Final Project- Text analysis

Extracting topics from New York Times Articles

Topic modelling using LDA python (NLTK, SKLEARN)

A close up of a newspaper

Description automatically generated

-Srivishnu Maganti

Your research question or analysis task and the motivation/justification for your task:

**Introduction and motivation:**

Basically, the main task of this project is to carry out topic modelling for New York times articles. Here 30 articles from New-York Times website have been considered. These articles are of different topics. In this era of advanced technology and development there are many aspects that are needed to be discussed and acknowledge common people, Media is playing a key role in extracting opinions and information from different parts of world and presenting Infront of people through articles, television, internet etc. But people would like to read and understand articles in the topics that they are interested in, out of thousands of articles they would like to read topics that they are interested in. So, my main motivation behind this project is to get the topics out of these articles. This project will be useful for readers by making them choose to read based on their interest. Here we considered each article as document, and we created a dataset which contains all the articles. This modelling is done using python, in which few natural language-based libraries are being used. We applied LDA using SKlearn module of python. After performing topic modelling, we are able to get satisfying topics of these articles.

A justification of your choice of either a rule-based or statistical/machine learning approach to text analysis that includes the strengths and weakness of the approach.

Reference-1:

This reference explains that rule-based approach is a kind of approach where systems employ rules such as if, elseif, then as a major representation paradigm. Rule based approaches has been designed by two different motivations which as phycological modelling and expert system creation.

Basically, there are three components in rule-based approaches which are working memory the rule base and the inference engine. Firstly, the working memory presents the working ana execution of system, where it knows everything about system simply, we can name as storage space. Apart from that, this helps the rules to communicate with each other. Rule base portrays the set of rules and knowledge about the integrating system. So finally, inference engine will take out rules from rule base and information about system from working memory and then execute the program.

Initially, working memory has the required information to run the program in inference engine, but to get the required results, it takes rules from rule base to compile the program. Here by we understand that it is an approach that rules are provided to the data and then system acts according to the rules that we applied. So, there are few advantages and dis-advantages on this approach.

So as per my understanding, in our lab we used profiler plus, XERF for running programs, so XERF is the rule base which will provide rules to the system, and profiler plus initial has working memory and it has inference engine which is able to run our programs.

Reference 2:

This article explains that rule-based approach is basically a hand-crafted approach which depends on system of rules. In this approach human is involved in the process of stepwise implementation and development, so if there is any issue that needs to be fixed, we can easily be able to fix it. Since all the rules are written by human so it is easy to localize the bug.

Furthermore, this system doesn’t require any kind of training corpus, as this system is about querying the data and improvising based on new data to obtain required output.

When it comes to disadvantages, we can tell that working on rule-based approach requires a skilled person with who has the knowledge on the input data, rule base and inference engine.

for instance, when we are working with sentiment analysis assignment in profiler plus, we can tell that without proper understanding of input corpus it will be difficult to write generalized rules, sometimes rules might be contradicting.

Furthermore, for this topic modelling project rule-based approach will be very hard, writing rules to figure out topics from the corpus is a difficult task. This is a reason for selecting machine learning approach.

There are very few references explaining topic modelling using rule-based approach.

Reference-2 explains machine learning approach as an algorithmic method which is based on probabilistic results. In this the system starts analyzing the input corpus with its own set of rules and classifiers, system starts improvising itself. Initially, the system requires training data where it learns the training data and it tests it with the testing set. There are few advantages listed such as learnability, fast development. Where we do not require skilled professional and manual grammar coding.

Also, when it comes to generalizing statistical approach will be more effective, for input data with different forms, we need to more generalized rules and it is hard in rule based approach.

For this project specifically statistical approach works more because there are lot of keywords, which makes classifier easy to learn the stats behind the words to obtain the topics.

Reasons why I am considering statistical learning overrule-based are:

1. Topic modelling with rule-based approach is very difficult.
2. As this is a kind of clustering method which contains as lot of data points, so machine learning method would be a better approach.
3. Using python with NLTK, SKlearn libraries it is easy to preprocess, vectorize and implement LDA technique for extracting topics out of corpus.

Because of all these reasons I have implemented machine learning approach for this project rather than rule-based approach.

A description of the selected technique

**Topic modelling:**

To understand the basic working of topic modelling and mathematics behind that, few articles have been reviewed.

