Assignment 3

Preetesh Verma
2018EEB1171

Abstract

This document demonstrates the various observations made from the fourth lab assignment given in the course CS521 i.e. Fundamentals of Data Science.The assignment was divided into several parts with the first part focusing on the importing and loading of the dataset into the Gensim Package which is extensively used for NLP(Natural Language Processing). The dataset used for this task is "State of the Union Speech". The second part of the assignment was to use perform Topic Modelling on the given dataset using the two algorithms namely LSI and LDA and compare their performances on the dataset.We also have to annotate randomly selected topics.In the next part of the assignment we have develop an algorithm to extract the themes of the topics with temporal frequency in decades and try to relate it various historical events.The last part focuses on the usage of the LDA algorithm to perform Topic Modelling on a different dataset and compare its performance on the two datasets.

At each level of the assignment proper observations were to be made from the result and conclusions were to be drawn thereafter.The task also involved computing the appropriate values of hyper parameters and check the differences that occur because of that.

Therefore, the main tasks and learning's of the assignment focuses on Topic Modelling using LDA and LSI.

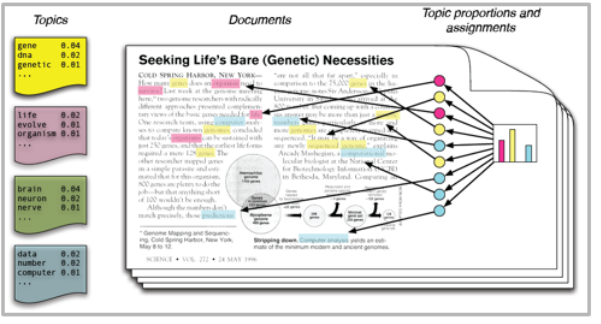
1 About Topic Modelling

Text mining, also known as text analysis, is the process of transforming unstructured text data into meaningful and actionable information. Text mining utilizes different AI technologies to automatically process data and generate valuable insights, enabling companies to make data-driven decisions. One such technique in the field of text mining is **Topic Modelling**. As the name suggests, it is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus. Thus, assisting better decision making.

Topic modelling is unsupervised machine learning method to understand a corpus that is to big to read.This is known as 'unsupervised' machine learning because it does not require a predefined list of tags or training data that's been previously classified by humans.It is a suite of algorithm that aim to discover and annotate large archives documents with thematic information.

Topic modelling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them.They do not require any prior annotation.Topics are not pre-defined but mined and are a distribution over words.It is a cluster of words representative of informative contained within set of documents. Goal of topic

modelling is to automatically discover the topics from a collection of documents.**Topics are therefore latent!**The central computation problem for the topic modelling is to use the observed documents to infer hidden topic structure.



Topic Models are very useful for the purpose for document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. In topic Modeling process we divide a corpus of documents in two:

1.A list of the topics covered by the documents in the corpus.

2.Several sets of documents from the corpus grouped by the topics they cover.

The underlying assumption is that every document comprises a statistical mixture of topics, i.e. a statistical distribution of topics that can be obtained by "adding up" all of the distributions for all the topics covered. What topic modeling methods do is try to figure out which topics are present in the documents of the corpus and how strong that presence is.Thus we try to compute **Topic Assignment and Topic Proportion** i.e. assigning words to topics and topics in documents respectively.

2 BOW and Tf-IDF

The most simple form of representation of words in the form of vectors is Bag of Words where each word in the vocab is treated as a feature and each row as a sentence.An extension of the same model where instead of single words a combination of several words is considered is called as Bag of N-grams where N is the number of words considered.There is a drawback in the BOW model that it just considers the absolute frequency of the words without taking into consideration the number of documents in which it appears.Tf-IDF considers this issue and tries to normalize it.Tf-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. Here we multiply two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

3 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is one of the most frequent topic modeling methods analysts make use of. It is based on what is known as the distributional hypothesis which states that the semantics of words can be grasped by looking at the contexts the words appear in. In other words, under this hypothesis, the semantics of two words will be similar if they tend to occur in similar contexts.

That said, LSI computes how frequently words occur in the documents – and the whole corpus – and assumes that similar documents will contain approximately the same distribution of word frequencies for certain words. In this case, syntactic information (e.g. word order) and semantic information (e.g. the multiplicity of meanings of a given word) are ignored and each document is treated as a bag of words. The input given generally to the LSI algorithm is a Document-Term Matrix which it decomposes into three sub matrices using **SVD(Singular Value Decomposition) (U*S*V).**
The U matrix is known as the Document-topic matrix which would be the feature matrix we are looking for and the V matrix is known as the Term-topic matrix which helps us in looking at potential topics in the document.

4 Latent Dirichlet Allocation

This is a generative probabilistic model which works with the underlying assumption that each document consists of a combination of several topics and each word could be assigned to a specific topic.Latent Dirichlet Allocation (LDA) and LSI are based on the same underlying assumptions: the distributional hypothesis, (i.e. similar topics make use of similar words).

The main difference between LSI and LDA is that LDA assumes that the distribution of topics in a document and the distribution of words in topics are Dirichlet distributions. LSI does not assume any distribution and therefore, leads to more opaque vector representations of topics and documents.

