SENTIMENT ANALYSIS & TOPIC MODELING ON NEWS ARTICLES

A Mini-project report Submitted to the Department of Computer Applications,

Bharathiar University, in the partial fulfilment of the requirements for the Award of degree of

Master of Science in Data Analytics

Submitted by

MARELLA VENU GOPAL REDDY

(18CSEG027)

Under the guidance of

Mr.K.MOORTHY, M.C.A.,

DEPARTMENT OF COMPUTER APPLICATIONS

SCHOOL OFCOMPUTERSCIENCEANDENGINEERING

BHARATHIAR UNIVERSITY

COIMBATORE-641 046

NOVEMBER-2019

DECLARATION

DECLARATION

I hereby declare that this mini-project work titled, "**SENTIMENT ANALYSIS & TOPIC MODELING ON NEWS ARTICLES**" submitted to Department of Computer Applications, Bharathiar

University, in partial fulfilment of the requirements for the award of the degree of **Master of Science in Data Analytics,** is a record of original work done by me, under the supervision and guidance of **Mr.K.MOORTHY, M.C.A.,**Department of Computer Applications, Bharathiar University, and that this project work has not formed the basis for the award of any Degree /Diploma /Associateship /Fellowship or similar title to any candidate of any University.

Place : Coimbatore Signature of the Candidate

Date : (M. VENU GOPAL REDDY)

Countersigned by

Project Guide

CERTIFICATE

CERTIFICATE

This is to certify that, this mini-project work entitled, "**SENTIMENT ANALYSIS & TOPIC MODELING ON NEWS ARTICLES**" submitted to Bharathiar University, in partial fulfilment of the requirements for the award of the degree of **Master of Science in Data Analytics,** is a record of original work done by **M. VENU GOPAL REDDY (18CSEG027)**, during his period of study in the Department of Computer Applications, Bharathiar University, Coimbatore, under my supervision and guidance and that this project work has not formed the basis for the award of any Degree /Diploma /Associateship /Fellowship or similar title to any candidate of any University.

Place : Coimbatore

Date :

Project Guide

Head of the Department

Submitted for the Project VIVA-VOCE Examination held on _____

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I express my respectful thanks to our Professor & Head of the Department, **Dr. T. DEVI, M.C.A., M.Phil., Ph.D. (UK),** Department of Computer Applications, Bharathiar University, for permitting me to carry out my mini project work in "**SENTIMENT ANALYSIS & TOPIC MODELING ON NEWS ARTICLES**".I express heart-felt gratitude to my project guide **Mr.K.MOORTHY, M.C.A, M.phil**, Department of Computer Applications, Bharathiar University, for his valuable support during research work.

I would like to heartily thank all respected faculty members, laboratory staff of Department of Computer Applications, Bharathiar University, Coimbatore- 641046.

I thank my parents for their encouragement and moral support. I thank everybody whose names are mentioned and not mentioned here, who has directly or indirectly contributed in bringing out this project successfully.

ABSTRACT

The project is entitled "**SENTIMENT ANALYSIS & TOPIC MODELING ON NEWS ARTICLES".** Dataset is taken from kaggle repository it contains American publications articles and their authors. Articles are in unstructured format. We are using text summarization, sentiment analysis and topic models. News articles plays crucial role in society, these articles will effects the peoples opinions. This will helps you to know that the articles are in positive, negative or neutral.

Sentiment analysis refers to the use of natural language processing, text analysis, extract, quantify, and subjective information. Using sentiment analysis to find positive, negative, neutral values by using sentiment intensity analyzer, From that sentiment intensity analyzer to find the polarity of the sentiment of words.

Topic modelling is a powerful technique for unsupervised analysis of large document collection. Topic models conceive latent topics in text using hidden random variables, and discover that structure with posterior inference. Topic models have a wide range of applications like tag recommendation, text categorization, keyword extraction and similarity search in the board fields of text mining, information retrieval, statistical language modelling.

In this project, a dataset with 50,000 articles collected from 2016&2017 of American publications. It is difficult to find out each topic from the documents, to make clusters for each topic and finding how words are frequently occurred. Based on the word frequency it create clusters. The document models are built using LDA (Latent Dirichlet Allocation), LSI(Latent Semantic Analysis) and HDP (Hierarchical Dirichlet process). The Performance of the built models are analysed by the evaluation measure perplexity and cohariance. Based on the cohariance score we will decide which model has formed clusters in better way.

II

TABLE OF CONTENTS

TABLE OF CONTENTS

1. INTRODUCTION

1.1 MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human interaction or assistance and adjust actions accordingly.

1.2 PYTHON

Python is a popular platform used for research and development of production systems. It is a vast language with number of modules, packages and libraries that provides multiple ways of achieving a task. Python and its libraries like NumPy, SciPy, Scikit-Learn, Matplotlib are used in data science and data analysis. They are also extensively used for creating scalable machine learning algorithms. Python implements popular machine learning techniques such as Classification, Regression, Recommendation, and Clustering.

Python offers ready-made framework for performing data mining tasks on large volumes of data effectively in lesser time. It includes several implementations achieved through algorithms such as linear regression, logistic regression, Naïve Bayes, k-means, K nearest neighbor, and Random Forest.

1.3 BIG DATA

Big Data is also data but with a huge size. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

- Structured
- Unstructured
- Semi-structured

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.

Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

1.4 ARTICLE

Article is a piece of writing usually intended for publication in a newspaper, magazine, or journal. It is written for a wide audience, so it is essential to attract and retain the reader's attention. It may include amusing stories, reported speech and descriptions. It can be formal or informal, depending on the target audience. It should be written in an interesting or entertaining manner. It should give opinions and thoughts, as well as facts. It is in a less formal style than a report.

1.4.1.News article

A news article is an article published in a print or internet news medium such as a newspaper, newsletter, news magazine, news-oriented website, or recent news directory that discusses current or recent news of either general interest(i.e. daily newspapers) or on a specific topic(i.e. political or trade news magazines, club newsletters, or technology news websites).

1.4.2. What makes a good news article?

A good story about something the audience decides is interesting or important. A great story often does both by using storytelling to make important news interesting. The public is exceptionally diverse. Though people may share certain characteristics or beliefs, they have an untold variety of concerns and interests. So anything can be news. But not everything is newsworthy. News articles are storytelling to make a subject newsworthy.

1.4.3 News article format:

A typical news article contains five parts:

1. Headline : This is a short, attention-getting statement about the event.
2. Byline : This tells who wrote the story.
3. Lead paragraph: This has all of the who, what, when, where, why and how in it. A writer must find the answers to these questions and write them into the opening sentence of the article.
4. Explanation : After the lead paragraph has been written, the writer must decide what other facts or details the reader might want to know. The writer must make sure that he/she has enough information to answer any important questions a reader might have after reading the headline and the lead paragraph. This section can also include direct quotes from witnesses or bystanders.
5. Additional Information: This information is the least important. Thus, if the news article is too long for the space it needs to fill, it can be shortened without rewriting any other part. This part can include information about a similar event.

1.5 TEXT ANALYTICS

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

The term text analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation. The term is roughly synonymous with text mining; indeed, Ronen Feldman modified a 2000 description of " text mining" in 2004 to describe "text analytics".

1.5.1 Unstructured text data

Unstructured data (or unstructured information) is information that either does not have a pre-data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well.

Techniques such as data mining, natural language processing (NLP), and text analytics provide different methods to find patterns in, or otherwise interpret, this information. Common techniques for structuring text usually involve manual tagging with metadata or part-of-speech tagging for further text mining-based structuring.

The unstructured text collected from social media activities plays a key role in predictive analytics for the enterprise because it is a prime source of sentiment analysis to determine the general attitude of consumers towards a band or idea.

1.5.2 TEXT PREPROCESSING

Preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process.

Converting to lowercase : Convert text to lowercase format using str.lower () to convert all text into lowercase.

Removing numbers : Remove numbers if they are not relevant to your analyses. Usually, regular expressions are used to remove numbers.

Removing punctuation : The following code removes this set of symbols [!”#$%&’()*+, /:;<=>?@[\]^_{\}~].

Removing whitespaces : To remove leading and ending spaces, use the strip () function.

Lemmatization : Lemmatization reduces words to their base word, which is linguistically correct lemmas. It transforms root word with the use of vocabulary and morphological analysis. Lemmatization is usually more sophisticated than stemming. Stemmer works on an individual word without knowledge of the context.

Tokenization : Tokenization is the first step in text analytics, the process of breaking down a text paragraph into smaller chunks such as words or sentence is called Tokenization. Token is a single entity that is building blocks for sentence or paragraph.

Stopwords : Stopwords considered as noise in the text. Text may contain stop words such as is, am, are, this, a, an, the, etc.

Lexicon Normalization : Lexicon normalization considers another type of noise in the text. For example, connection, connected, connecting word reduce to a common word “connect”. It reduces derivationally related forms of a word to a common root word.

Stemming : Stemming is a process of linguistic normalization, which reduces words to their word root word or chops off the derivational affixes. For example, connection, connected, connecting word reduce to a common word "connect".

POS Tagging : The primary target of part-of-speech (POs) tagging is to identify the grammatical group of a given word. Whether it is a NOUN, PRONOUN, ADJECTIVE, VERB, ADVERBS, etc. based on the context.

Word cloud : An image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

1.6 SENTIMENT ANALYSIS

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Polarity

Polarity in sentiment analysis refers to identifying sentiment orientation (positive, neutral, and negative) in written or spoken language. ... Language can contain expressions that are objective or subjective.