Reference-3 determines that it is a type of statistical analyses tool which will help us to explore the topics behind the documents. It is a frequently used tool in text mining, text analyses to understand the hidden deep semantics behind the large amount of text. For instance, if we are required to understand a corpus of data then we can do topic modelling where it results out document-topic and topic-word semantic. Also, in topic modeling the output is presented as topic to word such as topic A contains cats, dogs, lion etc. topic B contains words like jeans, JCPenney, H&M etc. Then we can conclude that topic A is referring to animas and topic B is referring to fashion and clothing. So, this way we are going to understand the text data that is present.

Topic modelling is also called probabilistic topic models because the computation of data is based on statistics, probability. It uses statistical algorithms to understand and interpret the hidden relationships in text data that is given. In this information age we are presented with huge amount of unstructured data where there are lot of stories, lot of text data available, so we need to make these understand and arrange in an order for further study, so we need these techniques to organize and understand the insights of the text data. Previously, this is used in text mining, but it is being used to analyze complex data like genetics information, images, networks etc. This has huge development in bioinformatics and computer vision.

Earlier topic models were invented in 1998, this was the era which has been developed topic models using latent semantic analyses, then in 1999 PLSA was invented which is a probabilistic assumption of LSA and later on LDA was invented, which is the most common method used these days for performing topic modelling, basically the difference between LDA, PLSA is that LDA introduces Dirichlet prior distributions over document-topic and topic-word distributions that are used in PLSA and LSA. There is an intuition used which is documents are made by topics and topics are made by words. Furthermore, there are few extended versions of LDA such as Pachinko allocation, here the feature that is being modelled is that it uses correlation between topics along with correlation between words to construct topics. Hierarchical latent tree analysis is an alternate suggestion for LDA, here it considers word cooccurrences using trees of latent variables and the states of the latent variables, which will result in soft clusters of documents, these are interpreted as topics.

Topic modelling is an unsupervised learning technique, which is similar to k-means, this is a kind of clustering techniques which is based in probabilistic assumptions and intuitions to get the result of text data.

After a brief introduction we are interested in finding the mathematics behind the algorithms, for referring different algorithms used in topic modelling I have referred a medium article which explains LSA, PLSA and LDA and importance of LDA algorithm compared with LSA and PLSA.

Reference-4 explains topic modelling in a comprehensive way, this summarizes topic modelling: In text mining and text analyses there is a hierarchy of path through which we are going to extract meaning this is from first words to sentences to paragraphs to documents. At document level, one of the most useful way to understand text is by analyzing its topics. The process to understanding and examining topics from collection of documents is nothing but topic modelling.

Basic intuition:

1. Each topic consists of collection of words.

2. Each document consists of mixture of topics.

Topics models are built in this intuition which are in hidden layer, our primary goal is to find out these hidden sematic layers, there are different way to uncover these layers from topics which are being discussed below.

Latent semantic analyses:

This is one of the fundamental techniques used for topic modelling, the basic idea behind this is to form a matrix of documents and terms and decompose that matrix into document-topic matrix and topic-term matrix. This is one of the foundational techniques in topic modelling,

Document term matrix where each column represents a word and each row represents a document that is generated, primarily basic LSA includes each entry to be a raw count of number of times the word is being repeated in specific document. But later with only count we cannot be able to specify the word importance in the whole document, so instead LSA uses “Term frequency- inverse document frequency” weight score. So, document term matrix will be formed using these weights. TF-IDF weights for term (j) and document (I) can be computed by using the following equation:

w(i,j)=term frequency(i,j)\*log(total documents)/(document frequency(j))

Here, the term will have larger weight if it occurs frequently across the document and infrequently across the corpus. For instance, let us consider word Vishnu occurs more frequently across the document but it does not appear in other document then the weight for this word will be less compared with other word which is repeated more in document. Our basic goal is to decompose this matrix in to two different matrices which are document topic and topic word matrix. So basically, this matrix will be noisy, and sparse, to obtain required latent semantics behind this matrix we need to do dimensionality reduction using truncated SVD (singular value decomposition), this is a common technique used in machine learning to decompose datasets in to required form and to reduce the dimensionality, reducing dimensions are very important in statistical algorithms to reduce the noise and sparsity in the dataset.