In the above mentioned algorithm we first initialize all the necessary parameters.Then for each document we randomly initialize each word to one of the topics(number of topics is a hyper-parameter) then we iterate the following procedure several times:
For each document D:

For each word W in document:

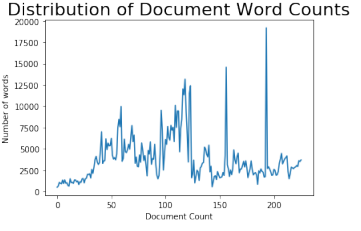
For each topic t:

compute P(T|D) i.e. topic proportion and P(W|T) i.e. topic assignment

Reassign the word W to the topic T with probability P(T|D) * P(W|T) (considering all other words and their topic assignments).

5 State Of the Union Dataset

For the first part of the assignment the dataset used is the State of the Union Dataset which compromises of the speeches given by the US Presidents from 1790 to 2012 i.e. from George Washington to Barack Obama and has a total of 226 speeches which can be treated as separate documents.The State of the Union Speech is a tradition in the USA where the president discusses about several topics during his speech. On exploring the dataset we can clearly see that the number of words per speech is relatively similar with a few spikes occurring as well.



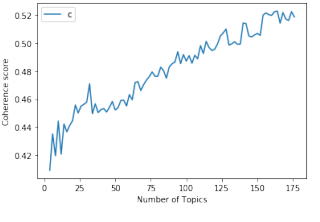
Tokenization

A major step in text pre-processing is to tokenize the documents into sentences and words which then could be easily converted to BOW and Tf-IDF distributions.Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.For the process of tokenization I have used NLTK library and in the procedure I have as well removed all the stop-words, punctuation,special characters and converted the speech text into lower case. Thereafter I have used the Gensim Package to load the tokenized text and convert it into BOW and Tf-IDF Document-Term matrices.

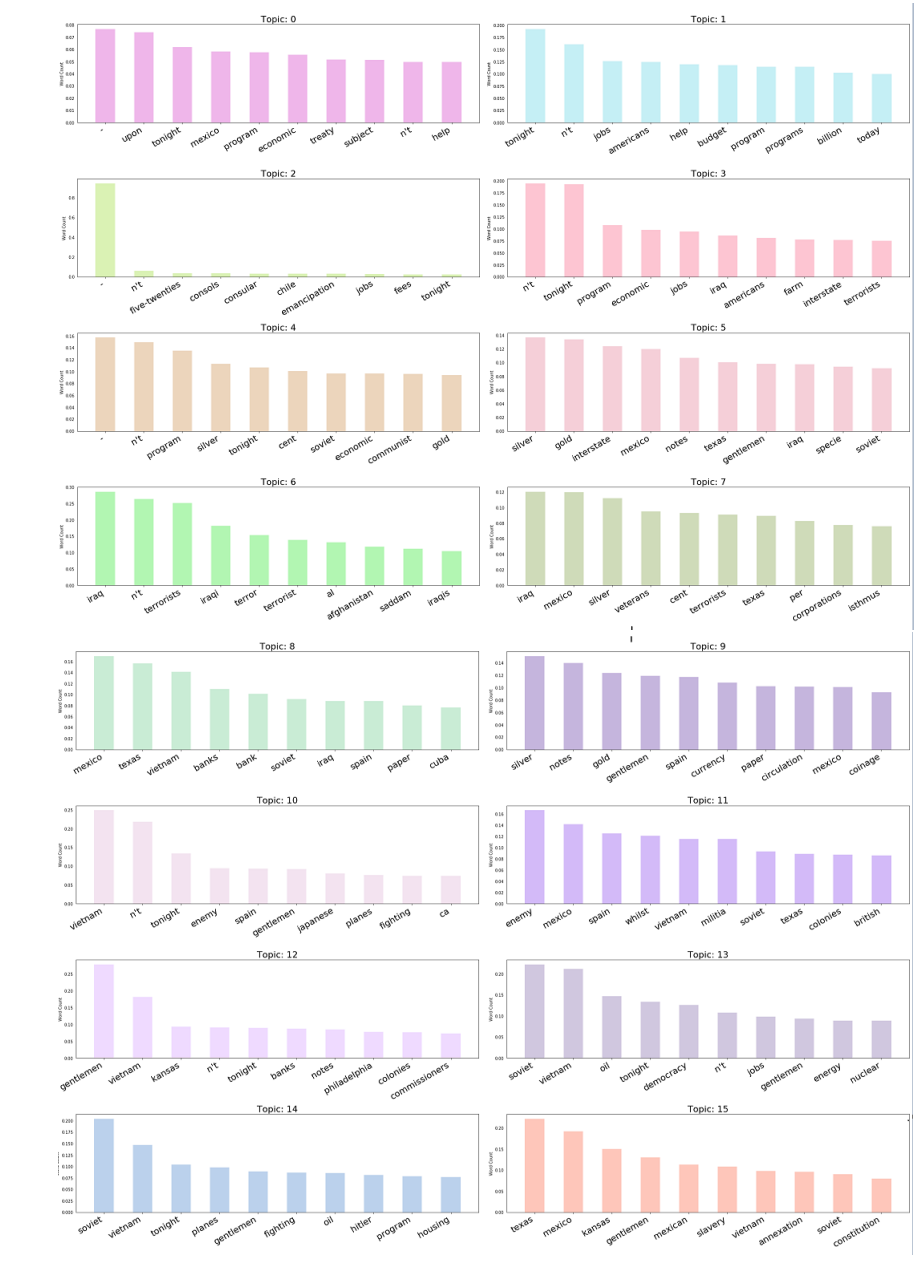
LSI Model After performing the tokenization the next step undertaken by us was to perform LSI where the number of topics were varied from 2 to 180 and coherence values for each of the number of topics was computed to find out the most appropriate number of topics for the dataset.

Topic Coherence– 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.Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.