1.7 SYSTEM ENVIRONMENT

The following are the hardware and software specifications used in the mini project development

Software specifications

Language	:	Python
Tool	:	Jupyter Notebook

Hardware specifications

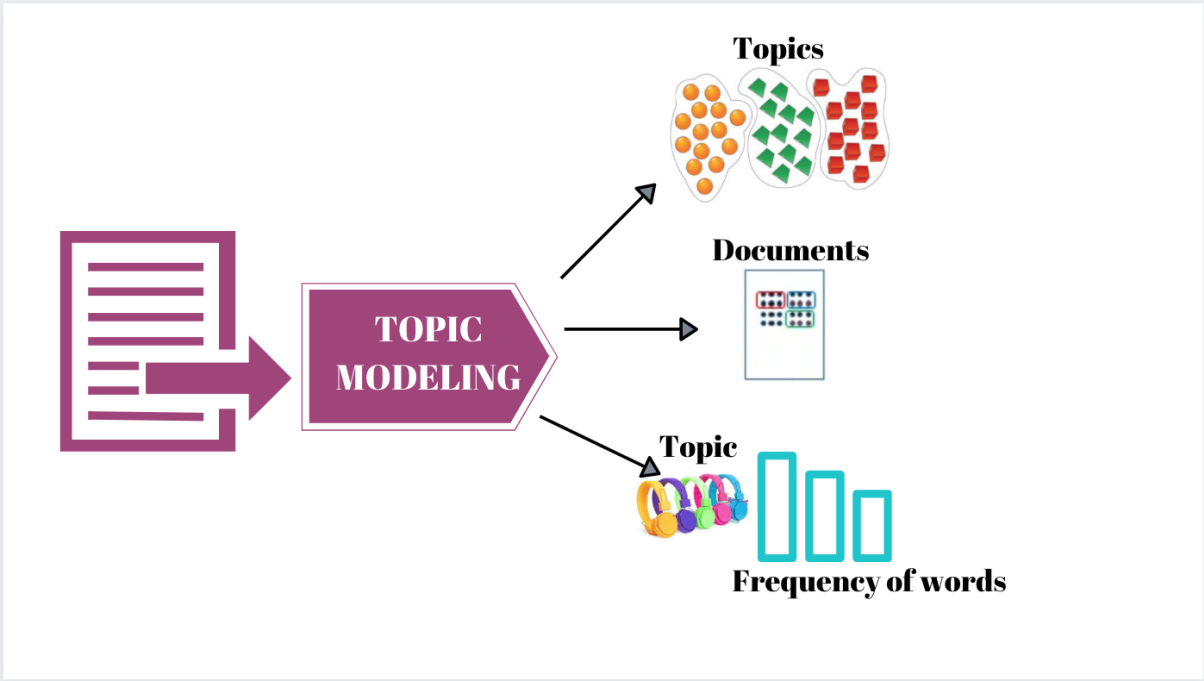
Processor	:	Intel core i3
RAM	:	4GB RAM
Hard disk drives	:	1TB
Monitor	:	15 inches colour monitor

2. TOPIC MODELING

2.1 TOPIC MODELING

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 modelling 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. "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words.

The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.



2.2 LATIN DIRICHLET ALLOCATION (LDA)

Topic modelling is a type of statistical modelling for discovering the abstract “topics” that occur in a collection of documents. Latin Dirichlet Allocation (LDA) is an example of

topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

2.3 LATENT SEMANTIC INDEX (LSI)

LSA (Latent Semantic Analysis) also known as LSI (Latent Semantic Index) LSA 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. LSA learns latent topics by performing a matrix decomposition on the document-term matrix using Singular value decomposition. LSA is typically used as a dimension reduction or noise reducing technique.

2.4 HIERARCHICAL DIRICHLET PROCESS (HDP)

In statistics and machine learning, the hierarchical Dirichlet process (HDP) is a nonparametric Bayesian approach to clustering grouped data. It uses a Dirichlet process for each group of data, with the Dirichlet processes for all groups sharing a base distribution which is itself drawn from a Dirichlet process.

The HDP mixture model is a natural nonparametric generalization of Latent Dirichle allocation, where the number of topics can be unbounded and learnt from data. Here each group is a document consisting of a bag of words, each cluster is a topic, and each document is a mixture of topics. The HDP is also a core component of the infinite hidden Markov model, which is a nonparametric generalization of the hidden Markov model allowing the number of states to be unbounded and learnt from data.

2.5 PYTHON LIBRARIES

Pandas & Numpy : Pandas used to perform in dataframe & Numpy is used for linear algebra method.

```
Import pandas as pd

Import numpy as np
```

Visualization Packages : Visualization packages to plot the word frequency method.

```
Import matplotlib.pyplot as plt

Import seaborn
```

import pyLDAvis

import pyLDAvis.gensim

NLP& NLTK : Using Natural Language Processing used for text document it analysis large amount of natural language.

Import NLP

Import NLTK

Stopwords Removal : To remove the stopwords from clean the dataset from documentation method.

From nltk.corpus import stipwords

From nltk.stem import WordNetLemmatizer

Sentiment Analyzer : To calculate the sentiment polarity by using sentiment intensity analyzer.

From nltk.sentiment.vader import sentiment Intensity analyser

2.6 TOPIC MODELLING LIBRERIES

Text summarization involves generating a summary from a large body of text which Somewhat describes the context of the large body of text. Text summarization using different library.

Gensim Library : Gensim offer used for text rank summarization method. It is based on Page Rank algorithm.

import gensim

Test Teaser : Text Teaser is an automatic summarization algorithm that combines the power of natural language processing and machine language to produce good results.

From textteaser import TextTeaser

PyTeaser : PyTeaser takes any news article and extracts a brief summary from it.

From Pyteaser import summarizeUrl

Latent Dirichlet Allocation : Using Latent Dirichlet Allocation used for Topic modeling from documentation to summarize to select the topics.

from gensim.models import LdaModel

Latent Semantic Index (LSI)

from gensim.models import LsiModel

Hierarchical Dirichlet process (HDP)

from gensim.models import HdpModel

Coherence score

from gensim.models import CoherenceModel

3. OBJECTIVE AND METHODOLOGY

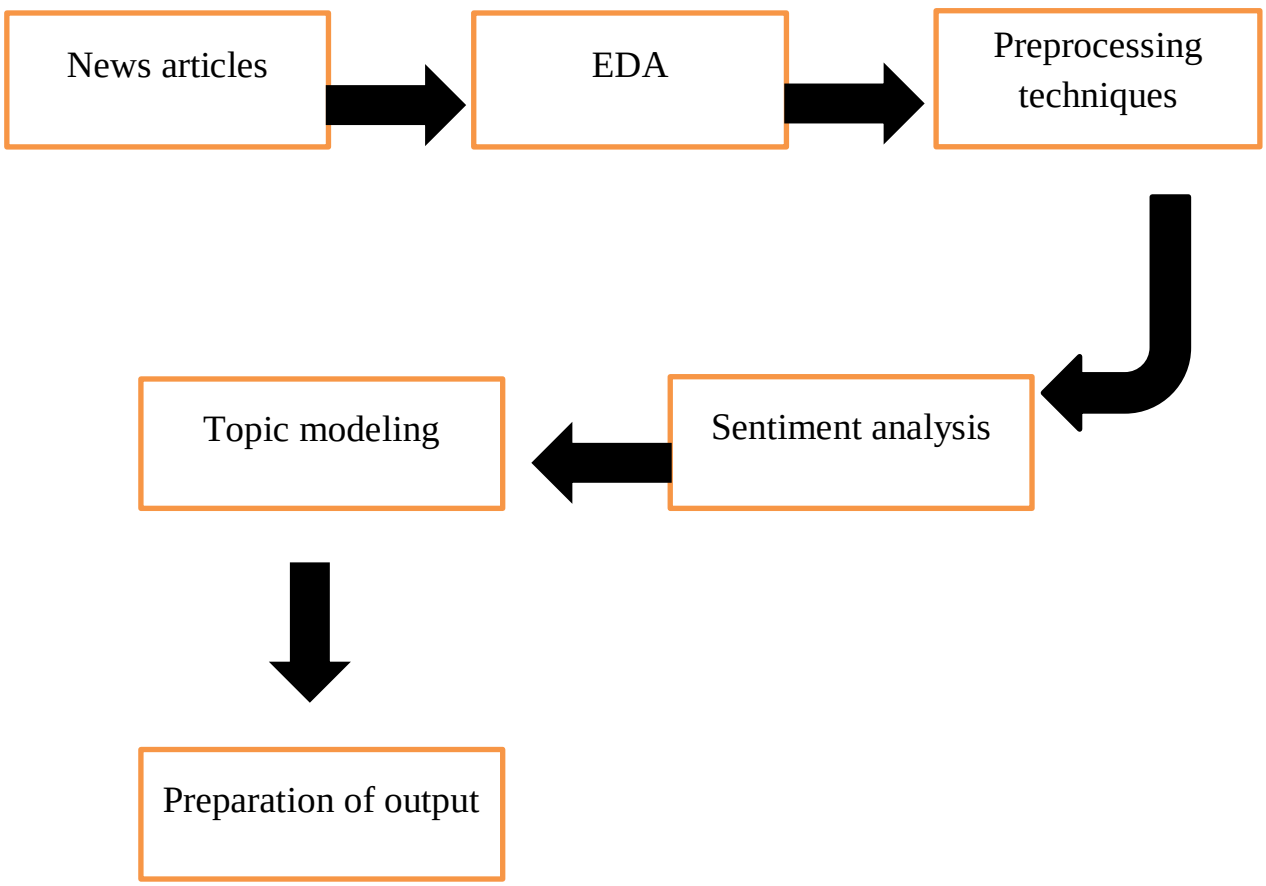
3.1 OBJECTIVE

Objective of this project is to perform topic modelling on news articles and sentiment analysis. The specific objective of this project is

- Perform preprocessing on the data.
- Do exploratory data analysis(EDA).
- Perform sentiment analysis to characterize the polarity of reviews.
- Perform topic modelling (LDA, LSI, GDP…..)

3.2 METHODOLOGY

These are the 5 steps to analyse the news articles and here is the graphical representation of methodology.



3.3 DATASET DESCRIPTION

The dataset that I have taken is Articles from 15 american publications. I wanted to see how articles clustered together if the articles were rendered into document-term matrices---would there be greater affinity among political affiliations, or medium, subject matter, etc. The publications include the New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Reuters, Vox, and the Washington Post.

Sampling wasn't quite scientific; I chose publications based on my familiarity of the domain and tried to get a range of political alignments, as well as a mix of print and digital publications. The data primarily falls between the years of 2016 and July 2017, although there is a not-insignificant number of articles from 2015, and a possibly insignificant number from before then.

