JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org **Vol.2, Iss.2 Pages:32-41**

TOPIC MODELLING USING LATENT DIRICHLET ALLOCATION

[1]V Ranjitha, [2]V Srividhya

1,2Dept of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India - 641043

Article Received: April 2021 Published: July 2021

Abstract

Topic modeling is an unsupervised machine learning system that is accomplished of scanning extra-large documents, phrase patterns and discovering words with them, and manually related words and clustering words that greatest characterize a group of documents. It is started from extracting the topics and text-mining techniques for discovering the latent semantic structure in a set of news documents. This conception of extracting the topics in each document is produced from a collection of topics. In this paper, propose a technique to categorize the news to topics using LDA (Latent Dirichlet Allocation) Model for Extraction the topics. The objective is to identify key themes or topics from the Twenty Newsgroups dataset. The dataset contains 20,000 news articles and extracted text. LDA is the probabilistic method that works to study the topic illustration in each text or file and the word collection of each topic. Topic models to analyze newsgroups dataset with LDA model to visualize the topics.

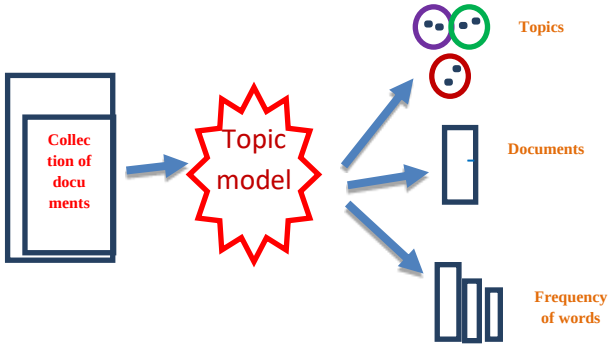
Keywords: Topic Modeling, LDA, visualization, Pre-processing, 20 newsgroups

I INTRODUCTION

Topic models are a statistical outline that supports handlers to recognize huge groups of documents just to find individual documents but to know the common subjects current in the collection. With growth of the internet and the development of blogs, reviews, and social networks, techniques like Topic Modelling, Opinion Mining, and Text Summarization are useful in decision making for Individual Customers, Manufacturers, and Organizations. Topic Models, in a nutshell, are a kind of statistical language model used for discovery hidden structure in a group of texts.A topic is a set of arguments that occur frequently together, and topic modelling may connect terms with similar meanings and distinguish between word expenses across a number of implications. It is developed by finding latent data and searching for associations between text data. In the field of text mining, topic modelling is one of the newest techniques. It's used to locate topics inside a set of works.

Natural language processing (NLP) is a fascinating research paper in computer science and information technology that involves allowing computers to understand human language processing in text documents. subject modelling approaches are highly effective and elegant strategies for subject exploration from unordered or unlabelled documents that are often used in natural language processing [1]. This model can be used in a variety of fields, including computer engineering, politics, medicine, and geography. Object modelling refers to the algorithms used to determine the main topics. It might be able to compile a list of the discovered terms. Large collections of documents may benefit from topic modelling algorithms. The following Fig.1 illustrates the steps to perform topic modelling.

Fig 1 Topic modeling s



Topic modelling technique maps the text collection to a low dimensional topic subspace, which is a cluster of words, known as the topic. Topic models are statistical models to uncover the hidden structure through large text collection, digital library, and web content and search interface through their ability to automatically infer the topics and categorize the documents accordingly. It can explain the corpus in a variety of ways such as topic proportion in a document, how many documents the specific topics span across, what fraction of the collection falls into various categories of the theme.

Present-day advance in this method allows to analyse stream collections. Topic modelling algorithms can alter to several kinds of documents, along with their applications, they have been used to find patterns in social networks, genetic data and images [2].

The topic model's basic premise is that a text can be represented as a set of topics, each of which is a probability distribution over the terms in the document. LDA (Latent Dirichlet Allocation) is a very common subject modelling technique.LDA is a correctly generative model that is achieved in generalising topic distributions so that it can be used to render invisible documents as well. It is a combination of topic models expected. Consider Dirichlet priors on text distributions over topics and division over terms to improve the comprehensiveness of the generative process for documents.