The coherence parameter used by me is 'c-v'. C-v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point-wise mutual information (NPMI) and the cosine similarity.Here is the result of the same for the various number of topics:



With the coherence score seems to keep increasing with the number of topics, it may make better sense to pick the model that gave the highest CV before flattening out or a major drop.The coherence value score seemed to increasing steadily from 20 and flattening out close to 40 with a significant rise coming at 30 followed by a dip.Thus the number of topics selected was 30 based on the coherence values.The concepts captured by the algorithm have resemblance to human themes but they need to be carefully examined to derive them.



The above are the first 16 topics from the total showing the first 10 words for each topic

and their co-efficient.

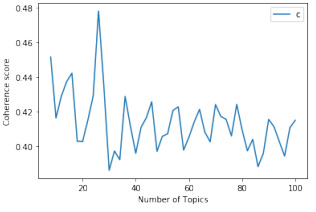
Following are the annotated topics.These are randomly selected 10 topics.

Topic 0 **Mexican Treaty and Economic Reforms**	'0.077*"-" + 0.074*"upon" + 0.062*"tonight" + 0.058*"mexico" + ' '0.058*"program" + 0.056*"economic" + 0.052*"treaty" + 0.051*"subject" + ' '0.050*"n\'t" + 0.050*"help" + 0.049*"budget" + 0.049*"cent" + ' '0.049*"treasury" + 0.048*"americans" + 0.048*"spain" + 0.047*"silver" + ' '0.047*"department" + 0.047*"duties" + 0.047*"programs" + 0.046*"territory"'
Topic 1 **Employment and Economic Reforms** (Budget)	'0.192*"tonight" + 0.161*"n\'t" + 0.126*"jobs" + 0.124*"americans" + ' '0.119*"help" + 0.118*"budget" + 0.114*"program" + 0.114*"programs" + ' '0.102*"billion" + 0.099*"today" + 0.099*"economic" + 0.093*"percent" + ' '0.084*"america" + 0.082*"soviet" + 0.081*"spending" + 0.075*"nuclear" + ' '0.072*"let" + 0.069*"--" + 0.067*"million" + 0.066*"children"'
Topic 3 **Iraq war and Terrorism**	'-0.195*"n\'t" + -0.193*"tonight" + 0.108*"program" + 0.098*"economic" + ' '-0.095*"jobs" + -0.086*"iraq" + -0.081*"americans" + 0.078*"farm" + ' '0.077*"interstate" + -0.075*"terrorists" + 0.075*"veterans" + 0.074*"cent" ' '+ 0.072*"industrial" + -0.064*"parents" + 0.062*"conference" + 0.061*"per" ; '+ 0.060*"production" + -0.058*"america" + -0.057*"ca" +-0.056*"medicare"'
Topic4 **Nuclear weapons and Cold War**	'0.157*"-" + -0.149*"n\'t" + 0.135*"program" + -0.113*"silver" + ' '-0.107*"tonight" + -0.101*"cent" + 0.097*"soviet" + 0.097*"economic" + ' '0.096*"communist" + -0.094*"gold" + 0.089*"programs" + -0.078*"per" + ' '-0.073*"iraq" + 0.073*"atomic" + 0.066*"billion" + -0.065*"terrorists" + ' '-0.062*"interstate" + 0.061*"militia" + 0.058*"gentlemen" + -0.056*"bonds"'
Topic 5 **Texas Revolutiona and Mexican relations**	'0.137*"silver" + 0.137*"gold" + -0.123*"interstate" + 0.120*"mexico" + ' '0.107*"notes" + 0.101*"texas" + -0.098*"gentlemen" + -0.098*"iraq" + ' '0.095*"specie" + 0.092*"soviet" + 0.090*"programs" + 0.087*"currency" + ' '-0.085*"terrorists" + 0.083*"paper" + 0.081*"circulation" + 0.080*"vietnam" ' '+ -0.077*"militia" + 0.077*"billion" + -0.072*"corporations" + ' '0.069*"coin"'
Topic 6 **Terrorism in Iraq and Afghanistan(Saddam Hussain and al- queda)**	'-0.288*"iraq" + 0.263*"n\'t" + -0.252*"terrorists" + -0.182*"iraqi" + ' '-0.153*"terror" + -0.139*"terrorist" + -0.133*"al" + -0.120*"afghanistan" + ' '-0.110*"saddam" + -0.105*"iraqis" + -0.093*"hussein" + -0.083*"qaeda" + ' '0.081*"jobs" + -0.080*"weapons" + -0.079*"regime" + -0.078*"coalition" + ' '-0.076*"homeland" + -0.075*"enemy" + -0.074*"enemies" + 0.074*"ca"'
Topic 7 **American History of Wars**	'-0.120*"iraq" + 0.119*"mexico" + -0.112*"silver" + -0.093*"veterans" + ' '-0.092*"terrorists" + -0.092*"cent" + 0.091*"texas" + -0.081*"per" + ' '0.077*"corporations" + 0.075*"isthmus" + 0.073*"japanese" + ' '0.073*"interstate" + -0.073*"gold" + 0.072*"man" + 0.072*"--" + ' '0.072*"fighting" + -0.071*"iraqi" + 0.070*"kansas" + 0.069*"vietnam" + ' '0.068*"forest"'
Topic 8 **Vietnam War and Economic Tensions**	'-0.172*"mexico" + -0.158*"texas" + 0.143*"vietnam" + -0.114*"banks" + ' '-0.101*"bank" + 0.089*"spain" + 0.088*"soviet" + -0.085*"iraq" + ' '-0.082*"paper" + 0.077*"cuba" + -0.074*"mexican" + 0.071*"silver" + ' '0.071*"gold" + -0.070*"veterans" + -0.069*"n\'t" + -0.069*"specie" + ' '-0.068*"oregon" + -0.064*"constitution" + -0.063*"terrorists" + ' '-0.060*"california"'
Topic 9 **Great Depression**	'-0.149*"silver" + -0.138*"notes" + -0.124*"gold" + 0.117*"spain" + ' '-0.116*"gentlemen" + -0.108*"currency" + -0.103*"paper" + ' '-0.102*"circulation" + 0.097*"mexico" + -0.089*"banks" + -0.089*"coinage" + ' '0.085*"colonies" + -0.080*"specie" + -0.076*"militia" + 0.073*"texas" + ' '-0.067*"bank" + 0.064*"canal" + 0.063*"panama" + 0.063*"cuba" + ' '-0.061*"democracy"'
Topic 10 **World War II**	'0.244*"vietnam" + -0.220*"n\'t" + 0.140*"tonight" + -0.095*"spain" + ' '-0.091*"enemy" + 0.083*"gentlemen" + -0.082*"planes" + -0.081*"japanese" + ' '-0.078*"fighting" + -0.075*"hitler" + -0.073*"production" + ' '-0.073*"british" + -0.073*"ca" + 0.073*"interstate" + -0.072*"jobs" + ' '0.071*"programs" + 0.068*"billion" + -0.063*"businesses" + -0.060*"savages" ' '+ -0.059*"atomic"'