Attribute description:

Id : Database ID

Title : Article title

Publication : Publication name

Author : Author name

Date : Date of publication

Year : Year of publication

Month : Month of publication

url : URL for article (not available for all articles)

content : Article content

3.3.1 Head and tail of the dataset

	Unnamed: 0	id	title	publication	author	date	year	month	url	content
0	0	17283	House Republicans Fret About Winning Their Hea...	New York Times	Carl Hulse	2016-12-31	2016.0	12.0	NaN	WASHINGTON — Congressional Republicans have...
1	1	17284	Rift Between Officers and Residents as Killing...	New York Times	Benjamin Mueller and Al Baker	2017-06-19	2017.0	6.0	NaN	After the bullet shells get counted, the blood...
2	2	17285	Tyrus Wong, 'Bambi' Artist Thwarted by Racial ...	New York Times	Margalit Fox	2017-01-06	2017.0	1.0	NaN	When Walt Disney's "Bambi" opened in 1942, cri...
3	3	17286	Among Deaths in 2016, a Heavy Toll in Pop Musi...	New York Times	William McDonald	2017-04-10	2017.0	4.0	NaN	Death may be the great equalizer, but it isn't...
4	4	17287	Kim Jong-un Says North Korea Is Preparing to T...	New York Times	Choe Sang-Hun	2017-01-02	2017.0	1.0	NaN	SEOUL, South Korea — North Korea's leader, ...

	Unnamed: 0	id	title	publication	author	date	year	month	url	content
0	53293	73471	Patriots Day Is Best When It Digs Past the Her...	Atlantic	David Sims	2017-01-11	2017.0	1.0	NaN	Patriots Day, Peter Berg's new thriller that r...
1	53294	73472	A Break in the Search for the Origin of Comple...	Atlantic	Ed Yong	2017-01-11	2017.0	1.0	NaN	In Norse mythology, humans and our world were ...
2	53295	73474	Obama's Ingenious Mention of Atticus Finch	Atlantic	Spencer Kornhaber	2017-01-11	2017.0	1.0	NaN	"If our democracy is to work in this increasin...
3	53296	73475	Donald Trump Meets, and Assails, the Press	Atlantic	David A. Graham	2017-01-11	2017.0	1.0	NaN	Updated on January 11 at 5:05 p. m. In his fir...
4	53297	73476	Trump: 'I Think' Hacking Was Russian	Atlantic	Kaveh Waddell	2017-01-11	2017.0	1.0	NaN	Updated at 12:25 p. m. After months of equivoc...

4. RESULT FOR AUTHOR-I

4.1. EXPLORATORY DATA ANALYTICS (EDA)

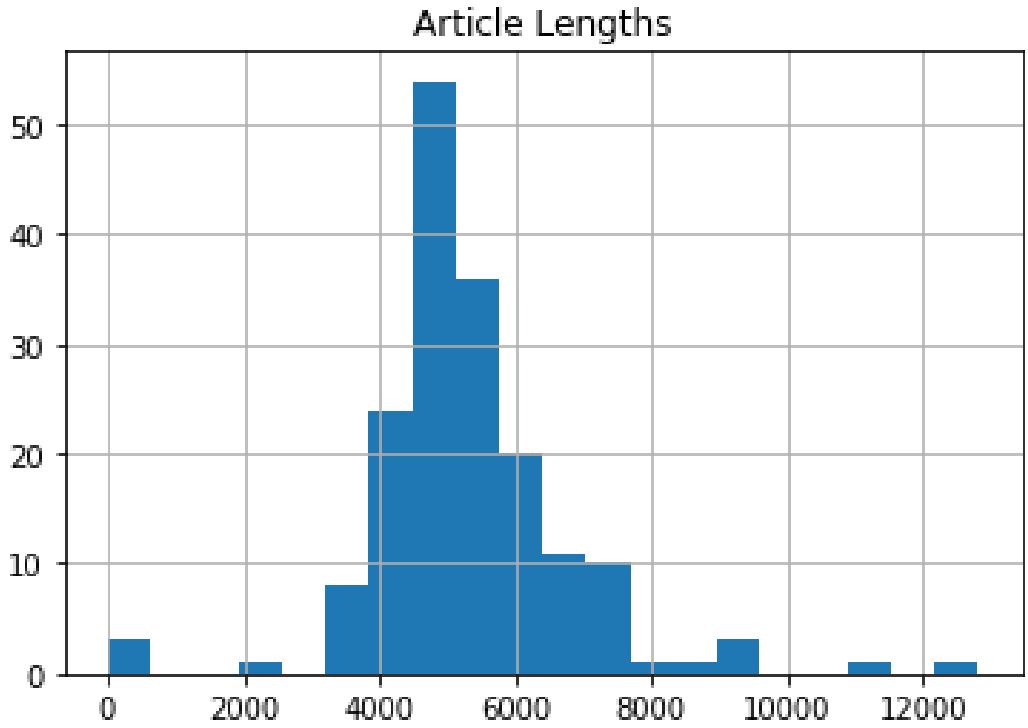


Figure 4.1(a) length of the articles content

Insights:

- From the above graph we can see the length of the articles that written by david sims.
- The plot says nearly 55 articles length lies between 3000 to 5000. There are very few articles written upto 12000 frequency.

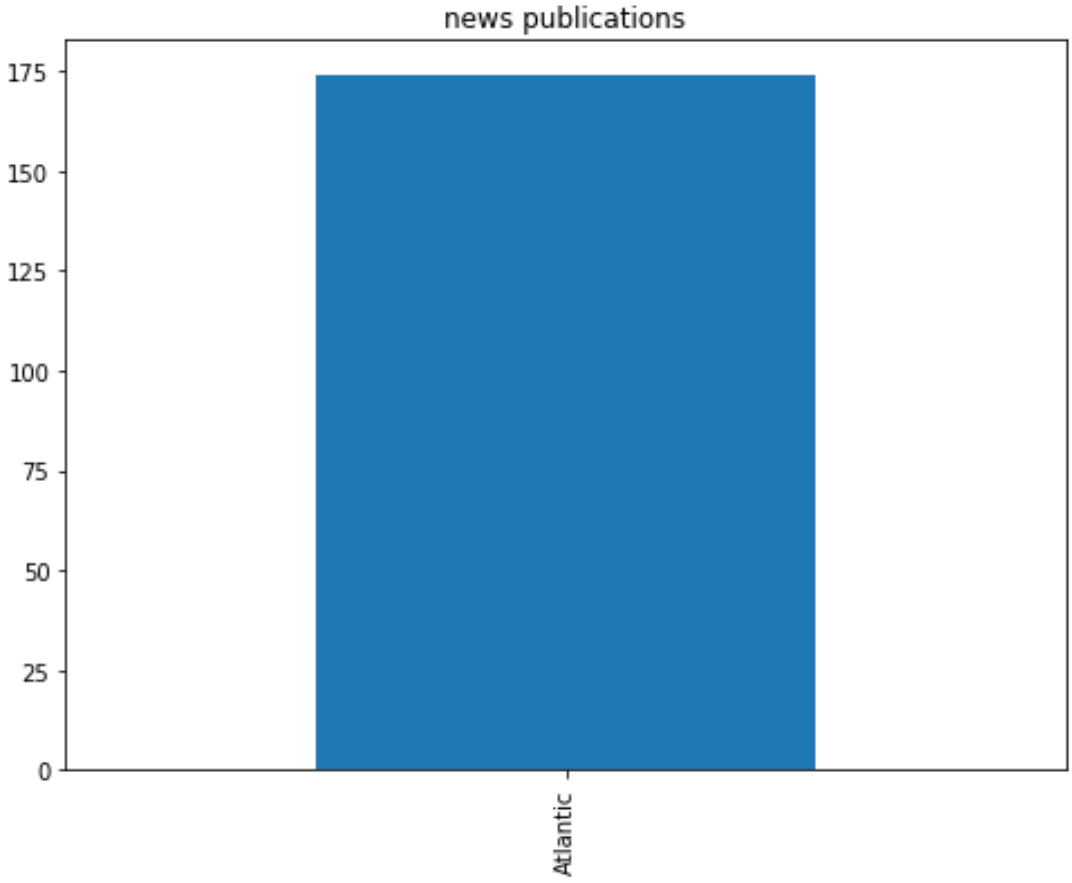


Figure 4.1(b) Bar chart for news publications

Insights:

- The above bar plot shows the count of articles in a particular publications.
- This plot shows that the author published his articles through only one publication that is atlantic news publications.

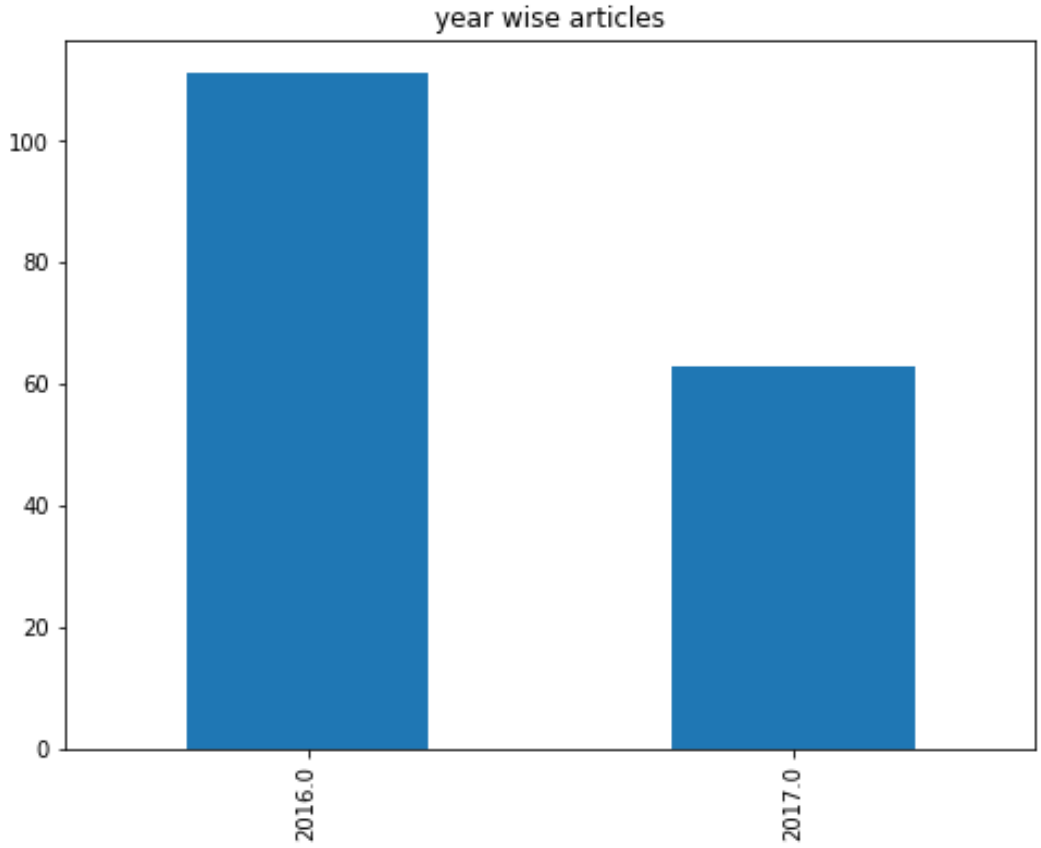


Figure 4.1(c) Bar chart for year wise articles

Insights:

- From the above bar plot we can tell in which year the author published the articles.
- Compare to 2017 most of the articles were published in 2016.
- in 2017 more than 100 articles were published, whereas in 2016 around 65 articles were published.