The use of the approach poses a number of analysis problems. Classification of a multinomial subject model Management of a large number of documents forming a large number of topics Scalability, the demonstration of a subject model and a mark are incompatible to define the importance of each subject with accuracy. To understand the arrangement of topics that aren't obvious. Discretization is defined by the constant dynamic subject model (CDTM).

II RELATED WORK

A topic model is very helpful in identifying multiple latent conceptions present in a text category and helping the handler to understand the topics addressed in each document separately [1]. Probabilistic generative methods, such as the subject model, are commonly used in the fields of information retrieval and text mining [3]. Topic models are used to find a set of terms that describe a single topic. They have a subject and a production that includes the likelihood of how powerful each term is linked to that topic. These models are used to explore a document's latent semantic constructs.

III METHODOLOGY

There are five steps to this initiative. The first step is data collection. Pre-processing takes just a second of the time. Exploration is the third phase's work. The model's development is the fourth step of work (LDA). The fifth step is visualisation. The following fig.2 shows the overall methodology diagram.

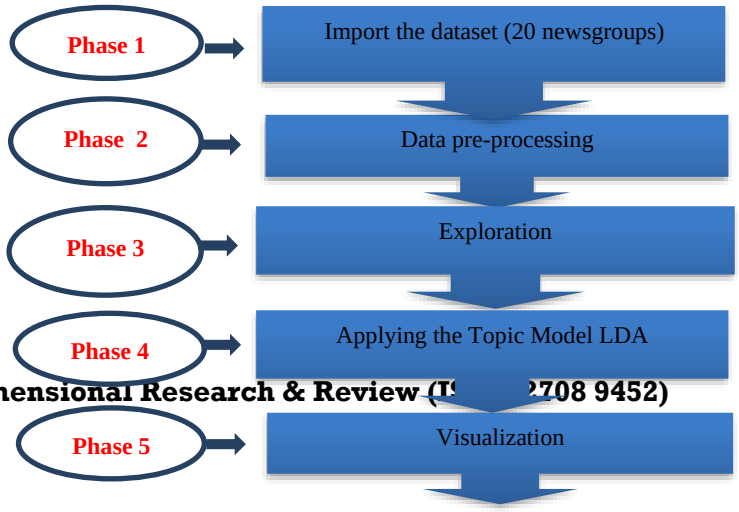


Fig 2 Methodology diagram

(i) Data set

To implement the work, 20 newsgroup data set has taken from UCI repository. This dataset consist of 20000 news article ranging from 1999 up to 2016 and number of instances is 20000.

(ii) Data pre-processing

The proposed extracting topics starts with pre-processing the data obtained from 20newsgroup.The newsgroup dataset are a document that is represented as a series of words and punctuation. Data pre-processing is a data mining technique for transforming raw data into a format that is both useful and effective.[2] translate to lowercase, [3] delete special characters and tokenize them into phrases, [3] remove stop words, [4] stemming, and [5] build term-document matrix are the five tasks in our data pre-processing. The details of each of these tasks are explained below.

In step 1 Tokenizes the document into words, excluding special characters such as punctuation marks (e.g.! percent $#&*?,/.;") percentages, as they never add to our text mining research.

In data pre-processing, step 2 is the most important task. Stop words are a collection of widely used words in any language that can be omitted from the text in order to retain the crude content, such as whether a customer submitted ranking scores on an online survey or created labels on an image. In certain instances, such a side flag contains supplementary information to help reveal the underlying mechanisms of the data under review.

Stemming's third aim is to eliminate text data variance by translating words to their standard base form/word stem. For instance, auto, auto mobile, and auto mobiles were all translated to auto, and universe, university, and universalization were all transformed to universe.

(iii) Exploration

Model development and data processing are also characterised by data discovery. You can't create professional analytical models without spending a lot of time sorting data and designing it. In the data science process, which involves pre-processing and data interpretation, data discovery takes a considerable amount of time.

(iv) Latent Dirichlet Allocation

The Latent Dirichlet classification (LDA) is a probabilistic general model that has been widely used to find dormant points (topics) that swarm a corpus. Each found theme commonly comprises of a set of related individual words, which are the real subjects examined or sentiments communicated in the corpus. In this way, theme models can be utilized in numerous imperative research regions [6]. Topic demonstrating offers a suite of valuable apparatuses that naturally take in the idle semantic structure of a gathering of reports or pictures, with inactive Dirichlet designation (LDA) [7] as a generally prevalent precedent.