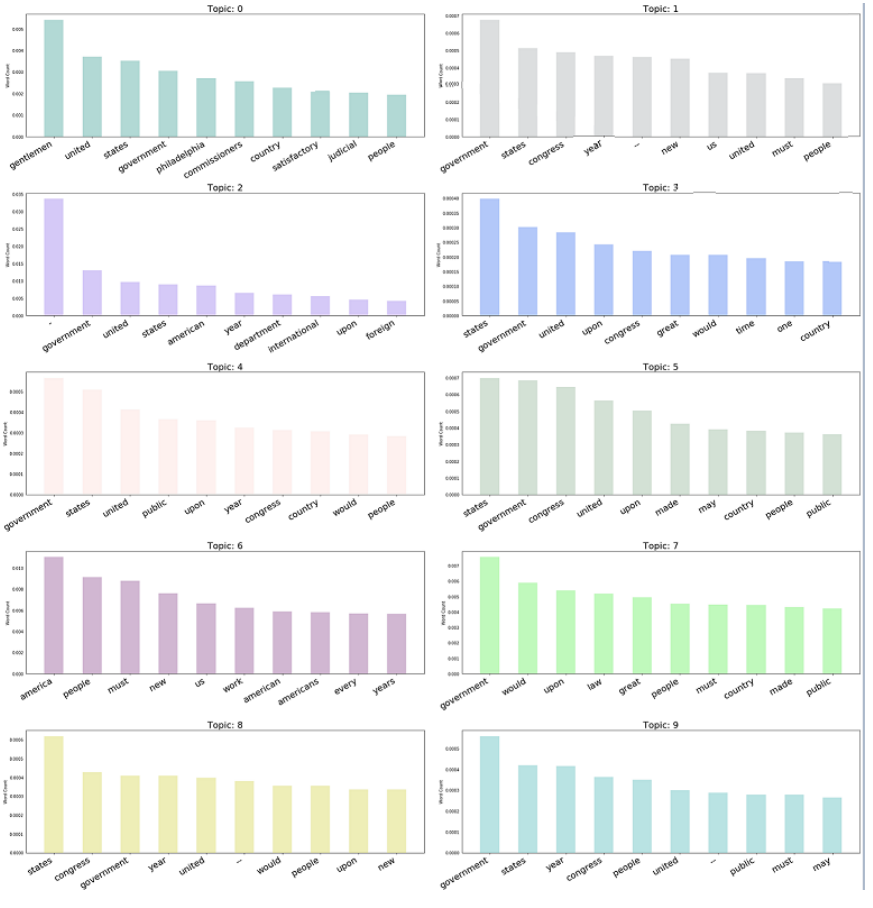
I have randomly considered 10 topics with 20 words each and tried to annotate them with topics which formed significant part of the US History. The topics here vary from various topics such as Vietnam War,Great Depression,World War 2, etc. which formed an essential part of the US history.

LDA Model– After performing the tokenization the next step undertaken by us was to perform LDA where the number of topics were varied from 2 to 180 and coherence values

for each of the number of topics was computed to find out the most appropriate number of topics for the dataset. The coherence value score seemed to increasing steadily early on and flattening out close to 40 with a significant rise coming at 26 followed by a dip.Thus the number of topics selected was 26 based on the coherence values.



The below are the first 10 topics from the total showing the first 10 words for each topic and their co-efficient.



One of the things which was clearly observable during running the code was that the LDA algorithm took significantly more time to run as compared to the LSI algorithm.The major distribution between the two Probabilistic algorithms is that these two assume different distributions over the prior in a Bayesian framework which varies their results.

Following are the annotated topics.These are randomly selected 10 topics.