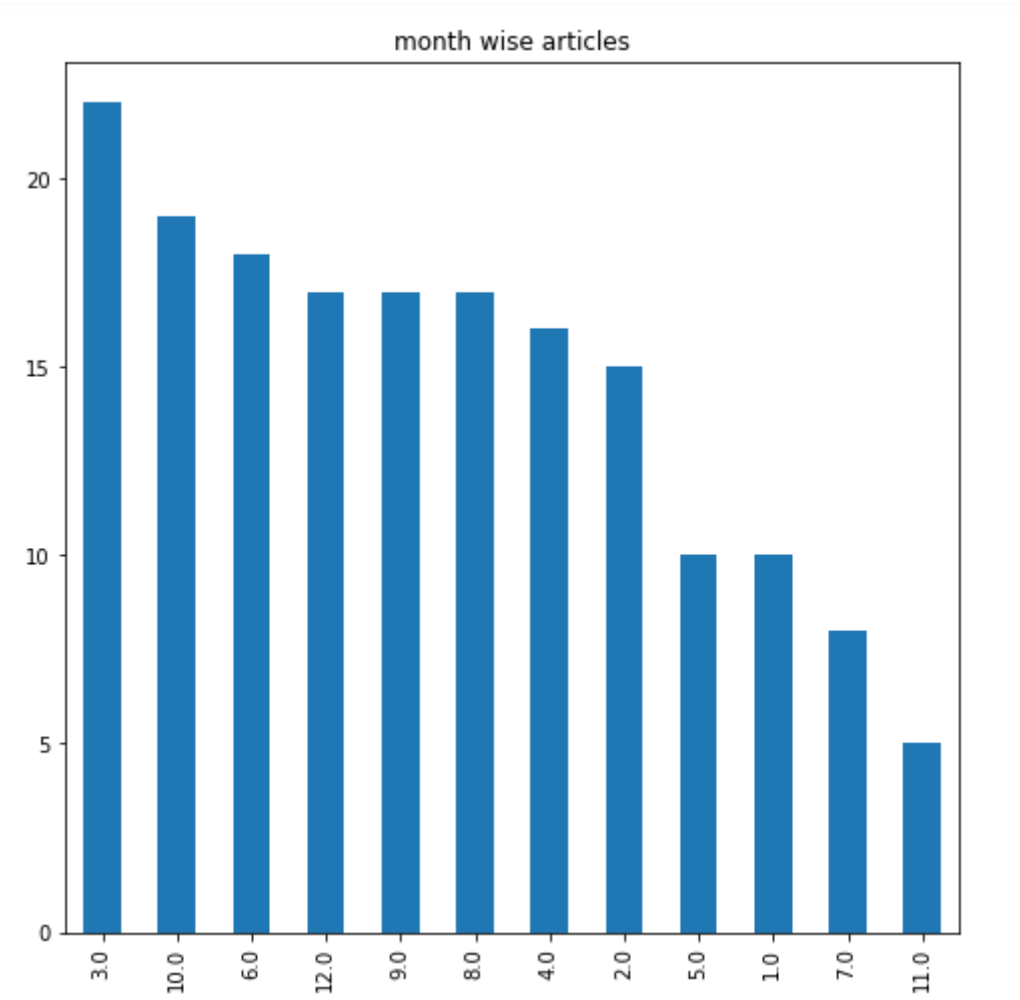


Figure 4.1(d) Bar chart for month wise articles

Insights:

- The above plot show number of articles published in every month.
- From the plot we can say most of the articles published in 3&10 months. In November only very few articles were published.

4.2.SENTIMENT ANALYSIS

4.2.1. Sentiment polarity

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
neutral
```

Insights:

- The above results is about sentiment polarity which tells about the articles sentiment like positive, negative or neutral.
- This author articles are perfectly in neutral, because an article suppose to expose the facts in news.

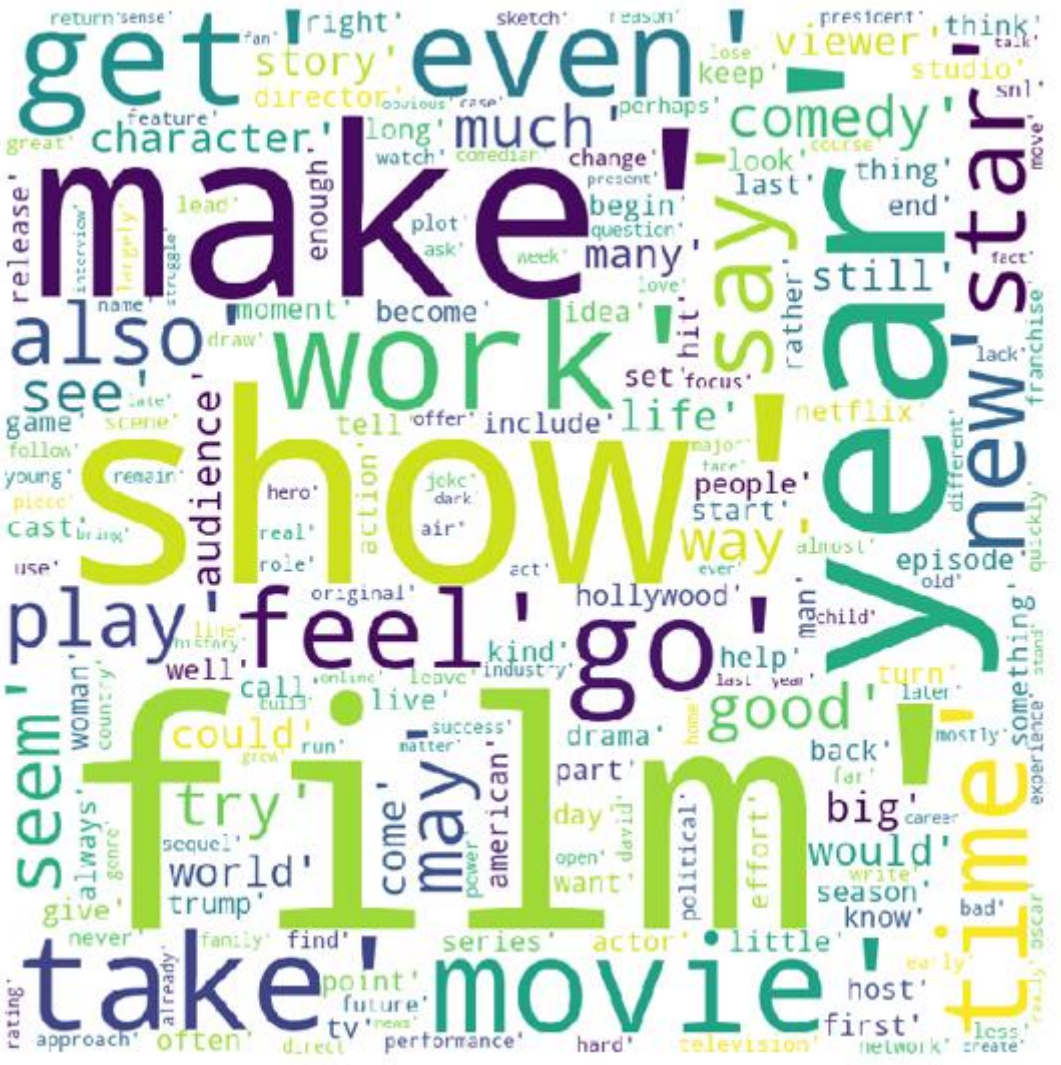


Figure 4.2(a) Word Cloud

Insights:

- The above plot is word cloud, it visualizes the most repeated words in the articles with big size.
- We can see here ‘show’, ‘film’, ‘make’, ‘year’ are very big compared to other words, it means author have used these words frequently in articles.
- In the same way, if words are looking very small means that words are used very rare in the articles.

4.3. MODEL PERFORMANECE

4.3.1. LDA model

```
[(0,
  '0.009*"film" + 0.008*"man" + 0.007*"work" + 0.007*"play" + 0.005*"gap" + '
  '0.005*"woman" + 0.004*"football" + 0.004*"movie" + 0.004*"tom" + '
  '0.004*"make"'),
 (1,
  '0.018*"film" + 0.011*"game" + 0.011*"hollywood" + 0.011*"movie" + '
  '0.010*"studio" + 0.007*"make" + 0.006*"market" + 0.006*"sequel" + '
  '0.006*"franchise" + 0.006*"hit"'),
 (2,
  '0.007*"film" + 0.007*"isis" + 0.007*"grave" + 0.005*"military" + '
  '0.004*"mosul" + 0.004*"fisher" + 0.004*"star_war" + 0.004*"state" + '
  '0.004*"government" + 0.003*"iraq"'),
 (3,
  '0.009*"zemecki" + 0.006*"song" + 0.006*"simmon" + 0.006*"tallulah" + '
  '0.005*"marshall" + 0.005*"film" + 0.004*"logan" + 0.004*"malick" + '
  '0.004*"walking_dead" + 0.003*"character"'),
 (4,
  '0.009*"film" + 0.008*"game" + 0.007*"lee" + 0.007*"barry" + '
  '0.006*"bioshock" + 0.006*"movie" + 0.006*"player" + 0.005*"star" + '
  '0.005*"big" + 0.004*"manchester"'),
 (5,
  '0.013*"show" + 0.007*"make" + 0.007*"year" + 0.006*"say" + 0.005*"take" + '
  '0.005*"go" + 0.004*"new" + 0.004*"time" + 0.004*"get" + 0.004*"feel"'),
```

