NEWS ARTICLE CLUSTERING USING TOPIC MODELLING APPROACHES

Sarthak S *Jagadish Rathod* *Sameer Kulkarni*

sarthakshivaraju@gmail.com rathodjagadish92@gmail.com sameer6kulkarni@gmail.com

Abstract - In many cases labelling the news articles according to labels by reading long unstructured news for supervised learning will consume large amounts of time and effort. Hence clustering news using unsupervised learning becomes very useful for these news organisations. Topic modeling is an unsupervised approach which is used to find topics from unstructured text. Here we are used LDA and NMF topic modelling Algorithms.

Keywords – Document Clustering, LDA, NMF, K-Means clustering, TF-IDF, News Article

INTRODUCTION

In this report we have built a model that performs the Clustering of the Documents using LDA and NMF Topic modelling methods. Our project helps many news organisations in the world to cluster similar news articles together. It helps to save time by reading long news articles and then grouping articles into categories. LDA model is a probabilistic based model that can be used for document clustering. Initially we randomly assign the number of topics to this algorithm. Then the documents are randomly assigned to topics and each word in the document is randomly assigned to every topic as a probability based on Dirchlet Allocation.

NMF is used for topic modelling that is based on dimensionality reduction technique.It decomposes matrix V into matrix U and W. For implementation we used python and imported required necessary modules.

PREVIOUS WORK

As a previous work some of the researchers used different methods for topic modelling.In all of the work first document is converted to either BOW or TF - IDF representation. Fiona Martin and Mark Johnson used only noun word in the corpus and topic modelling is performed using Latent Dirichlet Allocation (LDA) algorithm. This method reduced number of junk topics as well as time to train the model. But topic coherence used on top 10 words does not validate if they truly represent the themes across corpus.In Leveraging Unstructured Information

Using Topic Modelling[2] author demonstrated basic working of LDA which was useful to understand the basics of LDA.Avashlin,Vukosi[4] and Sara,Habib[5] compared two popular topic modelling algorithms named LDA and NMF based on coherence score. But author assumed fixed number of topics which can’t be optimal number all the time.

Dhendra,Sunarna[3] used k-means clustering with elbow method to find optimal number of topics and perform topic modelling. Optimal number of cluster is decided based on SSE(Sum of Squared Errors). The fault with this method is K-means does not work well with clusters of different size and density and Different initialization of centroids can lead to different final clusters. Sigit,Yuita,M.Ali,Putra [6] performed K-means on seeds which were selected based on pillar algorithm. This overcomes the problems of K-means and perform as optimized K-means.

DATASET

A total of 12008 News articles without any labels were taken from NPR News Dataset for our analysis of clustering using topic modelling.

TOPIC MODELLING

1. Latent Dirchlet Allocation:

LDA model is a probabilistic based model that can be used for document clustering. Initially we randomly assign the number of topics to this algorithm. Then the documents are randomly assigned to topics and each word in the document is randomly assigned to every topic as a probability based on Dirchlet Allocation.

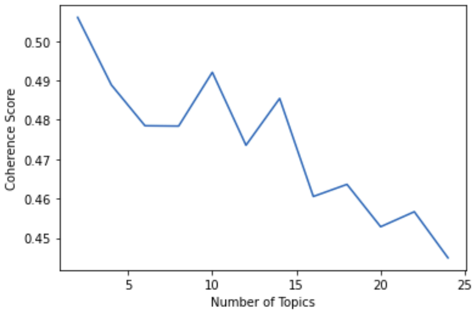
This algorithm works by finding the probability of each document belonging to a topic and the probability of each word belonging to a topic. This process continues for a number of iterations till we get a set of topics with acceptable related words in it.

DATA PREPROCESSING

All words having document frequency less than 2 and greater than 95% of the documents are removed. Stop words are also removed.
News documents are converted to lower-case. Lemmatization is being applied to remove set of words with common suffixes and to get the root form of the word.
Bag of words model has been used for word vector representation for LDA since it is a probabilistic model.

MODEL

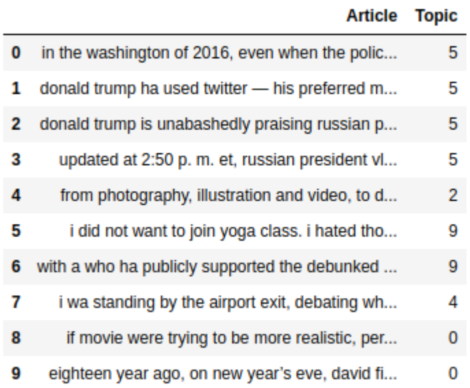
Since LDA Model requires the user to pre-define the number of topics, we used coherence metric to identify the optimal number of topics required for our analysis.
Coherence is calculated by finding the top words of the topic and finding the relation between each pair of the top words and then finding the mean for all topics of the model.
C_v measure from Gensim library is being used as coherence measure which uses normalised pointwise mutual information and cosine similarity based on one-set segmentation of top words and sliding window.
We calculated the coherence measure for the number of topics ranging from 2 to 24 as shown in the graph below for LDA Model using Gensim library.