Topic 0 **Administration**	'0.000*"government" + 0.000*"states" + 0.000*"congress" + 0.000*"united" + ' '0.000*"year" + 0.000*"people" + 0.000*"must" + 0.000*"may" + 0.000*"would" ' '+ 0.000*"country" + 0.000*"upon" + 0.000*"great" + 0.000*"--" + ' '0.000*"made" + 0.000*"public" + 0.000*"one" + 0.000*"new" + ' '0.000*"american" + 0.000*"time" + 0.000*"every" + 0.000*"war" + ' '0.000*"state" + 0.000*"present" + 0.000*"last" + 0.000*"national"
Topic 4 **Government policies on war and peace**	'0.009*"public" + 0.008*"may" + 0.007*"government" + 0.006*"states" + ' '0.006*"country" + 0.005*"united" + 0.005*"war" + 0.004*"made" + ' '0.004*"would" + 0.004*"state" + 0.004*"us" + 0.004*"great" + 0.004*"peace" ' '+ 0.004*"congress" + 0.004*"time" + 0.004*"every" + 0.003*"citizens" + ' '0.003*"general" + 0.003*"millions" + 0.003*"upon" + 0.003*"last" + ' '0.003*"present" + 0.003*"necessary" + 0.003*"also" + 0.003*"treasury"
Topic 9 **Government Laws**	'0.008*"government" + 0.007*"would" + 0.006*"upon" + 0.006*"states" + ' '0.005*"law" + 0.005*"great" + 0.005*"congress" + 0.004*"people" + ' '0.004*"country" + 0.004*"may" + 0.004*"public" + 0.004*"one" + 0.004*"men" ' '+ 0.004*"made" + 0.004*"united" + 0.003*"must" + 0.003*"present" + ' '0.003*"power" + 0.003*"time" + 0.003*"business" + 0.003*"service" + ' '0.003*"shall" + 0.003*"state" + 0.003*"war" + 0.003*"national"
Topic 12 **National Unity**	'0.014*"states" + 0.010*"government" + 0.009*"united" + 0.007*"congress" + ' '0.007*"upon" + 0.007*"may" + 0.006*"would" + 0.005*"public" + ' '0.005*"country" + 0.005*"great" + 0.005*"made" + 0.004*"last" + ' '0.004*"state" + 0.004*"year" + 0.004*"people" + 0.004*"war" + 0.003*"time" ' '+ 0.003*"present" + 0.003*"subject" + 0.003*"citizens" + 0.003*"power" + ' '0.003*"mexico" + 0.003*"treaty" + 0.003*"act" + 0.003*"part"
Topic 14 **War and agriculture**	'0.007*"would" + 0.006*"government" + 0.005*"made" + 0.004*"war" + ' '0.004*"public" + 0.004*"country" + 0.004*"much" + 0.004*"states" + ' '0.004*"year" + 0.004*"great" + 0.003*"national" + 0.003*"agriculture" + ' '0.003*"congress" + 0.003*"necessary" + 0.003*"part" + 0.003*"one" + ' '0.003*"people" + 0.003*"land" + 0.003*"state" + 0.003*"power" + 0.003*"law" ' '+ 0.003*"farmer" + 0.003*"us" + 0.002*"property" + 0.002*"may"
Topic 15 **Government administration and laws**	'0.011*"government" + 0.009*"states" + 0.007*"year" + 0.007*"united" + ' '0.007*"upon" + 0.006*"congress" + 0.005*"made" + 0.004*"may" + ' '0.004*"american" + 0.004*"-" + 0.004*"last" + 0.004*"country" + ' '0.003*"department" + 0.003*"would" + 0.003*"people" + 0.003*"law" + ' '0.003*"service" + 0.003*"great" + 0.003*"time" + 0.003*"secretary" + ' '0.003*"public" + 0.003*"foreign" + 0.003*"present" + 0.003*"commission" + ' '0.003*"new"
Topic 20 **Spanish and cuban tensions and International Peace Treaties**	'0.015*"united" + 0.014*"states" + 0.009*"government" + 0.009*"spain" + ' '0.006*"may" + 0.006*"war" + 0.004*"cuba" + 0.004*"made" + 0.004*"congress" ' '+ 0.004*"spanish" + 0.004*"great" + 0.004*"every" + 0.004*"country" + ' '0.003*"act" + 0.003*"treaty" + 0.003*"shall" + 0.003*"part" + 0.003*"peace" ' '+ 0.003*"vessels" + 0.003*"commerce" + 0.003*"without" + 0.003*"necessary" ' '+ 0.003*"state" + 0.003*"nations" + 0.003*"present"
Topic 22 **War against Terrorism**	'0.006*"terrorists" + 0.006*"every" + 0.006*"america" + 0.006*"world" + ' '0.006*"tonight" + 0.006*"americans" + 0.006*"terror" + 0.005*"many" + ' '0.005*"us" + 0.004*"people" + 0.004*"freedom" + 0.004*"country" + ' '0.004*"states" + 0.004*"united" + 0.004*"great" + 0.004*"ask" + ' '0.004*"american" + 0.004*"come" + 0.003*"terrorist" + 0.003*"war" + '
	'0.003*"together" + 0.003*"may" + 0.003*"citizens" + 0.003*"seen" + ' '0.003*"americas"
Topic 24 **Administration and Economy**	'0.010*"year" + 0.009*"war" + 0.009*"government" + 0.008*"federal" + ' '0.008*"program" + 0.007*"dollars" + 0.006*"fiscal" + 0.006*"economic" + ' '0.006*"congress" + 0.006*"expenditures" + 0.006*"national" + ' '0.006*"million" + 0.005*"legislation" + 0.005*"united" + 0.005*"public" + ' '0.005*"must" + 0.005*"production" + 0.004*"--" + 0.004*"states" + ' '0.004*"billion" + 0.004*"world" + 0.004*"policy" + 0.004*"administration" + ' '0.004*"employment" + 0.003*"veterans"
Topic 25 **National and international peace and Security**	'0.011*"--" + 0.009*"must" + 0.007*"people" + 0.007*"world" + 0.007*"new" + ' '0.006*"us" + 0.006*"america" + 0.005*"congress" + 0.005*"years" + ' '0.005*"year" + 0.005*"american" + 0.005*"government" + 0.005*"nation" + ' '0.004*"one" + 0.004*"make" + 0.004*"every" + 0.004*"work" + 0.004*"peace" + ' '0.003*"time" + 0.003*"help" + 0.003*"nations" + 0.003*"americans" + ' '0.003*"country" + 0.003*"security" + 0.003*"war"