```
Perplexity:  -8.477648254987894
```

Insights:

- From the above LDA model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- In results we can see most the words are related to films, so we can say that this David Sims published articles are related to film industry.
- Perplexity is a measure of how good the model is. Lower score indicates the good model.

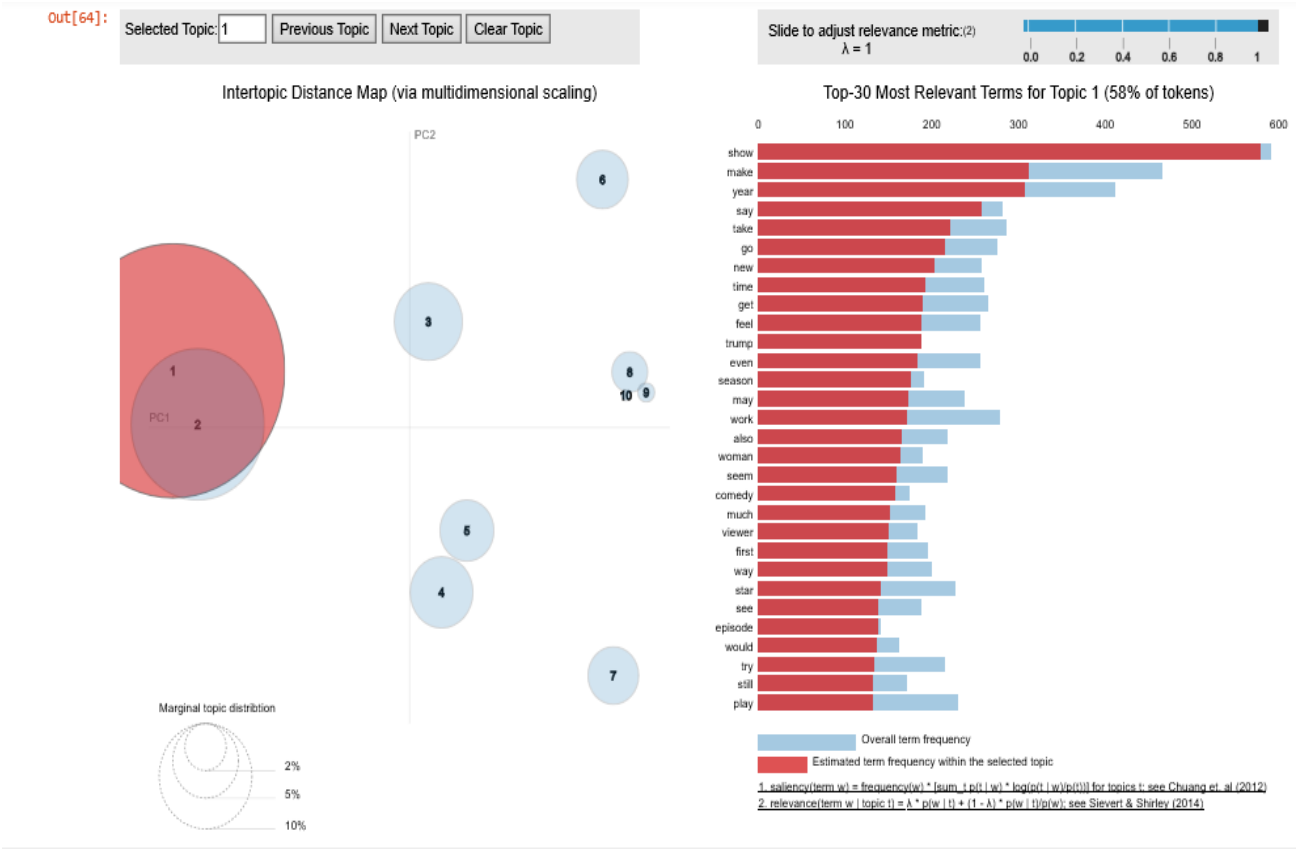


Figure 4.3(a) LDA model

Insights:

- The area of circle represents the importance of each topic over the entire corpus, the distance between the centre of circles indicate the similarity between topics.
- For each topic, the histogram on the right side listed the top 30 most relevant terms.
- I found out that majority of the words do related to one filed. Most of the words like ‘film’, ‘song’, ‘show’, hollywood’ etc.

4.3.3. LSI model

```
[(0,
  '0.346*"film" + 0.256*"show" + 0.209*"make" + 0.204*"year" + 0.139*"movie" + '
  '0.131*"work" + 0.127*"go" + 0.125*"take" + 0.124*"get" + 0.115*"time"'),
 (1,
  '0.561*"film" + -0.537*"show" + 0.201*"movie" + -0.185*"trump" + '
  '-0.146*"season" + -0.123*"episode" + -0.108*"snl" + 0.102*"hollywood" + '
  '-0.101*"woman" + 0.091*"studio"'),
 (2,
  '0.564*"woman" + 0.380*"man" + 0.198*"gap" + -0.175*"show" + -0.142*"season" '
  '+ 0.130*"say" + 0.122*"color" + 0.116*"married" + 0.105*"gender" + '
  '0.104*"trump"'),
 (3,
  '-0.265*"show" + -0.227*"film" + 0.217*"go" + 0.203*"wilson" + '
  '0.161*"character" + 0.153*"get" + 0.147*"want" + -0.144*"woman" + '
  '0.143*"think" + 0.141*"say"'),
 (4,
  '-0.464*"trump" + 0.197*"game" + -0.166*"say" + -0.158*"snl" + '
  '0.156*"season" + 0.150*"man" + -0.147*"year" + -0.143*"film" + 0.136*"tv" + '
  '0.136*"woman"'),
 (5,
  '0.304*"year" + 0.250*"good" + 0.215*"test" + -0.188*"character" + '
  '-0.188*"trump" + 0.163*"student" + 0.163*"college" + 0.155*"go" + '
  '0.154*"rorison" + -0.148*"film"'),
```

Insights:

- From the above LSI model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- In the results we can see most the words are related to films, so we can say that ‘David Sims’ published articles are related to film industry.
- In this model there few articles about trump also but compare to movies other articels are very few.

4.3.4. HDP model

```
[(0,
  '0.008*film + 0.003*make + 0.002*year + 0.002*show + 0.002*movie + 0.002*time + 0.002*take + 0.002*feel + 0.002*get + 0.002*say + 0.002*go + 0.002*barry + 0.002*episode + 0.002*may + 0.002*work + 0.001*character + 0.001*even + 0.001*audience + 0.001*series + 0.001*viewer'),
 (1,
  '0.004*film + 0.002*year + 0.002*say + 0.002*also + 0.002*make + 0.002*get + 0.002*go + 0.002*take + 0.002*new + 0.002*noah + 0.002*movie + 0.002*feel + 0.002*time + 0.002*work + 0.002*even + 0.001*first + 0.001*star + 0.001*try + 0.001*jackie + 0.001*game'),
 (2,
  '0.005*make + 0.004*film + 0.004*show + 0.002*year + 0.002*tv + 0.002*movie + 0.002*star + 0.002*last + 0.002*sketch + 0.002*television + 0.002*hollywood + 0.002*may + 0.002*america + 0.002*series + 0.001*take + 0.001*see + 0.001*time + 0.001*way + 0.001*play + 0.001*go'),
 (3,
  '0.004*show + 0.004*film + 0.003*make + 0.002*studio + 0.002*movie + 0.002*take + 0.002*even + 0.002*year + 0.002*time + 0.002*audience + 0.002*new + 0.002*release + 0.002*hit + 0.002*last + 0.002*wilder + 0.002*get + 0.002*wilmore + 0.002*also + 0.001*go + 0.001*first'),
 (4,
  '0.004*woman + 0.002*david + 0.002*film + 0.002*man + 0.002*snowden + 0.002*year + 0.002*may + 0.001*show + 0.001*say + 0.001*married + 0.001*story + 0.001*feel + 0.001*much + 0.001*new + 0.001*warcraft + 0.001*work + 0.001*point + 0.001*come + 0.001*stone + 0.001*vote'),
 (5,
  '0.002*film + 0.002*show + 0.002*comedy + 0.002*pete + 0.002*metzger + 0.001*make + 0.001*comedian + 0.001*star + 0.001*year + 0.001*johnson + 0.001*many + 0.001*woman + 0.001*would + 0.001*even + 0.001*seem + 0.001*get + 0.001*take + 0.001*robert + 0.001*also + 0.001*ceremony'),
```

Insights:

- From the above HDP model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- If we see the results all waords are like movies, actors, story etc. this means most the topics were related movies.
- At some point only we found words related to politics and elections.