Its success stems from the fact that it structures each text as a multi-membership mixture of K corpus-wide topics, with each subject being a multi-membership mixture of the words in the corpus vocabulary.

In LDA, each document is considered as combination of words. LDA systematically estimates both of these at the same time. It finds the combination of language that is connected among all topic, although influential the mixture of topics that describes every document. LDA model requires text data in the form of pure words (No punctuation, wide spaces, etc.), But, the input of dataset is not in the pure form. It contains various spaces, punctuation, different cases of letters and many things. LDA generates topics from documents and each topic consists of multiple words. It is method which automatically, detects topics from the documents.

The process is performing topic detection on large datasets, determining what the main ‘topics’ are in large unlabeled set of texts. This will helps to find the topics from the news articles. The following fig. 3 shows the result of LDA.

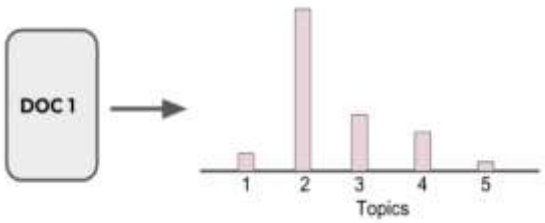


Fig. 3 Latent Dirichlet diagram

(v) Visualization

Finally, pyLDAvis is the most commonly used and a nice way to visualize the information contained in a topic model. It is considered to facilitate users understand the themes

in a topic model that has been well to a corpus of text data. The extraction sequence from a fixed LDA topic model is to tell an interactive web-based visualization.

IV RESULTS AND DISCUSSION

LDA is to extract the main topics from the 20news group dataset and results are given below in Table 1a and 1b.

Topic 0

(0, '0.051*"report" + 0.027*"black" + 0.020*"fire" + 0.020*"white" + ' '0.016*"trial" + 0.016*"cover" + 0.015*"medium" + 0.013*"vote" + ' '0.012*"minor" + 0.012*"title"'),

Topic 1

(1, '0.021*"god" + 0.020*"accept" + 0.016*"member" + 0.015*"man" + ' '0.014*"israeli" + 0.014*"season" + 0.012*"publish" + 0.012*"lebanese" +' '0.012*"jewish" + 0.011*"brain"')

Topic 2

(2, '0.017*"package" + 0.016*"press" + 0.015*"item" + 0.015*"break" + ' '0.011*"level" + 0.010*"edge" + 0.009*"hole" + 0.007*"eye"+' '0.007*"contribute"+ 0.007*"equipment"')

Topic 3

(3, '0.025*"pc" + 0.022*"contain" + 0.020*"input" + 0.020*"reality" + ' '0.017*"picture" + 0.016*"object" + 0.016*"level" + 0.015*"box" + ' '0.015*"quality" + 0.013*"greek"')

Topic 4

(4, '0.089*"" + 0.076*"max" + 0.032*"space" + 0.021*"launch" + 0.018*"di_di" + ' '0.017*"orbit" + 0.016*"sphere" + 0.015*"satellite" + 0.014*"plane" + ' '0.014*"mission"')

Topic 5

(5, '0.019*"people" + 0.017*"kill" + 0.015*"child" + 0.015*"government" + ' '0.012*"attack" + 0.012*"year" + 0.012*"die" + 0.011*"country" + 0.010*"say" ' '+ 0.009*"war"')

Topic 6

(6, '0.035*"window" + 0.032*"card" + 0.020*"image" + 0.020*"driver" + ' '0.020*"problem" + 0.019*"run" + 0.018*"sale" + 0.018*"machine" + ' '0.017*"color" + 0.016*"scr/een"')

Topic 7

(7, '0.025*"people" + 0.021*"say" + 0.014*"reason" + 0.014*"believe" + ' '0.012*"may" + 0.012*"evidence" + 0.010*"make" + 0.010*"think" + ' '0.009*"many" + 0.009*"mean"')

Topic 8

(8, '0.032*"book" + 0.023*"physical" + 0.021*"science" + 0.017*"choose" + ' '0.016*"explain" + 0.015*"create" + 0.011*"author" + 0.011*"earth" + ' '0.010*"study" + 0.010*"nature"')

