"welfare" + 0.071*"peacful"’)

(2, ’-0.182*"recommend" + -0.124*"covenant" + 0.103*"internet" + 0.092*"aggression" + 0.077*"saddam" + -0.075*"revenue" + 0.070*"teaching" + 0.068*"kristen" + 0.066*"iraq" + 0.064*"grade"’)

(3, ’0.179*"recommend" + -0.110*"covenant" + 0.106*"revenue" + -0.094*"where" + 0.087*"pri ority" + 0.082*"cocaine" + 0.082*"ocean" + -0.078*"panama" + 0.073*"involve" + 0.072*"laughter"’)

(4, ’0.183*"aggression" + -0.174*"where" + -0.145*"panama" + 0.128*"iraq" + 0.105*"struggle" + 0.096*"withdrawal" + 0.093*"blessing" + 0.093*"persian" + -0.087*"anchor" + -0.087*"markwell"’)

major topic related to iraq president, persian,insurance and internet

decade 22

(0, ’0.095*"gun" + 0.084*"institution" + 0.083*"surplus" + 0.076*"saddam" + 0.070*"qaeda" + 0.065*"hussein" + 0.060*"st" + 0.060*"teacher" + 0.056*"young" + 0.055*"senior"’)

(1, ’-0.205*"gun" + -0.151*"surplus" + -0.111*"st" + -0.108*"ought" + -0.093*"tobacco" + -0.079*"incentive" + -0.079*"internet" + 0.076*"saddam" + -0.075*"teacher" + -0.075*"district"’)

(2, ’0.159*"alqaida" + 0.153*"islamic" + -0.091*"institution" + -0.090*"qaeda" + 0.083*"muslim" + -0.079*"hopeful" + 0.078*"camp" + 0.073*"strike" + -0.073*"retreat" + 0.071*"taliban"’)

(3, ’0.167*"saddam" + -0.154*"qaeda" + 0.148*"hussein" + 0.127*"inspector" + -0.114*"alqaida" + -0.114*"shia" + -0.103*"oil" + -0.095*"julie" + -0.090*"extremist" + 0.089*"hes"’)

(4, ’0.186*"steven" + 0.149*"josefina" + 0.149*"trillion" + 0.149*"picture" + 0.141*"mayor" + -0.122*"gun" + 0.111*"joe" + 0.111*"teaching" + 0.096*"surplus" + 0.095*"road"’)

major topic related to muslims, terrorist alqaida and taliban, internet

decade 23

(0, ’0.124*"thats" + 0.107*"value" + 0.091*"youre" + 0.087*"recovery" + 0.085*"division" + 0.082*"empower" + 0.074*"theyre" + 0.074*"got" + 0.073*"mission" + 0.072*"race"’)

(1, ’0.292*"empower" + 0.159*"terror" + 0.133*"iraqis" + 0.133*"vital" + 0.121*"liberty" + 0.106*"militia" + 0.106*"lebanon" + 0.105*"extremist" + 0.101*"enemy" + 0.095*"iraqi"’)

(2, ’-0.135*"restart" + -0.105*"recovery" + 0.100*"race" + 0.099*"youre" + 0.091*"brandon" + -0.085*"buy" + -0.081*"pushed" + -0.081*"inherited" + -0.081*"layoff" + 0.080*"were"’)

(3, ’-0.162*"youre" + -0.153*"got" + -0.128*"bet" + -0.111*"mission" + 0.110*"brandon" + -0.104*"manufacturing" + -0.102*"hightech" + -0.101*"unit" + -0.099*"manufacturer" + 0.098*"race"’)