```
LDA Coherence Score:  0.3980443622872856

LSI Coherence Score:  0.4148576437828798

HDP Coherence Score:  0.6139373352693358
```

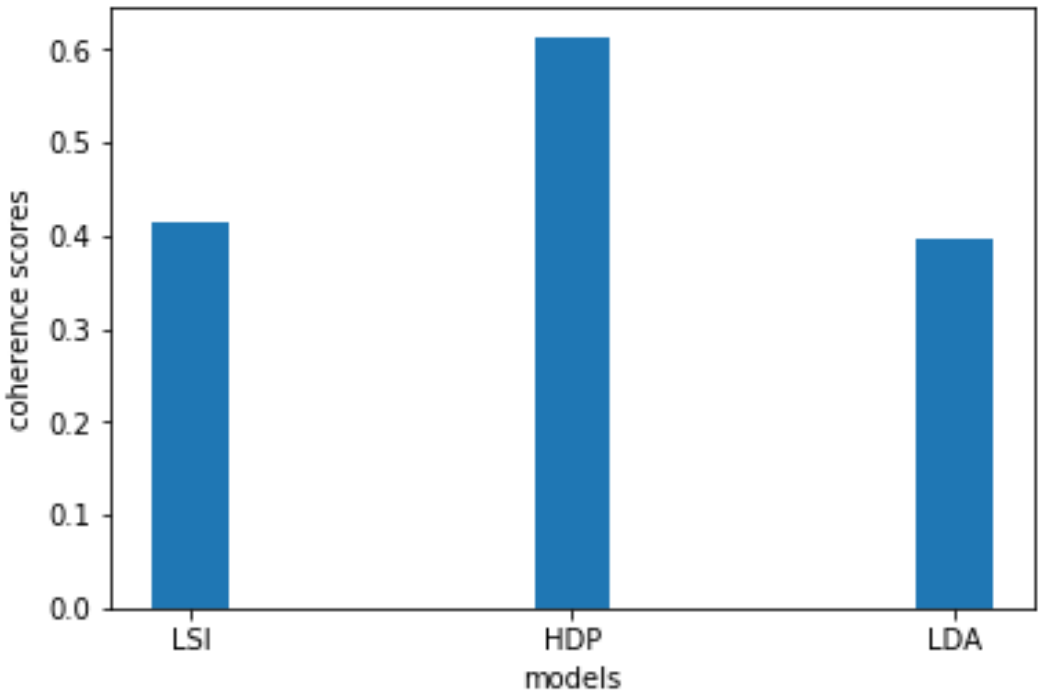


Figure 4.3(b) Accuracy scores (LDA, LSI & HDP Models)

Insights:

- In the above plot we have coharience score of three models. This tells how good the models are.
- HDP model gave high score of coharience to our data compare to LDA&LSI models.
- In this three models Hierarchical Dirichlet process performed better tha the other models.
- We can say that our data is fitted in Hierarchical Dirichlet process model in better way compare to LDA&LSI models.

5. RESULT FOR AUTHOR-II

5.1. Exploratory Data Analytics (EDA)

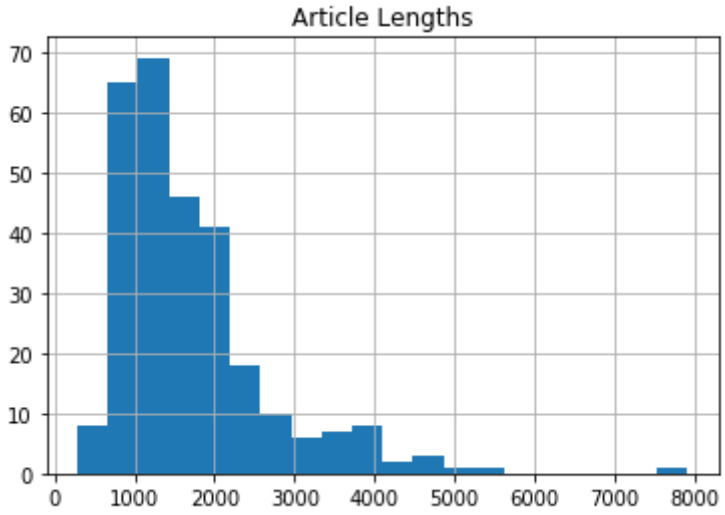


Figure 5.1(a) length of the articles content

Insights:

- From the above graph we can see the length of the articles that written by Dan Riehl.
- The plot says nearly 70 articles length lies between 1000 to 1500. Most of the articles length between 500 to 2000.
- The author wrote very few articles within the length of 500, 3000, 4000, 5000 & 7500.

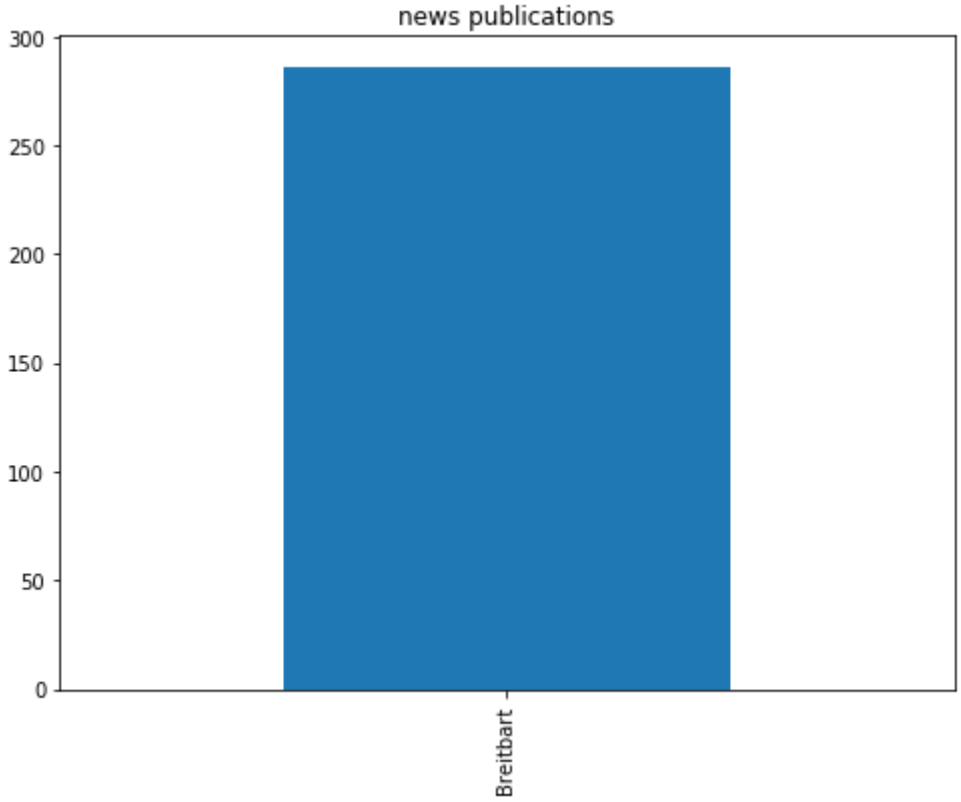


Figure 5.1(b) Bar chart for news publications

Insights:

- The above bar plot shows the count of articles in a particular publications.
- This plot shows that the author published his articles through only one publication that is Breitbart news publications.
- The total articles are around 280.

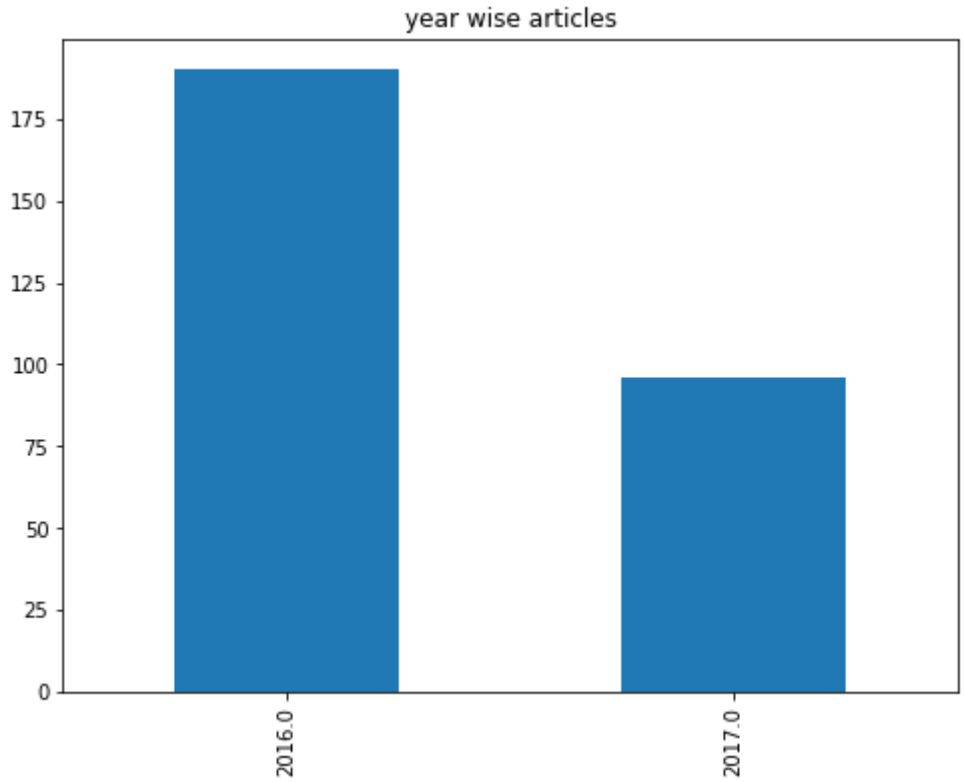


Figure 5.1(c) Bar chart for year wise articles

Insights:

- From the above bar plot we can tell in which year the author published the articles.
- Compare to 2017 most of the articles were published in 2016.
- in 2017 more than 180 articles were published, whereas in 2016 around 100 articles were published.

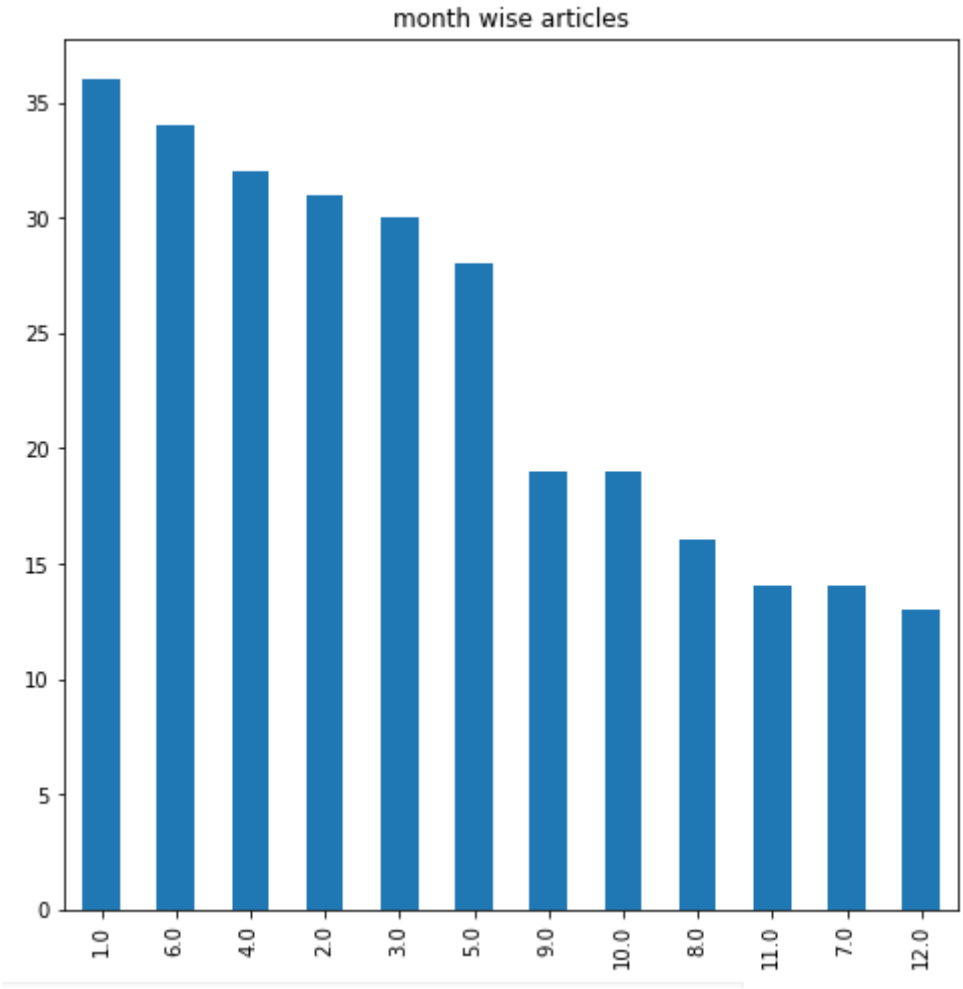


Figure 5.1(d) Bar chart for month wise articles

Insights:

- The above plot show number of articles published in every month.
- From the plot we can say most of the articles published in 1-6 months. In remaining months only few articles were published.

5.2. SENTIMENT ANALYSIS

5.2.1. Sentiment polarity

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
neutral
```