Topic 9

(9, '0.033*"mail" + 0.028*"file" + 0.027*"send" + 0.026*"program" + ' '0.025*"thank" + 0.024*"information" + 0.021*"software" + 0.021*"list" +' '0.019*"include" +0.019*"address"')

Table 1a LDA topic list

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
report	god	package	pc	ax
black	accept	press	contain	max
fire	member	item	input	space
white	man	break	reality	launch
trial	israeli	level	picture	orbit
cover	season	edge	object	sphere

mediu m	publish	hole	level	plane
Vote	lebanes e	eye	box	mission
minor	brain	equipm ent	greek	satellite

Table 1b: LDA topic list

Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
People	window	people	book	mail
Kill	card	say	science	file
child	image	may	choose	send
attack	driver	make	create	program
country	run	think	author	thank
die	sale	many	earth	list
say	machine	mean	study	include
war	screen	reason	explain	nature

PyLDAvis

pyLDAvis is a Python package for immersive LDA visualisation. The difference between the centres of circles reflects the similarities between topics for each subject, and the histogram on the right side listed the top 30 most important words. LDA facilitate extraction of ten main topics .The result as shown in Figure 4.

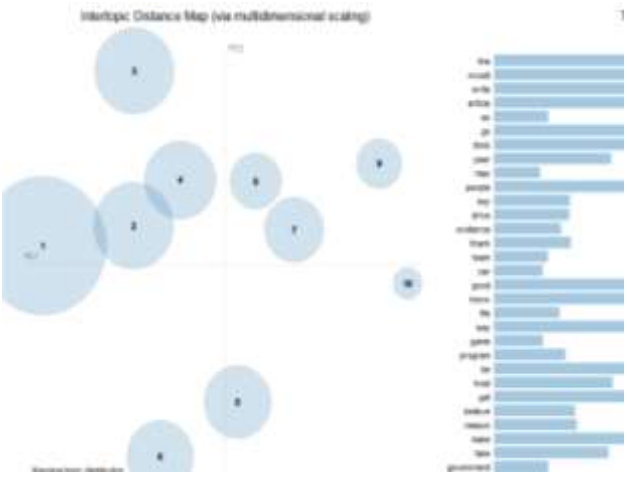


Fig 4 LDA visualization

V CONCLUSION

LDA is used to help accomplish topics or commonly significant phrases. It's conceivable that the number of topics defined as an excited parameter to LDA is manually set. This LDA-based approach was proposed to make valuable knowledge known from 20 newsgroup results. LDA creates semantically significant topic/clusters to review terminology into a grouping of subjects that would be impossible to explain using discrete explanations since analysing large amounts of data and understanding them is time intensive and complicated. The popularity of these issues and their time trends can be used to assess the newsgroup data's utility. The effects of extracting the key topics from a newsgroup dataset using LDA are the subject of this article.

REFERENCES

1. Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation.Journal of machine learning research, 2003. 3(Jan): p. 993-1022.
2. Hamed Jelodar,Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, Liang Zha, Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey
3. David M, ,Blei Surveying a suite of algorithms that offer a solution to managing large document archives, 2012
4. David M.Blei,Andrew Y, Latent Drichlent Allocation, 2003
5. Amir Karami Aryya Gangopadhyay Bin Zhou Hadi Kharrazi, Fuzzy Approach Topic Discovery in Health and Medical Corpora, 2013
6. Zhou Tong,Haiyi Zhang ,A text mining research based on lda topic modelling, 2016
7. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints Kaveh Bastani, Hamed Namavari, Jeffry Shaffer, 2016
8. H.M. Wallach, Topic Modelling: Beyond Bag-of-words, proc. of23rd INT. ICML, PP. 977–984, 2006
9. David M. Blei, Probabilistic topic models, ACM 0001- 0782/12/0, 2012

10. Lin Liu, Lin Tang, Wen Dong, Shaowen Yao and Wei Zhou, An overview of topic modeling and its current applications in bioinformatics, Springer Plus 2016 5:1608
11. David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3, 2003
12. T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl. 1):5228–5235. 2004 https://doi.org/10.1073/pnas.0307752101

Cite this article:

V Ranjitha, V Srividhya, "Topic Modelling Using Latent Dirichlet Allocation", Journal of Multidimensional Research and Review (JMRR), Vol.2, Iss.2, pp.32-41, 2021.