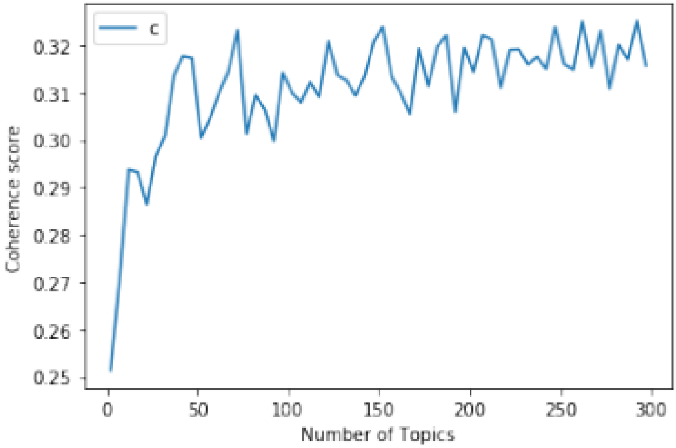
(4, ’0.174*"division" + -0.110*"restart" + 0.094*"value" + 0.087*"interested" + 0.087*"despite" + 0.087*"wouldnt" + 0.087*"setback" + 0.087*"hated" + 0.086*"lobbyist" + -0.077*"buy"’)

major topic related to hightech, manufacturing, liberty of country

I annotate every decade but we can easily see from the topic that they are fighting with in north america for land in 19th century and then they start war with almost every strong country and also took part in both world war but they involved less in world war I but there is huge role of america in world war II. In 20th century they also try to hold on bussiness with top counrty and sell weapons to them. In 21th century there are terrorist attacks and there is fear of life and they are also focusing on increasing the technology of the country. I may relate issue like job and ecomoy to civil rights of the people in country during 19th century. The Watergate scandal related to investgation of wrong thing related to oil and energy this is represent in decade 19. vietnam related issues represent the war with vietnam in 20th century.

# 7 Task 5

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

1->('0.026*"corn" + 0.017*"farmer" + 0.016*"the" + 0.014*"hasselbring" + 0.014*"acre" + 0.013*"his" + 0.013*"yield" + 0.010*"crop" + 0.009*"contest" + 0.008*"bushel" + 0.007*"you
+ 0.007*"hasselbrings" + 0.006*"farmers" + 0.006*"harvest" + 0.005*"campen" + 0.005*"gutwe
+ 0.005*"illinois" + 0.005*"per" + 0.005*"herd" + 0.005*"growing"')
topic may be related to farmers and fields

2->('0.016*"ton" + 0.013*"sundstrand" + 0.011*"accident" + 0.011*"the" + 0.011*"filing" + 0.008*"india" + 0.007*"boat" + 0.007*"sundstrands" + 0.006*"canal" + 0.006*"news" + 0.005*"of" + 0.005*"delhi" + 0.005*"drowned" + 0.005*"pradesh" + 0.005*"andhra" + 0.004*"contract" + 0.004*"uni" + 0.004*"overcharging" + 0.004*"district" + 0.003*"agency"')
topic may be realted to india and its beauty

3->('0.014*"inheritance" + 0.009*"city" + 0.009*"new" + 0.009*"york" + 0.007*"the" + 0.007*"william" + 0.006*"john" + 0.006*"calif" + 0.005*"macmillan" + 0.005*"estate" + 0.005*"inc" + 0.005*"real" + 0.004*"a" + 0.004*"to" + 0.004*"pa" + 0.004*"publishing" + 0.004*"jr" + 0.004*"and" + 0.004*"cargill" + 0.004*"oil"')
topic may be related to real-estate in newyork city

4->('0.000*"lawmakers" + 0.000*"accepts" + 0.000*"fax" + 0.000*"underwood" + 0.000*"appropriate" + 0.000*"undisputed" + 0.000*"irene" + 0.000*"walid" + 0.000*"manage" + 0.000*"brewery" + 0.000*"restructured" + 0.000*"threejudge" + 0.000*"burnout" + 0.000*"replanted" + 0.000*"arresting" + 0.000*"autumn" + 0.000*"watergate" + 0.000*"shrove" + 0.000*"koppers" + 0.000*"counter"')
topic may be related to law and order