A~U(t)S(t)(V(t)) ^T

Truncated SVD is a kind of technique used in linear algebra that factorizes the matrix into product of three different matrices. Here s(t) is a diagonal matrix which contains singular values of A. Truncated SVD only works with T columns of U (document term matrix) and V (topic word matrix) so here the t columns are number of topics which are required (hyper parameter) which can be changed according to the requirement. After performing SVD for document term matrix we will obtain the product of document-topic matrix and term-topic matrix. Furthermore, cosine similarity can be applied to obtain the similarity of different documents, similarity of different words, and similarity of terms and documents.

There are few drawbacks listed for LSA which are

1. Lack of interpretability, as we don’t know what the topics are.

2. For getting desired results, large amount of data sets is required.

3. Less efficient in presentation.

Probabilistic latent semantic analysis (PLSA): The basic difference is that previously, in latent semantic analysis for dimensionality reduction it uses SVD, but in PLSA it uses probabilistic method.

Basically, we have two common assumptions for performing topic modelling which are, documents as a mixture of topics and topics as collection of words, this model will spin the two intuitions with probabilistic assumptions given a document C and topic T is present in that document as p(T/C) and given a topic T and word W, word W is drawn from topic T with probability p(W/T).

Diagram

Description automatically generated

Figure 1 image taken from reference 4

The basic core idea to find a probabilistic model with hidden topics that can generate data that we can observe in our document-term matrix.

By this we can find the join probability of P(D,W), mathematically we can represent this as

P(D,W)=P(D)ΣP(Z/D)\*P(W/Z)

The right-hand side of equation will determine the probability of seeing a document, then when we follow the distributions, we can understand how likely to see a certain word within that document. All the parameters of our model are presented in right hand side, where P(D) can be determined easily from our corpus, the other two parameters are estimated using expectation-maximization algorithm. This statistical algorithm is nothing but used to find the parameter which depends on topics.

Also, there is one more method used to find P(D,W), this method is modelled as similar method with LSA because of similarity in the phenomenon.

In the first parametrizing we are starting with documents P(D) then topic to documents P(Z/D) then word to topic which is P(W/Z). So the other parametrized method is that we start with topic P(Z) and then independently generating documents P(D/Z) and then word with P(W/Z).

Although, this model looks very simple and quite different approach, but this method has few flaws that are needed to be worked on:

1. If we have new documents, we cannot be able to approximate p(D) for them.

2. Number of documents grows rapidly with number of documents we have so this leads to over fitting.

Addressing these issues that are being faced by PLSA, LDA has been invented.

Latent Dirichlet allocation (LDA): So, in LSA we used TF-IDF weights and SVD for distributions, in PLSA we used probabilistic models to distribute topic-document and word-topic distributions, but here in LDA we are using Dirichlet distributions, which will lead to better generalization.

For simple way to understand let us say we have three topics in our project which are business, fashion and political, if we are relying on Dirichlet assumptions, we are going to get results like:

Mixture X: 88% political, 6% fashion, 6% business.

Mixture Y: 6% political, 88% fashion, 6%business.

Mixture Z: 6%political, 6% fashion, 88% political.

If we are trying to use weighted Dirichlet distributions, having large weight on specific topic we are going to get the required results. But if we consider probabilistic distribution, we are going to get our results like 33% political, 33% fashion, 33% business, it is hard to predict results.

Here the method varies the way data is being sampled, from a Dirichlet distribution we first draw a random sample which will represent the topic distribution, or topic mixture. From this initial topic distribution, we are going to select particular topic.

We continue the same step for next random sample of data which represents the same topic and from that we choose the words this is a kind of generating process.

There are few points which explains why LDA works better than PLSA because

1. LDA can generalize new documents easily.

2. In PLSA document probability is fixed point.

3. Sampling of new documents can be done very easily when compared with PLSA.

With using LDA we will be able to extract human interpretable topics which can be displayed in the form of words. Results of this project displays in the same manner which each topic contain few words which are interpretable by humans and we can understand the topics easily if our preprocessing is perfect.

Even though this article explains the phenomenon easily, but there are few more clarifications needed to understand, so there is a blog which was posted in LinkedIn website by an employee, he explained LDA in a simple form which we can understand easily, mentioned as reference 5.

He explains that the purpose of LDA is to understand the representation of fixed number of topics, and from these topics learn the topic distribution from each document from corpus. There are few steps that we need to work to get the required output.

