Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 468 – 474

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

Prediction of Online Lectures Popularity: A Text Mining Approach

Kavita S. Oza[a], Poornima G. Naik[b] *

[a]*Shivaji University, Kolhapur, Maharshtra-416004 ,India*

[b]*CSIBER, Kolhapur, Maharshtra-416004 ,India*

Abstract

Text mining is an emerging area of research with flavors of opinion mining, sentiment mining, document classification, content mining etc. Another flavor of text mining is text clustering. Proposed work is based on clustering the comments posted by users to online learning. The dataset is prepared using comments posted by users for text mining video lectures using R and Weka. In the proposed work learners comments for online text mining lectures have been clustered to observe the popularity of the lectures by analyzing the terms in each cluster.

Keywords: Text Mining, Clustering, R studio, Video lectures, online learning

1) INTRODUCTION

* Corresponding author. Tel.: +9850766660 ; fax: +0-000-000-0000 .

*E-mail address:*kso_csd@unishivaji.ac.in

Peer-review under responsibility of the Organizing Committee of ICCC 2016

doi:10.1016/j.procs.2016.07.369

Text mining is a specialized branch of Data mining. Data mining deals with mining hidden knowledge but in text mining information is plainly present in the text there is no concept of hidden information. Text mining main objective is to get text in computer understandable form directly so that it can be processed without human intervention. Data mining works with structured data like databases, data warehouse, online shopping data, mobile usage data etc. Text mining works with unstructured or semi-structured natural language data. Example of dataset for text mining is data generated by social media, which is natural language unstructured data. So biggest hurdle to text mining is natural language processing.

There are technologies coming up which will help computers to understand natural language to analyze and interpret. Some of the techniques used in text mining is information retrieval, summarization, clustering and classification. Information retrieval may be to mine some interesting patterns within the text. Human experts do exist who can summarize the text with fewer sentences and core concept. Attempts have been made to develop techniques which can help

computers in summarization of text. It may deal with summarization of one document or group of documents. Summarization is condensation of text into smaller version. Here information and meaning of text is maintained.

Clustering is another popular unsupervised technique of Data mining. Text data clustering has applications in customer categorization, document classification, pattern evaluation etc. In the proposed work learners comments for online text mining lectures have been clustered to observe the popularity of the lectures by analyzing the terms in each cluster. Paper is organized into four sections with section 1 dedicated to introduction of the text mining, section-2 deals with literature survey , section- 3 talks about data set creation followed by data analysis and paper is concluded with conclusion.

2) LITERATURE REVIEW

New technologies are emerging to make text mining more comfortable like Data mining. Its not only text data which mined but also text stream mining is a new research area where new techniques are proposed for text stream classification and evolution analysis of the same[1],[2] . Clustering is widely studied data mining problem in the text domains with applications in number of domains. A detailed survey of the problem of text clustering has been carried out with text domain in focus [3]. A comparative study of document clustering using various techniques on twitter data has been carried out by[4]. Improvement in the clustering co-citation models by using full text along with bibliographic information has been proposed by[5]. Real world applications of text mining and its complexity in implementation has been studied by [6]. Lots of work has been carried out in clustering and analysis of complete biomedical article texts. An algorithm is introduced by [7] for Semi-supervised Affinity Propagation (SSAP) to improve

analysis efficiency, using biomedical journal names as an evaluation background. Application of textual clustering for Defect Resolution Time (DRT) with focus on accuracy has been proposed by [8]. An algorithm with combination of classical partitioning algorithms with probabilistic models is designed by [9] in order to create an effective clustering approach. Text Summarization is another promising area of research where text is condensed into shorter version without any change in its meaning and information content. Survey of Text Summarization Extractive techniques has been carried out by [10].

3) DATASET CREATION

Dataset is prepared by taking the comments posted for the video lectures viz.

Text Mining in R Tutorial: Term Frequency & Word Clouds, deltaDNA; Text Mining for Beginners, Linguamatics; Text Analytics and Text Mining Explained by OdinText; Text Mining - Part I, FlávioClésio

Weka Text Classification for First Time & Beginner Users, Brandon Weinberg etc.

All the comments are stored in two text files comments and comments1. This datasets is loaded into R-studio[11] for processing. Dataset is preprocessed to remove numbers, conversion of text to lowercase, elimination of common words like 'the', 'is',' a' etc, and also punctuations from the dataset. Now the preprocessed dataset is ready for analysis.

4) DATA CLUSTERING

The preprocessed dataset is used to create document term matrix. Following is output of Document Term matrix:

DocumentTermMatrix (documents: 2, terms: 402)
Non-/sparse entries: 449/355
Sparsity : 44%
Maximal term length: 24
Weighting : term frequency (tf)

The two documents are comments and comments1, with number of terms 402 and maximum term length as 24. Sparse terms are further removed from the dataset and final dataset has :

DocumentTermMatrix (documents: 2, terms: 47)
Non-/sparse entri es: 94/0
Sparsity : 0%
Maximal term length: 10
Weighting : term frequency (tf)

This final dataset is clustered into three clusters using kmeans algorithm with Euclidean distance. Following figure 1 and figure 2 shows the plotting of clusters using clusplot as well as R interface.