I have randomly considered 10 topics with 20 words each and tried to annotate them with topics which formed significant part of the US History.LDA generates more human-interpretable topics from document corpus.Thus making LDA a better suited algorithm to understand the topics hidden.

Decade summarizing Algorithm– For the decade summarizing part I had first performed the topic modelling algorithm using LDA and then for each document in the corpus I had calculated the most significant topic contributing to each document(This was found using the Gensim Package as well as I have annotated these topics)After finding the most significant topic for each document I have also mined the keywords present in the document responsible for the topic assigned to the document.After this I had manually inspected the documents with the temporal frequency of a decade. The major topics contributing to the documents had a significant relations with the major events in the United States history. Here is the list of the the major topics present in the State of the Union Speech for the previous century.

1900-1910–Government and Administration Laws formed the main crux of the speeches.

1910-1920–Government policies on War and Peace formed the main components along with administration .(World War 1 and Spanish Flu Peak)

1920-1930–Laws,Medical,War and Agriculture formed the main themes.(Re-Built Begins)

1930-1940–Discussions on Economic and Administrative Reforms.(Great Depression Period with unemployment at all time high)

1940-1950–Nazi suppression in World War2 and International Peace.(WW-2 and UNO formation)

1950-1960–International Peace and Democracy.(Korean and Cold Wars Begin)

1960-1970–Cuban and Vietnam Wars and Civil Rights movement.(Martin Luther King Jr. Assassination)

1970-1980–Cold War and Space Explorations against USSR(Apollo Missions)

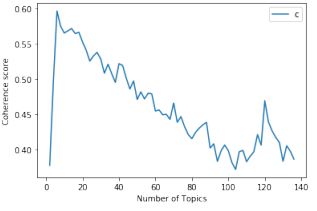
1980-1990–World Power and Military expansion(USSR Dissolution, American Economy wins)

1990-2000–International Peace and Unity.(Terrorism Attack rises around the world)

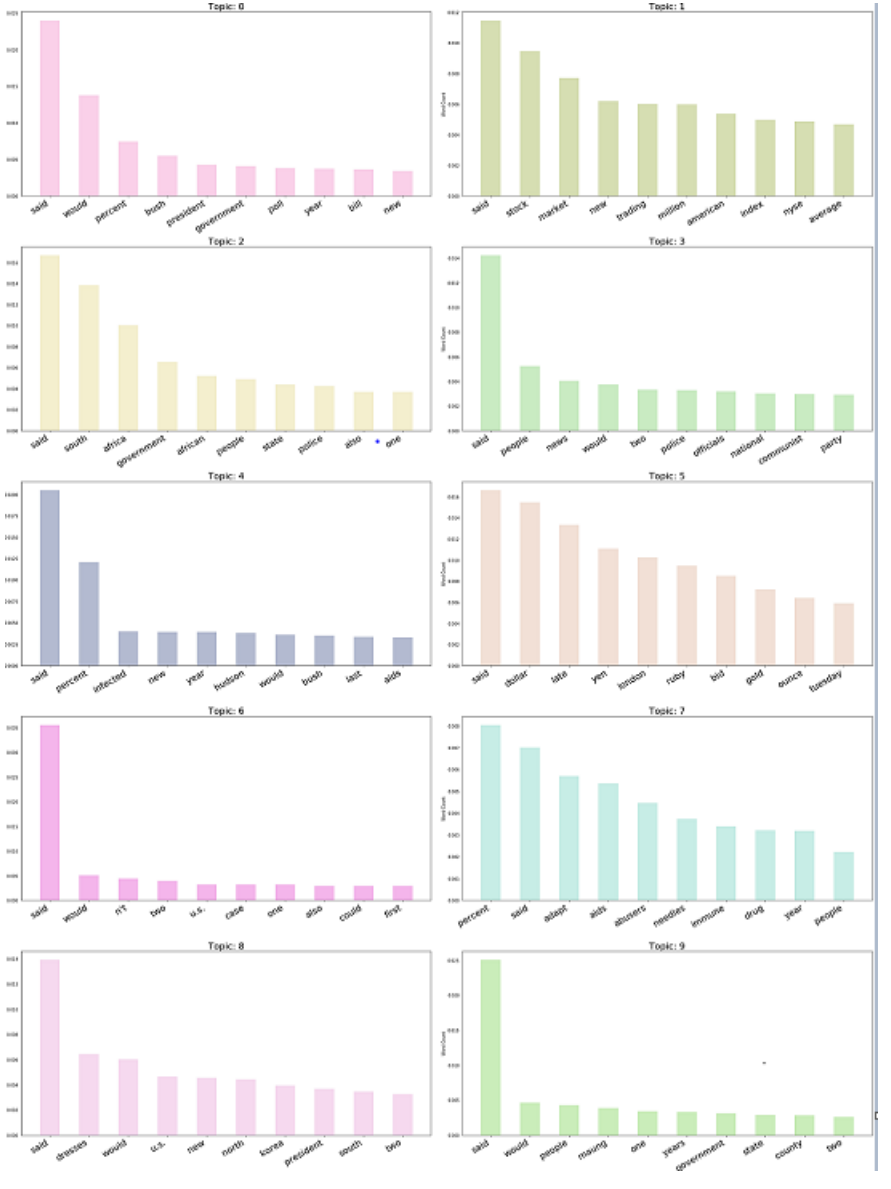
Thus in the topics of the state of the union speech in the previous century a clear transition from the discussion of just about US to talking about international Peace and Security could be seen.This also supports the fact that during this period the Country also transformed to World Power Status and started dominating the world talks and summits.Another shift which is quite relevant is the shift from the talks of economics to terrorism and war in the later half of the century which proves that the country did move ahead of the Great Depression Period and was facing new issues such as Cold War.