Firstly, we need to select number of topics that are needed to be considered. Then LDA will go through all the words of each document and randomly assign a topic to the word. After this LDA will have topic-word and document- topic distributions, after this step LDA will try to analyze per document by calculating two probabilistic assumptions which are p(topic/document) and p(word/topic) after calculating these LDA will reassign the topics if

P(topic1/document) \*p(word/topic1) < p(topic2/document) \*p(word/topic2)

If this happens then the topic will change from 1 to 2 for that word. Basically, first it is randomly assigning each topic to each word and then by using probabilistic calculations it will try to re assign until it satisfies. It will iterate until it gets done.

There is hyper parameter in which I used in project which is a method in LDA which is called online LDA here the difference between LDA and online LDA is that for LDA we need to calculate probability of each word to each topic in order to re-assign the topics with new topics (explained above topic1, topic2). So, calculating this probability is very hard so we try to estimate this probability using online stochastic optimization, which has given better results than other optimizing techniques. After trying with different sampling and optimizing methods finally online stochastic is the one which followed better results.

These articles above explains difference between LSA and LDA and specifies why LDA is better approach than LSA.

Reference 6 explains about hyperparameters that are needs to be considered in LDA which are known as alpha and beta respectively, it explains that lower value of alpha will assign fewer topics to each document and lower value of beta results a topic with fewer words.

The main difference between LSA and LDA is that LDA will assume document to topic and topic to word distributions are Dirichlet, whereas PLSA considers the distributions are normal and LSA does not consider any distribution. So, as LDA is considering Dirichlet assumptions, therefore the representation of LDA topics and documents are more transparent.

So after careful study of all these articles, basically LDA is better for performing topic modelling when compared with other algorithms.

Finally, I will try to summarize LDA in points:

1. We need to select number of topics.
2. LDA assumes that document as mixture of topics and topics as mixture of words.
3. LDA will go through each document and assign a random topic to each word.
4. After assigning it will again repeat the process and calculate p(topic/document), p(word/topic) for each word and respective topic.
5. It re-assigns the word with new topic if the value of product of two probabilities are greater than value of product of two probabilities with other assumed(initial) topic.
6. Repeat this process until it reaches it gets done.

To understand the purpose of LDA and Dirichlet distribution one more medium article has been considered (reference 7) this article explains that there are total three hyper parameters in this topic modelling where one is alpha, one is beta, one is N. Here N is number of topics that we are going to give while programming in the beginning. Then we can differ alpha to increase or decrease the mixture of topics, like less alpha will have less mixture of topics and more alpha will have more mixture of topics. And then they explained beta hyper parameter which controls the distribution of words per topic, when beta value is less topics will have less words and beta is more than topics will have more words, these can be controlled by us depending on our requirement and dataset, ideally we need to set alpha and beta below one for better results as we need only few topics in our project.

Why do we need to use LDA: as this is a kind of clustering technique, we may be confused why do we need LDA of we can culture using K-means? But, if we consider topics as clusters, and probabilities as proportion of clusters, then using LDA is a way of simple clustering without any variation and disturbance in data we have.

Also, by using LDA we are reducing the number of dimensions of the dataset, if number of topics are less than documents. Apart from uncovering topics, this is one of the dimensionality reduction technique that can be used for text data, furthermore we can apply any machine learning model to investigate the data further.

As LDA learn more with number of iterations, here I have included maximum iterations are 50 so that the model learns the text more when it iterates more.

A flow chart of the overall ***text analysis*** process:

GETTING THE TOPICS FROM NEW YORK TIMES ARTICLES

IMPORTING THE DATA IN TO PYTHON IN THE FORM OF LIST OF STRINGS

PREPROCESSING THE DATA TO DESIRABLE FORMAT

FEATURE EXTRACTINNG USING COUNT VECTORIZER

LATENT DIRICHLET ALLOCATION

COMPARING THE RESULT WITH HAND NOTED TOPICS TO DISCUSS EFFICENCY OF MODEL

Whole topic modelling workflow will be as follows:

1. We need to look at the dataset we have and annotate with the potential topics that we are looking for in the topic modeling output. We need to first infer the data carefully and note down all the topics that are being discussed, here I have considered five topics for my project to be discovered and they are 1. Business, 2. Fashion, 3. Political, 4. Terrorism, 5. Sports.
2. Loading the dataset into python, so here my requirement is to load the dataset in the form of list of strings.
3. Then we need to preprocess the data according to our requirement, here i tried to implement expand contractions, remove special characters, tokenizing, stop words removal, n-gram inclusion, stemming and lowercasing.
4. Furthermore, we cannot implement raw data into the LDA process, so we need to extract vectors from that using count vectorization technique.
5. Then Latent Dirichlet allocation has been implemented in which topics will be extracted from the raw data. As explained above regarding LDA with changing few hyperparameters such as learning method, iterations, number of topics.
6. Then we can compare the results with step 1 so that we can discuss the potentiality of the model that was created. Each topic in the output will be presented as words and from that words we are going to estimate the meaning of that topic.

All these steps have been implemented in the project and preprocessing with feature extraction will be explained below in the following steps.

Here number of topics considered is 5, number of features are 2000 and number of top words are 30.

Code has been referred from the course material given for topic modelling assignment.

And preprocessing steps have been referred from text classification code given in course material.

**Preprocessing flow chart:**

EXPANDING CONTRACTIONS

REMOVE SPECIAL CHARECTERS

TOKENIZING

STEMMING

STOP WORD REMOVAL

N-GRAM INCLUSION

FEATURE EXTRACTING USING COUNT VECTORIZER

**Pre-processing steps:**

1. Expanding contractions: When we have a raw text data and there will be a lot of text data

Which is not preprocessed and there will be many words such as can’t, won’t, don’t etc.

And for these contractions if we try to remove special characters then all these words will go meaning less, so first we need to expand contractions. So, after performing expanding contractions for this raw text, lot of words have been expanded and they got the real meaning that is needed.

1. Remove special characters: After expanding contractions and there are lot of special characters in the data that is needed to be removed for making analysis more efficient.

If we do not remove these special characters and start extracting features then that will be an issue, because in feature extraction (count vectorizer) the vectorizer counts number of times the word has been repeated in the corpus and if the word has been repeated more than that will be in top features and considered for LDA, and our topics will be filled with special characters, because these characters will be most repeated in any text. For instance if we take my report as a text data and if we compare words according to number of times appeared then (, ? .) are the one more repeated and they will have more weightage. This will be difficult for analysis.

1. Tokenizing: After the above two preprocessing steps we need to tokenize each sentence and form into words. Tokenizing is important in topic modeling because text data cannot be directly imported in to count vectorizer, because count vectorizer will weight each and every word in the document, but if we directly apply count vectorizer with text data it will be hard for the system to analysis.
2. Stemming: Stemming is a process of converting the tokens into basic form, let us consider the tokens and if we did not do stemming to the tokens then the tokens will be in different forms of same meaning. Here there are many tokens in the corpus which will have different forms and that will have same meaning. For instance, if we consider the word pitching, the word pitch and pitching will have same meaning but different form and dealing with all these tokens with similar meaning in count vectorizer, vectorizer will try to give weightage to different words which has same meaning. So, while performing final LDA for all these words then we will have same words but in different forms. Like in features when we consider we will have same features, so analysis will not be done in a proper manner. But if we do stemming then new words will be added to the top features, here we only consider top 2000 features for topic extraction, so if we do stemming there is scope of getting new words to the analysis and we can get to extract any new topics that are needed
3. Stop words removal: Even after stemming and all other preprocessing, there are few more words which are not important to consider as features. Words like the, an, is, his etc. are few words which are being repeated in any text most. So, ones we do count vectorizer without removing stop words then words like the, an, as etc. will have more weightage and these words will be presented as topics, because topics are nothing but collection of words so if we are getting out topics which are completely filled with stop words then it will be hard to extract topics from that. Without removing stop words topics which are extracted will be meaningless.
4. N-gram inclusion: Here n gram inclusion means the way tokens are considered and weighted will be varied, for instance if we consider n grams as (1,2) then all the double words will be counted, because some words in our text will mean more if we consider the immediate next word. For example, there are few words like “do not” if we consider these words as independent vectors then they will have different meaning. Like “do” infers different meaning and “not” infers different meaning. Considering them independently will make completely different meaning of respective sentence, so few words are needed to be considered like this, that’s why I have included n grams as (1,2).
5. feature extraction:

To discuss what is going on in count vectorizer, I have referred a medium article, reference 8 this article states that in natural language processing when we are dealing with data, we cannot directly initiate the text data into the model, because text data cannot be imported into the model, so we need to convert the data into numbers (vectors), where we can directly initiate it into the model.