5->(’0.038*"the" + 0.028*"to" + 0.023*"document" + 0.015*"case" + 0.015*"north" + 0.015*"a" + 0.014*"of" + 0.013*"gesell" + 0.012*"trial" + 0.011*"that" + 0.011*"judge" + 0.011*"walsh" + 0.010*"butterfly" + 0.007*"secret" + 0.007*"specie" + 0.006*"want" + 0.006*"said" + 0.006*"he" + 0.006*"classified" + 0.006*"government"’)
topic may be related to government and court

6->(’0.000*"shouldnt" + 0.000*"probing" + 0.000*"cyclone" + 0.000*"spurning" + 0.000*"andersons" + 0.000*"wise" + 0.000*"closed" + 0.000*"disappearing" + 0.000*"unstable" + 0.000*"scowcrofts" + 0.000*"relieved" + 0.000*"hills" + 0.000*"conway" + 0.000*"reestablishing" + 0.000*"accommodate" + 0.000*"sidewalk" + 0.000*"rotunda" + 0.000*"sunglass" + 0.000*"tally" + 0.000*"cough"’)
topic may be related to cyclone and its effect

7->(’0.000*"maildrop" + 0.000*"hecht" + 0.000*"ogara" + 0.000*"funeral" + 0.000*"take" + 0.000*"stinson" + 0.000*"refinery" + 0.000*"judeochristian" + 0.000*"vivid" + 0.000*"duvalier" + 0.000*"mutalibov" + 0.000*"amhoist" + 0.000*"mx" + 0.000*"dove" + 0.000*"insures" + 0.000*"armenias" + 0.000*"palermo" + 0.000*"matthews" + 0.000*"geologist" + 0.000*"arbitrator"’)
topic may be related to funeral and geology

8->(90, ’0.025*"the" + 0.023*"percent" + 0.020*"below" + 0.020*"july" + 0.019*"normal" + 0.014*"gallon" + 0.013*"per" + 0.012*"billion" + 0.012*"day" + 0.011*"and" + 0.009*"averaged" + 0.009*"flow" + 0.009*"river" + 0.008*"record" + 0.006*"mississippi" + 0.006*"missouri" + 0.006*"lowwater" + 0.006*"ohio" + 0.005*"period" + 0.005*"survey"’)
topic related to flow of mississippi river in america

9->(19, ’0.030*"the" + 0.029*"and" + 0.021*"a" + 0.020*"sale" + 0.020*"store" + 0.019*"of" + 0.016*"to" + 0.014*"in" + 0.014*"dress" + 0.009*"are" + 0.009*"retailer" + 0.009*"percent" + 0.008*"be" + 0.007*"for" + 0.007*"is" + 0.006*"say" + 0.006*"have" + 0.005*"said" + 0.005*"than" + 0.005*"leonard"’)
topic may related to sale in a cloth shop

10->(5, ’0.015*"patient" + 0.014*"disc" + 0.009*"the" + 0.007*"study" + 0.007*"should" + 0.006*"surgery" + 0.006*"saal" + 0.006*"operate" + 0.006*"be" + 0.005*"root" + 0.005*"of" + 0.004*"whether" + 0.004*"fruehauf" + 0.004*"diagnostic" + 0.004*"and" + 0.004*"surgeon" + 0.003*"exam" + 0.003*"combs" + 0.003*"terex" + 0.003*"unnecessary"’)
topic may be related to patients and diagnostic surgery

LDA with Ap wire stories look very good to me. it is easy to find their topic. this is may happen beacuse it ha salmost 20000 token whereas state of union data has only 15386 token id’s. becaus eof more token id’s it is easy to lda work on it. i don’t check the result on other iteration but there is good chance of getting better more then 100 iteration.

I also compared the result with the previous lda result file given in assignment. most of the words are similar

# 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. 1. LDA give better result with large number of token id's than LSI
2. LSI is must faster than LDA because LDA is unsupervised lerning which based on iterations.
3. If we increase the iteration then LDA may give better result but it consume more time to computethe result.
4. for small datasets LSI give almost same result of LDA so i used LDA in task 4
5. From the graph of Coherence score you can easily predict that score is constant after certain no. of topics in lsi but it is increasing in LDA continuously