6 Second Dataset

The second dataset comprises of the AP wire stories.The dataset was smaller in size as compared to the previous dataset and as such is expected to have less number of topics. We had to perform the LDA and compare the performance of the algorithm on the two datasets.After performing the tokenization the next step undertaken by us was to perform LDA where the number of topics were varied from 2 to 120 and coherence values for each of the number of topics was computed to find out the most appropriate number of topics for the dataset. The coherence value score seemed to be steadily high early on and then decreasing before flattening out close to 40 with a significant drop coming at 20.Thus the number of topics selected was 20 based on the coherence values.



The below are the first 10 topics from the total showing the first 10 words for each topic and their co-efficient.



Following are the annotated topics.These are randomly selected 10 topics.

Topic 0 **Administration and law**	'0.027*"said" + 0.010*"police" + 0.006*"would" + 0.006*"u.s." + ' '0.006*"court" + 0.005*"government" + 0.004*"two" + 0.004*"one" + ' '0.004*"attorney" + 0.004*"case" + 0.004*"east" + 0.004*"military" + ' '0.004*"germany" + 0.004*"also" + 0.004*"officials" + 0.003*"united" + ' '0.003*"judge" + 0.003*"german" + 0.003*"former" + 0.003*"north" + ' '0.003*"charges" + 0.003*"west" + 0.003*"told" + 0.003*"trial" + ' '0.003*"Thursday"'
Topic 1 **Drug dealings**	'0.014*"said" + 0.006*"owen" + 0.005*"dresses" + 0.005*"government" + ' '0.004*"way" + 0.004*"police" + 0.004*"people" + 0.004*"dress" + ' '0.003*"united" + 0.003*"maung" + 0.003*"greyhound" + 0.003*"says" + ' '0.003*"spain" + 0.003*"gunter" + 0.003*"cocaine" + 0.003*"would" + ' '0.003*"organizations" + 0.003*"demonstrators" + 0.003*"students" + ' '0.003*"bridal" + 0.003*"bank" + 0.002*"u.s." + 0.002*"barahona" + ' '0.002*"two" + 0.002*"burma"'
Topic 2 **Prsident Bush and US -Soviet Summit**	'0.026*"said" + 0.005*"people" + 0.005*"new" + 0.005*"\n\t" + 0.005*"bush" + ' '0.004*"president" + 0.004*"would" + 0.004*"state" + 0.004*"one" + ' '0.003*"years" + 0.003*"south" + 0.003*"dukakis" + 0.003*"two" + ' '0.003*"soviet" + 0.003*"government" + 0.003*"also" + 0.003*"national" + ' '0.003*"first" + 0.002*"today" + 0.002*"last" + 0.002*"year" + 0.002*"city" ' '+ 0.002*"states" + 0.002*"house" + 0.002*"campaign"'
Topic 3 **Computer technology**	'0.025*"said" + 0.005*"\n\t" + 0.004*"one" + 0.004*"years" + 0.003*"new" + ' '0.003*"first" + 0.003*"time" + 0.003*"two" + 0.003*"would" + 0.003*"also" + ' '0.003*"could" + 0.003*"program" + 0.003*"office" + 0.003*"computer" + ' '0.003*"drug" + 0.003*"people" + 0.003*"aids" + 0.002*"national" + ' '0.002*"company" + 0.002*"may" + 0.002*"says" + 0.002*"last" + 0.002*"mrs." ' '+ 0.002*"get" + 0.002*"system"'
Topic 4 **Climatic calamities**	'0.025*"said" + 0.006*"water" + 0.006*"people" + 0.005*"one" + ' '0.004*"county" + 0.004*"ohio" + 0.004*"fire" + 0.004*"area" + 0.003*"miles" ' '+ 0.003*"flight" + 0.003*"two" + 0.003*"found" + 0.003*"eastern" + ' '0.003*"homes" + 0.003*"weather" + 0.003*"\n\t" + 0.003*"could" + ' '0.003*"feet" + 0.003*"river" + 0.003*"airline" + 0.003*"continental" + ' '0.002*"state" + 0.002*"night" + 0.002*"central" + 0.002*"thursday"'
Topic 5 **American stock market exchange**	'0.023*"percent" + 0.016*"said" + 0.015*"million" + 0.010*"year" + ' '0.009*"billion" + 0.009*"market" + 0.007*"sales" + 0.007*"new" + ' '0.007*"stock" + 0.005*"prices" + 0.005*"company" + 0.005*"inc." + ' '0.005*"rose" + 0.004*"business" + 0.004*"economy" + 0.004*"last" + ' '0.004*"york" + 0.004*"index" + 0.004*"economic" + 0.004*"exchange" + ' '0.004*"analysts" + 0.004*"week" + 0.004*"interest" + 0.003*"average" + ' '0.003*"shares"'
Topic 6 **Farming and Labor**	'0.012*"cents" + 0.009*"said" + 0.009*"cars" + 0.009*"cent" + 0.008*"lower" ' '+ 0.008*"futures" + 0.008*"would" + 0.007*"higher" + 0.007*"corn" + ' '0.006*"labor" + 0.006*"percent" + 0.006*"fuel" + 0.005*"party" + ' '0.005*"workers" + 0.005*"contract" + 0.005*"bushel" + 0.005*"standards" + '
	'0.004*"poll" + 0.004*"soybean" + 0.004*"controls" + 0.004*"last" + ' '0.004*"industry" + 0.004*"year" + 0.004*"mpg" + 0.004*"pollution"'
Topic 7 **United States International Tradings**	'0.023*"said" + 0.011*"would" + 0.008*"united" + 0.007*"states" + ' '0.006*"u.s." + 0.006*"trade" + 0.005*"president" + 0.004*"world" + ' '0.004*"agreement" + 0.004*"also" + 0.004*"talks" + 0.003*"japan" + ' '0.003*"billion" + 0.003*"iraq" + 0.003*"government" + 0.003*"one" + ' '0.003*"bush" + 0.003*"countries" + 0.003*"economic" + 0.003*"officials" + ' '0.003*"new" + 0.003*"nations" + 0.003*"two" + 0.003*"minister" + ' '0.003*"foreign"'
Topic 8 **Economy and Defense**	'0.016*"said" + 0.007*"year" + 0.007*"billion" + 0.007*"tax" + ' '0.007*"defense" + 0.006*"would" + 0.006*"state" + 0.006*"government" + ' '0.006*"percent" + 0.005*"money" + 0.004*"budget" + 0.004*"last" + ' '0.004*"spending" + 0.004*"fiscal" + 0.004*"million" + 0.004*"public" + ' '0.004*"federal" + 0.003*"people" + 0.003*"pacs" + 0.003*"also" + ' '0.003*"members" + 0.003*"law" + 0.003*"security" + 0.003*"report" + ' '0.003*"energy"'
Topic 9 **Space exploration and NASA**	'0.018*"dollar" + 0.015*"late" + 0.014*"said" + 0.011*"yen" + 0.010*"london" ' '+ 0.009*"gold" + 0.008*"bid" + 0.007*"tuesday" + 0.007*"venus" + ' '0.007*"ounce" + 0.006*"galileo" + 0.005*"bank" + 0.005*"friday" + ' '0.005*"compared" + 0.005*"price" + 0.005*"nasa" + 0.005*"earth" + ' '0.004*"space" + 0.004*"europe" + 0.004*"troy" + 0.004*"francs" + ' '0.004*"tokyo" + 0.004*"shuttle" + 0.004*"u.s." + 0.004*"rates"'

