International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-176 PAGE NO: 1075-1078
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

INTERPRETING DOCUMENT COLLECTIONS WITH TOPIC MODEL USING LATENT DIRICHLET ALLOCATION

C.R.Durga Devi,Research Scholar
Department of Computer Science
NGM College
Pollachi, India
deviswe@gmail.com
Dr.R.Manicka Chezian,Associate Professor
Department of Computer Science
NGM College
Pollachi, India
chezian_r@yahoo.co.in

ABSTRACT- Topic models come under the area of text mining and provide a way to analyse large text corpora. a topic contains a cluster of words with similar meaning. topics are created based on the strength of their use in analysis of large repositories of information. topic modeling can be used in various application areas like document clustering, text classification, characterizing core and distributed genes within a species, etc; there are a variety of models available for identifying topics in collection of large number of text documents. these methods are latent semantic analysis (lsa), probabilistic latent semantic analysis (plsa), latent dirichlet allocation (lda).this paper gives an overview of lda model and applies lda to sample corpora taken for study.

Keywords: Topic models; lda; lsa; correlation.

INTRODUCTION

The users search the web by querying method. the words used as query gain importance to find a relevant information [1]. topic modeling is a form of text analysis used to explore relationships between words within a document where the words are grouped together to form topics. there are a variety of different methods for topic modeling, using different sampling algorithms for word selection and topic creation.. latent dirichlet allocation is a basic topic model. it groups words together based on how likely they are to appear in a document together.

documents that come from this word w. reassign w a new topic, where you choose topic t with probability p(topic t | document d) * p(word w | topic t). update the assignment of the current word using this model. after repeating the previous step a large number of times, finally we get correct and good assignment. this iterative process is implemented using a technique called gibbs sampling.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-176 PAGE NO: 1075-1078
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

RELATED WORK

Shi-heng wang et al. applied lda for medical articles dataset collected from pubmed[2]. the articles are related to adolescent substance use and depression. conclusion from this study shows that topic modeling not only segregate large collection of articles but also discovers relevant unknown topics like risk factors, link between substance use and depression.

lijun sun, yafeng yin applied topic modeling (lda) on article abstract data from 1990 to 2015[3].in this study, focused on 22 scientific journals included in science citation index and social science citation index 303 under the category "transportation science & technology" and "transportation". as an unsupervised learning algorithm, lda does not require any prior knowledge and the inferred topics are purely resulted from the statistical structure of the abstract-word data. using the posterior document-topic and topic-word distribution, investigated the context of each topic and the evolution of topics across time and journals

massimo la rosa et al. presented a novel computational approach for gene sequence classification[4]. using the probabilistic topic models mainly adopted in text mining applications and developed a pipeline that, by means of the latent dirichlet allocation algorithm, is able to learn a probabilistic topic model from a dataset of 16s gene sequences.

wongkot sriurai applied latent dirichlet allocation algorithm to cluster the term features into a set of latent topics [5]. same topic will has terms that are semantically related. comparision between the feature processing techniques of bow model and the topic model is executed.

abdessalem kelaiaia and hayet merouani applied latent dirichlet allocation algorithm and k-means for clustering arabic texts [6]. the results consistently show a clear improvement of lda over k-means on raw, cleaned, and stemmed forms of document collections.

METHODOLOGY

The dataset contains collection of 30 posts from blogs of eight2late.wordpress.com as a corpus. topic modelling in r uses the packages tm, topic models, and ggplot. pre-processing of the corpus involves the steps of transforming to lower case, removing problematic symbols, punctuation, numbers, strip whitespace, stop words and stemming. stemming is the process of reducing such related words to their common root. we can also set our own collection of stop words to be removed from the collection of documents. a document term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. this dtm will be the main input for lda algorithm. in this document term matrix, convert rownames to filenames; collapse the matrix by summing over columns. arrange all terms in decreasing order of frequency. the total number of occurrences of each word across all documents is given in fig.4

fig 4.cummulative frequency of words across documents

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-176 PAGE NO: 1075-1078
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

rstudio

	WORD	FREQUENCY
1	manag	229
2	work	209
3	system	193
4	project	187
5	problem	171
6	exampl	164
7	differ	158

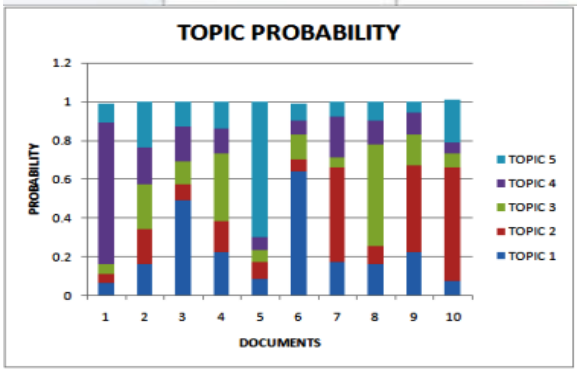

fig.8

Topic probability for first ten documents

CONCLUSION

Topic modeling provides a quick and convenient way to perform unsupervised classification of a corpus of documents.in this study, lda model is applied to sample corpora with 30 documents and got the result of topic probabilities associated with each documents.

REFERENCES

1. R.deepa and dr.r.manicka chezian,"an ontological approach for the semantic web search and the keyword similarity metrics", international journal of advanced research in computer and communication engineering, issn (online) 2278-1021,vol. 5, issue 3, march 2016

2. shi-hengwang et al.,"text mining for identifying topics in the literatures about adolescent substance use and depression", article in bmc public health, doi: 10.1186/s12889-016-2932-1, december 2016

3. lijun sun, yafeng yin "discovering themes and trends in transportation research using topic modeling", preprint submitted to transportation research, june 12, 2016

BIOGRAPHY

C.R.Durga devi received her bsc.computer technology from coimbatore institute of technology,coimbatore,india. she had her master of computer applications from bharathiar university,coimbatore, india. she holds mphil, in computer science from bharathiar university,

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-176 PAGE NO: 1075-1078
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

coimbatore, india. she has 9 years of experience in teaching. she is presently working as an assistant professor in ngm college, pollachi. her research interest includes data mining,big data analytics. now she is pursuing her ph.d computer science in dr.mahalingam center for research and development at ngm college pollachi.

Dr.R.Manickachezian received his m.sc., degree in applied science from p.s.g college of technology, coimbatore, india in 1987. he completed his m.s. degree in software systems from birla institute of technology and science, pilani, rajasthan, india and ph d degree in computer science from school of computer science and engineering, bharathiar university, coimbatore, india. he served as a faculty of maths and computer applications at p.s.g college of technology, coimbatore from 1987 to 1989. presently, he has been working as an associate professor of computer science in n g m college (autonomous), pollachi under bharathiar university, coimbatore, india since 1989. he has published 150 papers in international/national journal and conferences: he is a recipient of many awards like best computer science faculty of the year 2015, best research supervisor award, life time achievement award, desha mithra award and best paper award. his research focuses on network databases, data mining, distributed computing, data compression, mobile computing, real time systems and bio-informatics.