There are two steps we can introduce the data into the model, one is count vectorizer, and other is term frequency inverse document frequency. So, for this project we are focusing on count vectorizer.

So, after pre-processing steps that we consider to our data, the data has been modified, but this textual raw data cannot be directly imported to LDA, we need encoded version where the machine should understand the data.

So here we need to first understand bag of words model, Bow is a basic model which is used to allocate weights to the vectors. This method is very simple as it discards the order of text and just considers the occurrences of word. It converts the documents to fixed length of numbers where it has count of how many times the words are being repeated.

Now we will consider count vectorizer: So here count vectorizer itself tokenizes the terms and it performs few basic preprocessing such as n grams, lowercasing etc. A vocabulary of most common words is formed.

The encoded vector will return a vector which represents with the length of entire vocabulary and a numeric count of how many times the word has been repeated in the document itself.

So, the format will be like, the row represents the document, and the columns represent the word, and each entry will tell how many times the word is being repeated in the document.

Moreover, in count vectorizer the scikit-learn library has its own method of implementation, where we can import the required preprocessing steps required.

Here in my model, I have implemented:

1. Lowercasing, max\_features, stop words, n grams, min df, max df.
2. Stop words and n grams were explained in before section.
3. Min df specifies how much importance you want to give to less frequent words in our document, we do this because sometimes we might have few words which will be repeated only twice or ones. So, these words can be considered as outliers. As in our model outliers will be creating very bad influence on data.
4. Max df specifies how importance you are allocating to the most frequently repeated words, some time there are few words which might be repeated mostly, so we will think to ignore these words, that’s why we need to exclude these words. So, we consider these max df to indicate that.
5. And finally, max features indicate the vocabulary size that we want to consider. If we indicate max features, then we are instructing the machine what is the size of vocabulary that is needs to be considered.

**An evaluation of the results of the analysis approach as they relate to the purpose of the text analysis:**

As the motive of topic modelling is to get the hidden topics out of text, by using LDA. As per the given work flow all the required steps are being done, and finally I tried to infer five topics out from the NY times articles, so there are total thirty articles in the corpus and I have physically inferred that there are five topics that are being discussed commonly.

All the topics that are being discussed

1. Political: This is a generalized version of many topics such as international relations, politics, elections, law, rules made by government, emails between politicians, rules made by government to change rules, tweets and speeches given by politicians.
2. Business: Business deals, retail stores discounts, installing new stores, income, raising funds, company policy.
3. Sports: This is also a generalized version of baseball, tennis, basketball, Olympics. Because there are not many articles present for each sport, so consider each sport as each article will make no meaning.
4. Fashion: shopping, travelling, music, actors and their lifestyle, famous celebrities.
5. Terrorism ,crime and police action: This will come under few underlying topics which are like police cases, assassinations, court cases, ISIS related issues, gun firing etc.

So, these are the five topics that I want to produce in the output.

﻿﻿Topic 0:

said wa hi ha mr state year thi wednesday group citi like time team just new democrat right mani turkey islam unit nation union game peopl work european way member

Topic 1:

mr climat hi chang republican said support faison ha wa vote new run polit parti percent action democrat group american face obama spend arzberg charg plan guzmn portman extradit million

Topic 2:

mr said pigeon state student wa school clinton depart email ha year use jh lionsgat michalek 50 work hi percent new abedin privat person time space public server starz accord

Topic 3:

libyan islam state surt fighter origin command street libya like avenu fight moham citi forc push includ featur open style lauren week percent madison 150 artilleri brother bullet yard west

Topic 4:

fund rais ventur perkin billion firm invest year kleiner join thi percent mr file feng total knauf 12 data partner latestag quarter schlein ted capit growth compani privat accord work

These are the five topics that are in the output after respective preprocessing, feature extraction and LDA.

If we carefully observe the outputs here the each topic discuss about each topic, as we discussed earlier all the outputs are presented as clustered words, there are different words for each topic so we need to understand these words and simplify them as outputs.