Insights:

- The above results is about sentiment polarity which tells about the articles sentiment like positive, negative or neutral.
- This author articles are perfectly in neutral, because an article suppose to expose the facts in news

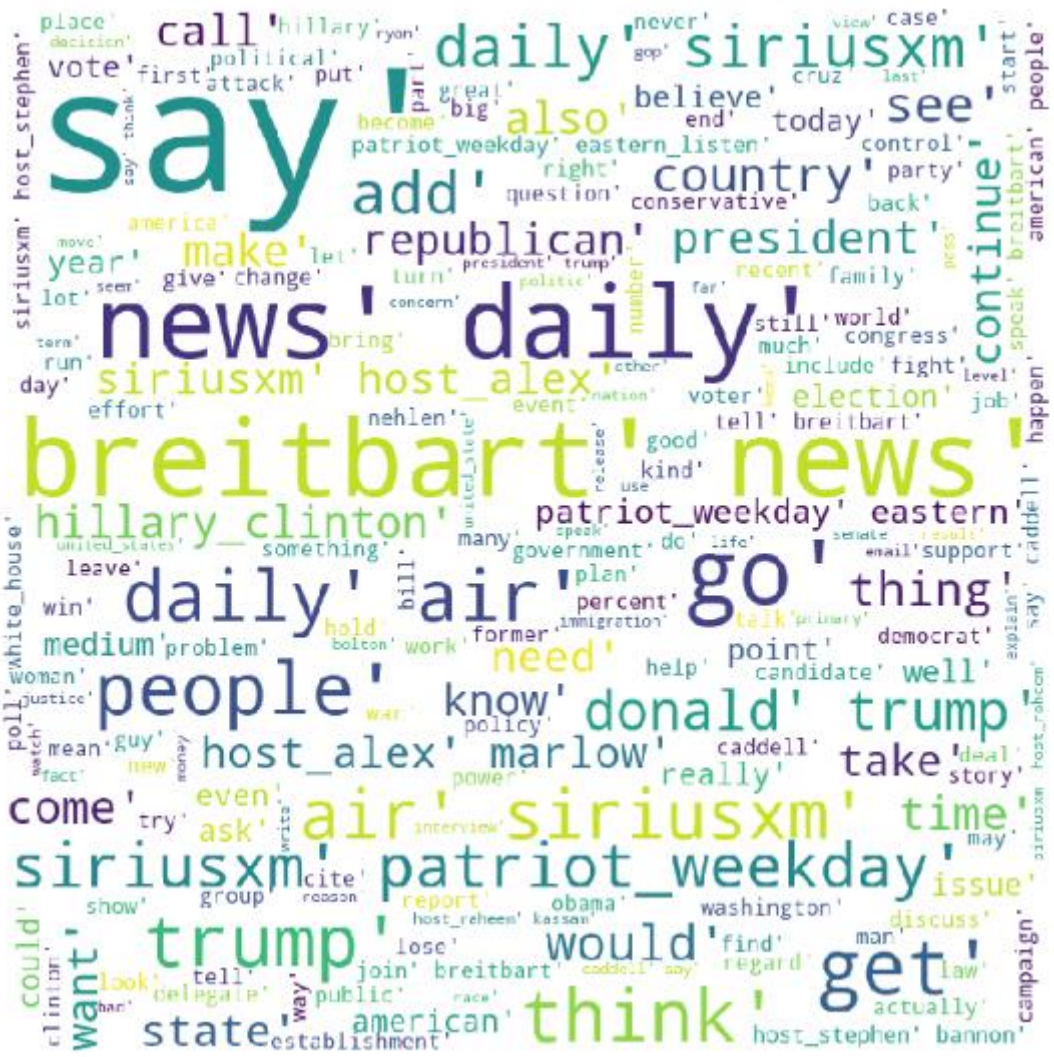


Figure 5.2(a) Word Cloud

Insights:

- The above plot is word cloud, it visualizes the most repeated words in the articles with big size.
- We can see here ‘say’, ‘trump’, ‘breitbart’, ‘news’ are very big compared to other words, it means author have used these words frequently in articles.
- In the same way, if words are looking very small means that words are used very rare in the articles.

5.3. MODEL PERFORMANECE

5.3.1. LDA model

```
[(0,
  '0.018*"say" + 0.013*"trump" + 0.012*"breitbart" + 0.011*"news" + 0.010*"go" + '
  '0.009*"daily" + 0.008*"nehlen" + 0.008*"republican" + 0.007*"siriusxm" + '
  '0.007*"people"'),
 (1,
  '0.016*"breitbart" + 0.014*"ryan" + 0.014*"news" + 0.013*"daily" + '
  '0.012*"siriusxm" + 0.009*"terrorist" + 0.009*"say" + 0.008*"go" + '
  '0.007*"border" + 0.007*"immigration"'),
 (2,
  '0.005*"tancredo" + 0.005*"ice" + 0.005*"kill" + 0.003*"edward" + '
  '0.003*"mayhem" + 0.003*"liable" + 0.003*"cream" + 0.003*"shop" + '
  '0.003*"marten" + 0.003*"hernandez"'),
 (3,
  '0.026*"say" + 0.018*"trump" + 0.015*"go" + 0.014*"news" + 0.013*"breitbart" + '
  '0.011*"daily" + 0.010*"people" + 0.010*"think" + 0.009*"donald" + '
  '0.009*"siriusxm"'),
 (4,
  '0.019*"news" + 0.018*"breitbart" + 0.017*"say" + 0.013*"daily" + '
  '0.012*"siriusxm" + 0.007*"people" + 0.007*"get" + 0.006*"go" + 0.006*"air" + '
  '0.006*"day"'),
```

```
Perplexity:  -7.63376330038478
```

Insights:

- From the above LDA model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- In results we can see most the words are related to politics, terrorist,kill. so we can say that this Dan Riehl published articles are related to trump, society, & crime.
- Perplexity is a measure of how good the model is. Lower score indicates the good model.

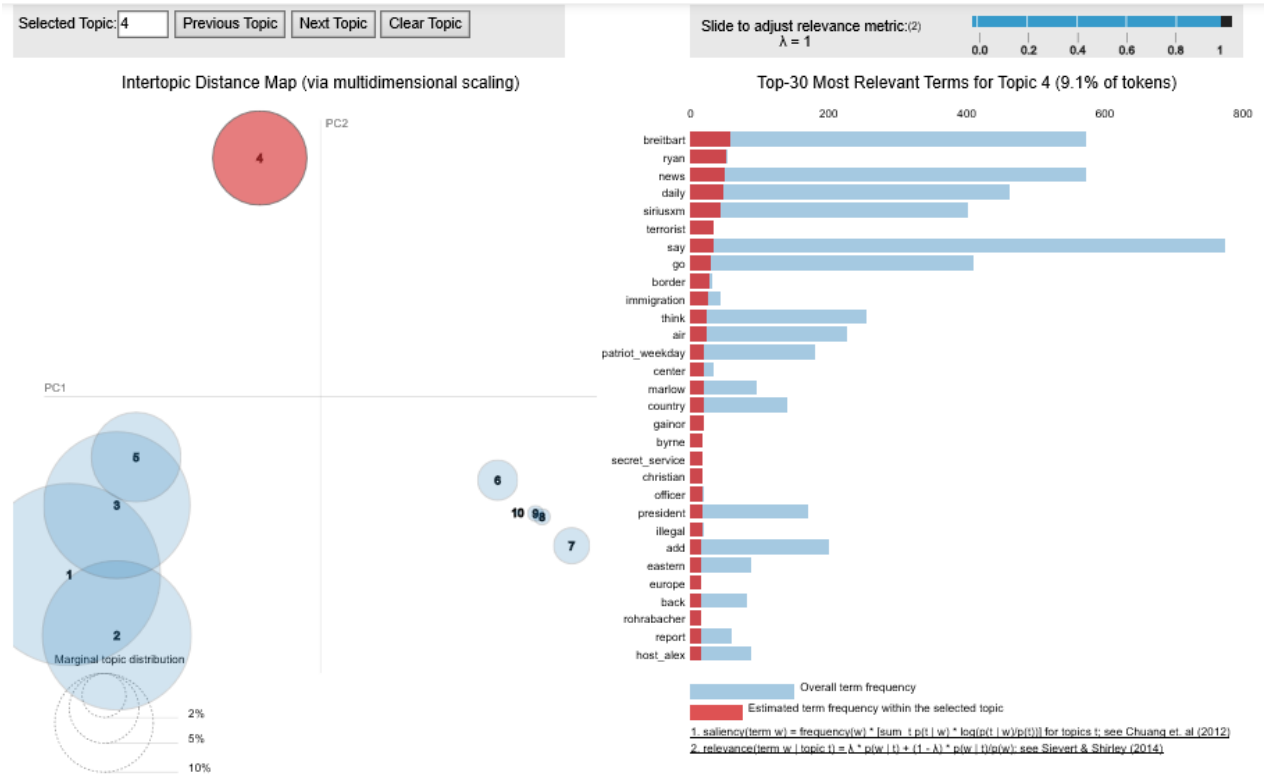


Figure 5.3(a) Visualization of LDA model

Insights:

- The area of circle represents the importance of each topic over the entire corpus, the distance between the centre of circles indicate the similarity between topics.
- For each topic, the histogram on the right side listed the top 30 most relevant terms.
- I found out that majority of the words do related to one filed. Most of the words like ‘kill’, ‘trump’, ‘terrorist’, ‘people’ etc.

5.3.3. LSI model

```
[(0,
  '0.434*"say" + 0.289*"trump" + 0.282*"news" + 0.268*"breitbart" + 0.216*"go" '
  '+ 0.215*"daily" + 0.189*"siriusxm" + 0.164*"people" + 0.154*"think" + '
  '0.147*"get"'),
 (1,
  '-0.649*"trump" + 0.271*"breitbart" + -0.242*"haye" + -0.216*"donald" + '
  '0.186*"news" + 0.186*"daily" + 0.184*"siriusxm" + -0.099*"weekly_standard" '
  '+ -0.096*"republican" + -0.092*"cruz"'),
 (2,
  '0.285*"say" + 0.247*"vatan" + 0.243*"nehlen" + -0.235*"breitbart" + '
  '-0.232*"news" + -0.217*"trump" + -0.185*"daily" + 0.178*"get" + '
  '-0.176*"siriusxm" + 0.162*"paul_ryan"'),
 (3,
  '0.385*"nehlen" + -0.288*"vatan" + 0.257*"paul_ryan" + 0.245*"ryan" + '
  '-0.202*"think" + -0.177*"get" + -0.161*"review" + 0.156*"district" + '
  '0.148*"percent" + -0.144*"qtc"'),
 (4,
  '-0.321*"haye" + 0.241*"caddell" + 0.225*"say" + 0.171*"donald" + '
  '0.170*"cruz" + -0.168*"vatan" + -0.155*"news" + 0.147*"delegate" + '
  '0.144*"get" + -0.136*"file"'),
```

Insights:

- From the above LSI model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- In the results we can see most the words are related to tump & about public, so we can say that ‘Dan Riehl’ published articles are related to trump.
- In this model it shows most of articles about trump compare to other articels are very few.

5.3.4. HDP model

[(0, '0.011*say + 0.010*trump + 0.009*news + 0.009*breitbart + 0.009*go + 0.008*daily + 0.007*siriusxm + 0.006*get + 0.005*issue + 0.005*see + 0.004*air + 0.004*thing + 0.004*cop + 0.004*patriot_weekday + 0.004*people + 0.004*know + 0.004*president + 0.003*caddell + 0.003*add + 0.003*donald'),

(1, '0.012*say + 0.005*news + 0.005*breitbart + 0.005*daily + 0.005*get + 0.004*caddell + 0.004*hillary_clinton + 0.004*also + 0.004*percent + 0.004*republican + 0.004*would + 0.004*people + 0.003*siriusxm + 0.003*huma_abedin + 0.003*try + 0.003*trump + 0.003*think + 0.003*believe + 0.002*host + 0.002*cite'),

(2, '0.008*trump + 0.006*donald + 0.005*say + 0.004*news + 0.004*breitbart + 0.004*bossie + 0.003*democrat + 0.003*daily + 0.003*siriusxm + 0.003*go + 0.003*hillary_clinton + 0.003*pac + 0.003*get + 0.003*see + 0.002*strength + 0.002*take + 0.002*ricker + 0.002*discuss + 0.002*think + 0.002*campaign'),

(3, '0.006*say + 0.005*dixon + 0.005*comic_strip + 0.004*caddell + 0.004*medium + 0.004*siriusxm + 0.003*news + 0.003*daily + 0.003*graphic_novel + 0.003*breitbart + 0.003*trump + 0.003*people + 0.002*tell + 0.002*make + 0.002*american + 0.002*clinton_cash + 0.002*think + 0.002*new + 0.002*money + 0.002*friday'),

(4, '0.007*say + 0.006*koch + 0.004*think + 0.004*trump + 0.003*call + 0.003*news + 0.003*kind + 0.003*muslim + 0.003*cruz + 0.003*people + 0.003*would + 0.003*gaffney + 0.002*breitbart + 0.002*least + 0.002*real + 0.002*jihad + 0.002*could + 0.002*nazi + 0.002*comment + 0.002*program'),

Insights:

- From the above HDP model result we can see most of topics are clustered with their word frequency.We can easily find that what kind articles are written by this author in the period of 2016&2017.
- If we see the results all waords are like people, trump, think etc. this means most the topics were related trump and their nation.
- In this model we can see different topics not only about trump.we can say that this model clustered the topics in better way.

```
LDA Coherence Score:  0.49904799384157145

LSI Coherence Score:  0.40818791221360745

HDP Coherence Score:  0.7096354558728941
```

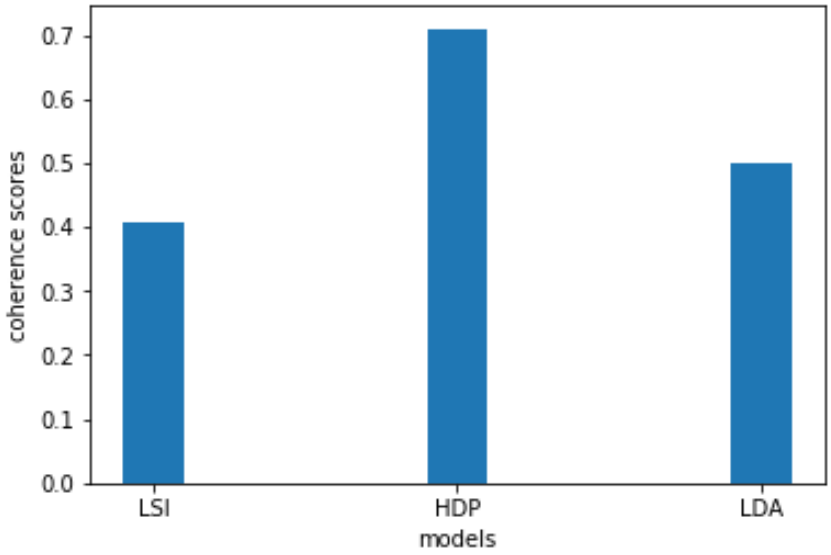


Figure 5.3(b) Accuracy scores (LDA, LSI & HDP Models)

Insights:

- In the above plot we have coharience score of three models. This tells how good the models are.
- HDP model gave high score of coharience to our data compare to LDA&LSI models.
- Somehow LDA performed it's best but LSI gave very poor score to our data.
- We can say that our data is fitted in Hierarchical Dirichlet process model in better way compare to LDA&LSI models.
- HDP model score is 0.70 we can say this is good model compare other.

6. CONCLUSION

This project deals with news articles by American publications with different authors during the period of 2015, 2016& 2017years. This is an unstructured data using text analytics techniques to extract the information from the document. Using text modelling to cluster the news articles written by authors. In topic modelling, the no of topics are equivalent to no of clusters and the cluster points are equivalent to the no of words that are present in topics. using sentiment polarity to find positive, negative, neutral measure for text.

All the articles are shown as neutral opinions in polarity results. From the given results all the authors wrote articles about related to trump. In 2016 elections were took place Donald Trump became the president of the united-states. He is the 45^{th} president of united-states. Before entering politics, he was a businessman and television personality. In this election period of time trump gave many statements the he want to do for united-states. So this may the reason why we see most of the articles about trump. There are many other articles related to movies, starts, crime, education compare to other fields most of the articles are related to trump. In the above three models Hierarchical Dirichlet process model gave us best result compare to other models.

6.1. SCOPE FOR FUTURE ENHANCEMENT

IOT: The Internet of Things (IoT) provide intelligence for the communication between people and physical objects. An important and critical issue in the IoT service applications is how to match the suitable IoT services with service requests. To solve this problem, researchers use semantic modeling methods to make service matching. Semantic modeling methods in IoT extract meta-data from text using rule-based approaches or machine learning techniques often suffer from the scalability and sparseness since text provided by sensors is short and unstructured. In recent years, topic modeling has been used in IoT service matchmaking.

However, most topic modeling methods do not perform well in IoT service matchmaking since the text is too short. In order to address the issues, this paper proposes a new topic modeling method to extract topic signatures provided by intelligent devices. The method extends the classical knowledge representation framework and improves the qualities of service information extraction, and this process is able to improve the effectiveness of service matchmaking in IoT service. The framework incorporates human cognition to improve the effectiveness of the algorithm and make the algorithm more robust in heterogeneous systems in the IoT. The usefulness of the method is illustrated via experiments using real datasets.

REFERENCE

- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
- https://radimrehurek.com/gensim/models/ldamodel.html
- https://radimrehurek.com/gensim/models/lsimodel.html
- https://radimrehurek.com/gensim/models/hdpmodel.html
- https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21
- https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908
- https://monkeylearn.com/blog/sentiment-analysis-with-python/