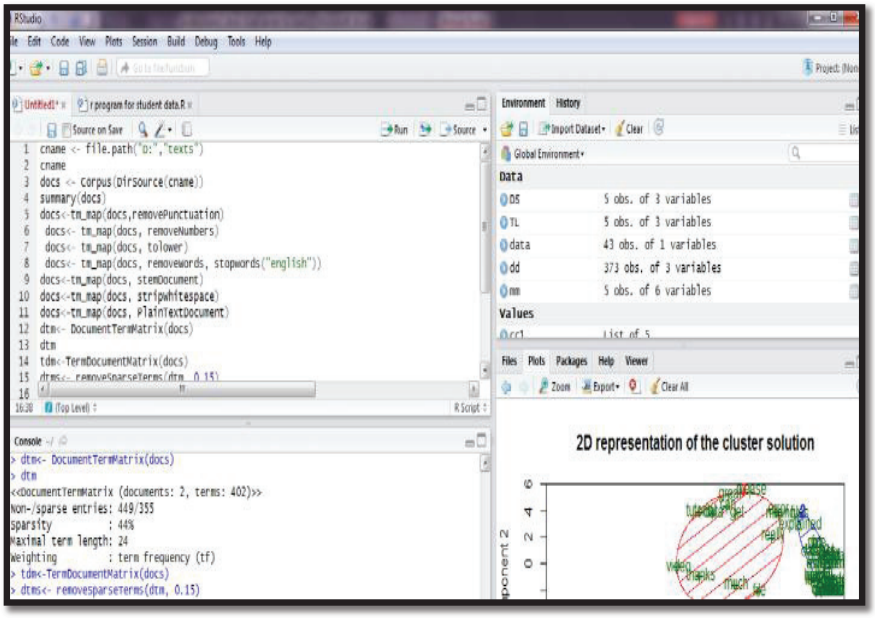

Fig.1 Text document clusters with R studio

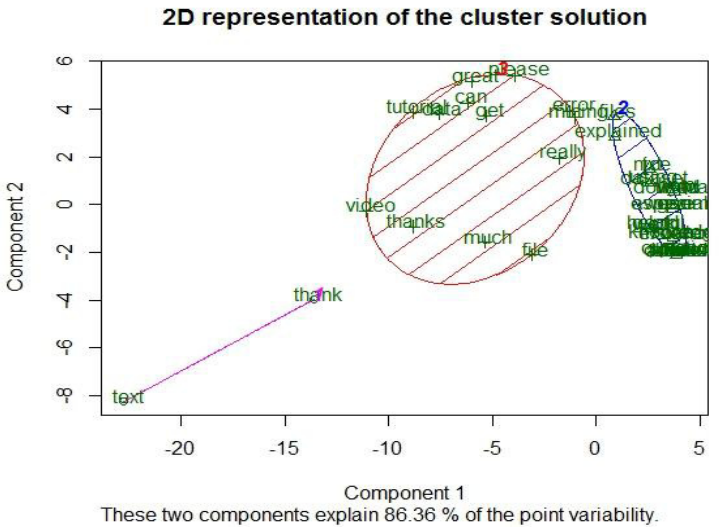

Fig.2 clusters in detail

Here component 1 and component2 are first two principal components which were derived from the dataset. 86.36 of variability indicates that almost 87% of the information about multivariate data is captured by this plot of components. Which is quite a good percentage indicating good Clusters. Cluster details are as follows:

Cluster1 → 2
Cluster2 → 32
Cluster3 → 13

Sample of cluster data is shown below with the term and the cluster number.

download	please	especially	excellent
2	3	2	2
great	help	helpful	instance
3	2	2	2
knowledge	line	lot	mining
2	2	2	3
much	nice	one	please
3	2	2	3
question	really	saved	text
2	3	2	1

thank	thanks	tutorial	txt
1	3	3	2
upload	use	used	using

It can be observed that cluster1 has only two items i.e. thank and text. Here thank is something which is not commonly used and word text may not be relevant to video lectures so we can conclude that cluster1 contains outliers.

Cluster2 contains the terms like nice, want, helpful,upload, knowledge etc. are positive terms which shows that users are happy with video lecture. Majority of comments i.e. out of 47terms 32 terms belong to this cluster.

Cluster3 has 13 terms viz. thanks, tutorial, video, great, much etc.Here user comments are better than the Cluster2.

5) CONCLUSION

Cluster analysis shows that learners are happy with online learning rather than traditional way of class room teaching. This analysis can be further extended to different topics also & can be one of the parameter to be considered for launching a new online course and also about the popularity of the resource person.

References:

[1] Aggarwal, Charu C., and Chandan K. Reddy, eds. *Data clustering: algorithms and applications*. CRC Press, 2013.

[2] Aggarwal, Charu C., and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[3] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." In *Mining Text Data*, pp. 77-128. Springer US, 2012.

[4] Rangrej, Aniket, Sayali Kulkarni, and Ashish V. Tendulkar. "Comparative study of clustering techniques for short text documents." In *Proceedings of the 20th international conference companion on World wide web*, pp. 111-112. ACM, 2011.

[5] Boyack, Kevin W., Henry Small, and Richard Klavans. "Improving the accuracy of co-citation clustering using full text." *Journal of the American Society for Information Science and Technology* 64, no. 9 (2013): 1759-1767.

[6] Miner, Gary. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.

[7] Guan, Renchu, Chen Yang, Maurizio Marchese, Yanchun Liang, and Xiaohu Shi. "Full Text Clustering and Relationship Network Analysis of Biomedical Publications." (2014): e108847

[8] Assar, Saïd, Markus Borg, and Dietmar Pfahl. "Using text clustering to predict defect resolution time: a conceptual replication and an evaluation of prediction accuracy." *Empirical Software Engineering* (2015): 1-39.

[9] Aggarwal, Charu C., Yuchen Zhao, and Philip S. Yu. "On the use of side information for mining text data." *Knowledge and Data Engineering, IEEE Transactions on* 26, no. 6 (2014): 1415-1429.

[10] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." *Journal of Emerging*

*Technologies in Web Intelligence*2, no. 3 (2010): 258-268.

[11] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.