\*(please consider referring my annotation guide before referring to explanation of outputs)

Topic 4:

fund rais ventur perkin billion firm invest year kleiner join thi percent mr file feng total knauf 12 data partner latestag quarter schlein ted capit growth compani privat accord work

This topic in the output has both influential and non-influential words such as billion, venture, capit (stemmed from capitol), company, quarter, private, partner. Some words might be in a common usage like billion, this word may be presented in many topics such as politics etc. But here billion refers to business. So, this topic finally indicate that they are referring to business.

Also, capitol, company, quarter, private, partner etc. are all referred to obviously business, so this topic might be referred as business.

Topic 3:

libyan islam state surt fighter origin command street libya like avenu fight moham citi forc push includ featur open style lauren week percent madison 150 artilleri brother bullet yard west

Here this topic contains words like Libyan, Islam, fighter, street fight, surt fighter. So, in the articles there are many topis which are discussed about turkey ISIS fighter, Libya army, attacks happened there, bombing schools. So, this article refers to terrorism and police force actions. I have mentioned police force actions because there are few articles which are discussing about firings happening, murder cases, victims etc. So I inferred all these articles under the same underlying topic named as police action and terrorism.

Topic 1:

mr climat hi chang republican said support faison ha wa vote new run polit parti percent action democrat group american face obama spend arzberg charg plan guzmn portman extradit

Here this topic has words like republican, democrat, Obama, voter, campaign etc. these words are mostly under the topic political so this topic is referred as political.

Topic 2:

mr said pigeon state student wa school clinton depart email ha year use jh lionsgat michalek 50 work hi percent new abedin privat person time space public server starz accord

Here this topic is very unclear and it is not good for interpretation, this topic does not give clarity on what it describes, like it has words like Clinton which describes about politics, Lionsgate, Starz, explains as business, Michalak as fashion. So this cannot be interpretated as a topic, because this topic itself is cluster of words which mean differently.

Topic 0:

said wa hi ha mr state year thi wednesday group citi like time team just new democrat right mani turkey islam unit nation union game peopl work european way member

Even this topic is unclear and there is word which has different meanings, because let us consider some words like turkey, nation, Islam as terrorism, European, democrat, state, members etc as political. So, there are two topic of words that represent this topic. So, it is very difficult to interpret the topics from this cluster of words.

As I have anticipated total of 5 topics, but I could only infer three of them. The three topics that I got are 1. Crime/terrorism 2. Political 3. Business, I could not infer fashion and sports.

There might be several reasons for this inefficient model, but as far as my knowledge, I have tried all the ways that I can. Like changing preprocessing, considering TF-IDF in the place of count vectorizer, changing number of topics, changing total number of features, number of topics. But this is the best output out of all other methods considered, it can be improvised by expertise. Also, there are only 2-3 articles which refer to fashion and comparatively less articles which talk about sports, this might also be the reason, they might not be in the features.

A discussion of potential changes to your approach that might improve the accuracy, recall, and/or precision, and/or the reasonableness/usefulness of the approach.

1. Try doing LDA with genism or any other potential method.
2. Working more on input data, if we understand the data more then we can further understand the topics easily.
3. Working more on annotation guide and choosing topics may resolve the issue

Here I have generalized the topics more, because there are more subtopics in political, considering subtopics as new topics might help to achieve good results.

**Annotation guide:**

Here there are total 30 articles and all the articles are discussing about different topic which are:  
1. First article discusses about baseball, pitching, Mets these are terms related to baseball and this has been classified as sports.

2 Second article discuss about mayor bill, resignation, city hall commitment and these words are referred to political.

3 Third article discuss about gunman, killing, early morning firing so this is categorized as police action

4 This article refers to an article about an apple store which has similar name as apple brand and its business after inventing apple product, so this is related to business.

5 This one speaks about swimming pool rules that has been kept by government, so this is political.

6 This specifies about Olympics- sports

7 Lionsgate and its dealing with Starz this is referring business.

8 This refers to matters of swimming pool and its rules- government

9 winter league- this refers to sports

10 This refers to politics

11 Libya and its gun fires and about ISIS so this discusses about terrorism.

12 This discusses about a pastor and gun control so this is referred as police action

13 bribing a judge this refers to political.

14 turkey attacks- ISIS so this is considered as terrorism

15 venture capitol and its funds release for K&P company, so this is referred to business

16 this article is on emails between Hillary and Huma so this is political

17 turkey and ISIS fighting so terrorism.