From the above graph, we took 10 as optimal number of topics since it reasonably has a good coherence score and there is a dip in coherence score as we move after 10 topics.
We then built our final LDA Model using 10 topics using Sklearn library.

The top 15 words for each of the 10 topics are found.

Later the probability of each document belonging to a topic is found using LDA and the topic having highest probability is assigned to that document for clustering the news articles.



	Article	Topic
0	in the washington of 2016, even when the polic...	5
1	donald trump ha used twitter — his preferred m...	5
2	donald trump is unabashedly praising russian p...	5
3	updated at 2:50 p. m. et, russian president vl...	5
4	from photography, illustration and video, to d...	2
5	i did not want to join yoga class. i hated tho...	9
6	with a who ha publicly supported the debunked ...	9
7	i wa standing by the airport exit, debating wh...	4
8	if movie were trying to be more realistic, per...	0
9	eighteen year ago, on new year's eve, david fi...	0

2. Non negative Matrix Factorization

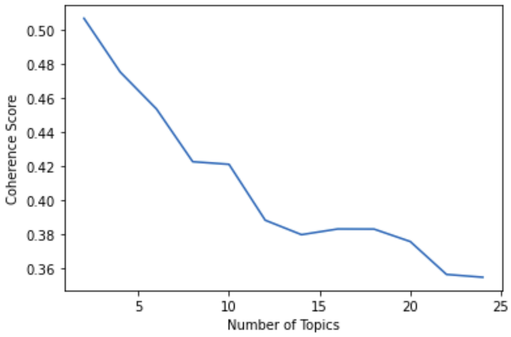
NMF is used for topic modelling that is based on dimensionality reduction technique. Given a matrix V(n x m), it decomposes into two matrices U(n x k) and W(k x m) where k is the number of topics. Here V is the document term matrix, U is document topic matrix and W is topic term matrix. The algorithm works by repetitively updating U and W to find the best set of co-efficients.

DATA PRE-PROCESSING

The same pre-processing steps applied for LDA Model are being applied for NMF Model except using Term Frequency Inverse Document Frequency (TF-IDF) for word vector representation.

MODEL

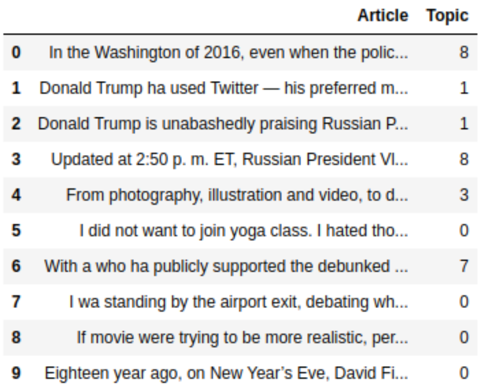
We again calculated the coherence measure for the number of topics ranging from 2 to 24 as shown in the graph below for NMF Model using Gensim library.



From the above graph, again we took 10 as optimal number of topics since it reasonably has a good coherence score and there is a dip in coherence score as we move after 10 topics.

We then built our final NMF Model using 10 topics using Sklearn library. The top 15 words for each of the 10 topics are found.

Later the probability of each document belonging to a topic is found using NMF and the topic having highest probability is assigned to that document for clustering the news articles.



	Article	Topic
0	In the Washington of 2016, even when the polic...	8
1	Donald Trump ha used Twitter — his preferred m...	1
2	Donald Trump is unabashedly praising Russian P...	1
3	Updated at 2:50 p. m. ET, Russian President Vl...	8
4	From photography, illustration and video, to d...	3
5	I did not want to join yoga class. I hated tho...	0
6	With a who ha publicly supported the debunked ...	7
7	I wa standing by the airport exit, debating wh...	0
8	If movie were trying to be more realistic, per...	0
9	Eighteen year ago, on New Year's Eve, David Fi...	0

CONCLUSION

Comparing both the models LDA and NMF, we can see that the coherence of LDA model is better than that of NMF model for various numbers of topics.
Hence, we conclude that LDA model is relevant than NMF model for our application of news article clustering.

REFERENCES

i. Fiona Martin,Mark Johnson. More Effificient Topic Modelling Through a Noun Only Approach,2019.

ii. J.W. Uys,N.D. du Preez,E.W. Uys.Leveraging Unstructured Information Using Topic Modelling,August 2008.