Here the topics are clearly more human interpretable as compared to the previous dataset

which is also supported by the fact that the coherence score for the second dataset i.e. AP wire dataset is significantly higher as compared to the State of the Union Speech along with the perplexity score(this should be low) which also lower for the second dataset.Thus proving that the second dataset produced better results (more human interpretable and easy to grasp) results.

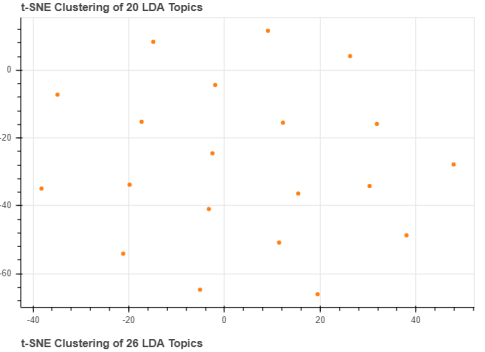
Perplexity: -9.522917658719486 for second dataset

Perplexity: -8.404044364429746 for fist dataset

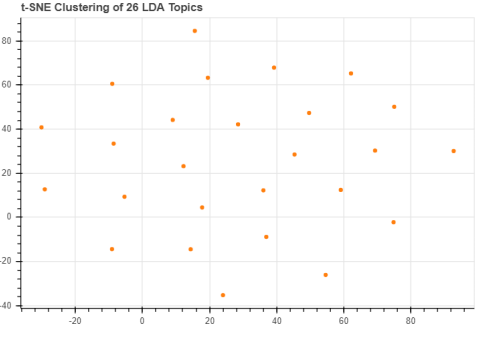
A major reason for this could be the document composition with each document in the first dataset providing little input for each of the topics and had several repeating phrases whereas the second dataset consisted of the documents which had focused information on topics and had fewer number of topics per document which meant each topic was covered significantly in the documents of the second dataset.The topics in the first dataset were also seemingly related to each other more as compared to the second dataset,uncorrelated random normal points generally improve the performance of the LDA algorithm.

Here is the t-SNE results on the topic vector of the two datasets LDA algorithm.

Dataset -2



Dataset -1



Clearly the topics in the dataset 2 are more separated as compared to the first dataset thus making the performance of the algorithm better.

7 Conclusion

Thus, I would like to conclude by saying that LDA is better topic modelling algorithm as compared to LSI as it produces more human understandable i.e. more coherent topics.The second dataset had cleared topics per document.The concepts captured by the algorithm have resemblance to human themes but they need to be carefully examined to derive them.