18 government and school policies and information referred as political.

19 Article about Bernie sanders this discusses about political

20 This is about yankees- baseball so sports

21 elections/republicans/votes changing due to climate change- Political

22 shopping/ sale this refers to fashion

23 cycle race and stats on that it refers to sports

24 European soccer and its league so sports

25 Murder case of a violinist so its related to crime and terrorism

26 extradition of Leora this is crime and terrorism

27 Residents of Kansas election results refers to political

28 Saudi Arabia human rights council, so referred as political

29 travel and new places discovered- fashion/travel

30 shopping mall opening and its budget referred as fashion/travel.

And few words I considered as words required for these topics to be present/ I have worked on only top few words:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SPORTS** | **BUSINESS** |  | **TERRORISM AND POLICE ACTIVITIES** | **POLITICAL AND ELECTIONS.** | **LIFESTYLE, FASHION, MUSIC** |  |
|  | Tampa-bay  Baseball  Younger  90th min-goal  Math  Teixeira  Began- focusing  Smith  Summers ago  Murray  Beltran  Game-Albania  ﻿decidedly unbowed  ﻿proud  ﻿comebacks  absenteeism  ﻿inconsistent  ﻿50one  ﻿berth every team  ﻿Alex Rodriguez  ﻿quarterfinals European  ﻿making Olympic  ﻿Barbara boxer  ﻿leads innings  ﻿best baseball  ﻿validates yankees  ﻿Hornacek | Trading  Account  Attributed  Department-business  pedagogy issue  ﻿ ﻿attorney  ﻿Deal  ﻿Firm  Documents from  ﻿Housing  ﻿Advocate  ﻿fund responds  Firm expansion |  | Bombing  Government Islamic  Sessions  Islamic -Housing  Eliminating  Custody  Wanted  Terrorist  Bone-spur  Warned  City wide  Doghole  165 students  ﻿pedagogy issue  ﻿main Syrian  ﻿troubled Brooklyn  ﻿baton said  ﻿allow airstrikes  ﻿Bullet  ﻿Suspects  ﻿Mediation  Fighter  ﻿MR. Arbitrager  ﻿throw investigators  ﻿getting leads  ﻿Isis related  ﻿constructive force  increase terrorist  Bipartisan  ﻿shooting | Gonzales  News-bulletins  Event  Documents  Election  Gregarious  Glaring  Losing-dedicated  Withdrawn-presidential  State management  Public assistance  Blasio’s -school  Democratic addition  InterVision’s students  Committed-supporting  Potential-candidate  Members-congress  Medical aid  Politicians  American  Royals protecting  Western officials  ﻿substantial improvement  ﻿foreign service  ﻿District- ﻿representatives  ﻿campaign Mrs.  ﻿Clayton  ﻿ ﻿labor pensions  ﻿ ﻿fellow senators let  ﻿Bashar  ﻿Position  ﻿Mr. Espaillat  ﻿race really  Governors  ﻿president Obama  ﻿president spent  ﻿trump Patrick Kennedy  ﻿Clintons campaign | Music  Genetically  Fashion  Offers  Lewis-howes  Outskirts  Footprints  Travel  ﻿package stays  ﻿Frayed  ﻿customers apple  ﻿does endorse  ﻿Sunday  ﻿traveling abrades  ﻿stay beach  ﻿little sore  ﻿men women  ﻿classroom music  ﻿bodegas discount  ﻿beach resort  ﻿music room |  |

References:

1. The pdf given in course material about rule-based systems, textbook about rule-based approaches.

2.https://medium.com/friendly-data/machine-learning-vs-rule-based-systems-in-nlp-5476de53c3b8\

3.https://en.wikipedia.org/wiki/Topic\_model

4.https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

5.https://www.linkedin.com/pulse/lda-explanation-gaurhari-dass/

6.https://monkeylearn.com/blog/introduction-to-topic-modeling/#:~:text=Topic%20modeling%20is%20an%20unsupervised,characterize%20a%20set%20of%20documents.

7.https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

8.https://medium.com/swlh/understanding-count-vectorizer-5dd71530c1b

9. Dataset has been considered from Kaggle, and here is the link of reference,

https://www.kaggle.com/nzalake52/new-york-times-articles