CLUSTERING AND INDEXING OF MULTIPLE DOCUMENTS USING FEATURE EXTRACTION THROUGH APACHE HADOOP ON BIG DATA

E.Laxmi Lydia[1*], G. Jose Moses[2], Vijayakumar Varadarajan[3], Fredi Nonyelu[4], Andino Maseleno[5]
Eswaran Perumal[6], K. Shankar[7]

[1]Professor and Big Data Consultant,Computer Science and Engineering,Vignan's Institute of Information Technology, India

[2]Professor, Computer Science and Engineering, Raghu Engineering College (Autonomous), Visakhapatnam (Andhra Pradesh), India

[3]School of Computer Science and Engineering, The University of New South Wales, Australia.

[4]Chief Executive, Briteyellow Ltd, United Kingdom

[5] STMIK Pringsewu, Lampung, Indonesia

[6,7]Department of Computer Applications, Alagappa University, Karaikudi, India

Email: elaxmi2002@yahoo.com[1*](corresponding author), josemoses@gmail.com[2], v.varadarajan@unsw.edu.au[3], fredi.nonyelu@briteyellow.com[4], andimaseleno@gmail.com[5], eswaran@alagappauniversity.ac.in[6], drkshankar@ieee.org[7]

DOI: https://doi.org/10.22452/mjcs.sp2020no1.8

ABSTRACT

Bigdata is a challenging field in data processing since the information is retrieved from various search engines through internet. A number of large organizations, that use document clustering,fails in arranging the documents sequentially in their machines. Across the globe, advanced technologyhas contributed to the high speed internet access. But the consequences of useful yet unorganized information in machine files seemto be confused in the retrieval process. Manual ordering of files has its own complications. In this paper, application software like Apache Lucene and Hadoop have taken a lead towards text mining for indexing and parallel implementation of document clustering. In organizations, it identifies the structure of the text data in computer files and its arrangement from files to folders, folders to subfolders, and to higher folders. A deeper analysis of document clustering was performed by considering various efficient algorithms like LSI, SVD and was compared with the newly proposed updated model of Non-Negative Matrix Factorization. The parallel implementation of hadoopdevelopedautomatic clusters for similar documents. MapReduce framework enforced its approach using K-means algorithm for all the incoming documents. The final clusters were automatically organized in folders using Apache Lucene in machines. This model was tested by considering the dataset of Newsgroup20 text documents. Thus this paper determines the implementation of large scale documents using parallel performance of MapReduce and Lucenethat generate automatic arrangement of documents, which reduces the computational time and improves the quick retrieval of documents in any scenario.

Keywords: ***Text Mining, Hadoop MapReduce, Indexing, Lucene, Clustering, NMF, K-means***

1.0 INTRODUCTION

Organizations leverage large databases for robust utilization of information retrieved from various sources. Well recognized databases for structured data can be relational, object-oriented or object-relational. Both unstructured as well as semi-structured data are reserved in huge volumes. Text mining and data mining integrate to extract the data and assemble the patterns for detection. TDM(Text and Data Mining) is an improvedapproach to read data.

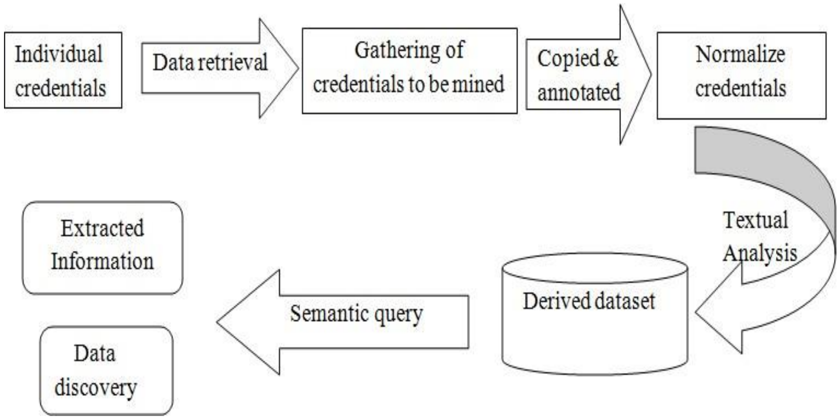


Fig.1: Procedure for text mining

There are four steps listed in the figure 1 above. Initially, all significant credentials are identified. Every single credential considered is then converted into a readable format. Later the data is mined to locate the information, test hypothesis and find the hidden associations. The primary task for gadget file process is to insert the files into folders and those folders in some other folders of highest level. Based on the file information, the documents are manually organized which is a troublesome process. Therefore, to overcome these cons of manual arrangement, computer-aided clustering of documents is being followed. Text mining is an analytics tool used for knowledge-consuming process. It classifies the incoming documents by following major steps like collection of documents and identification ofdocument features, especially the commonly used document features[29].

The main objective of text mining is to transform the text into information so that it can be analyzed. To achieve this, computer-intensive artificial intelligence algorithms and statistical techniques are necessary for the implementation of text documents. It uses a wide variety of tasks, which can be combined into a single working flux and can be distinguished through various technologies used for text data mining such as Information Retrieval, Natural Language Processing, Information Extraction, and Data Mining. A large collection of digital text documents in text mining is used to identify the documents. The information recovery systems used are intended to identify a subset of documents that correspond to the request of a user. The tools used in library searches for books in specific genre and web search engine (e.g. Google, Bing) for searchinginformation on the world wide web are two classic examples of information recovery systems. The character strings must be processed so as to be analyzed by computers, once a subset of text documents has been retrieved. Computers must be fed with a certain amount of input to allow natural languages to be understood by human beings. Generally, the tasks provided by NLP system are to determine the syntax, search for sentence and to determine the significance of the sentence semantically. The unstructured natural language documents must be transformed into data in a structured form so as to be extracted like any other types of data. In this phase, the data generated by NLP systems are referred toinformation extraction. During this phase, the most common task is to identify specific terms that may consist of one or more words, as in scientific research papers which contain lot of multiple complex words. The data extraction also allows one to connect names, entities and more complex facts such as events or name relationships.

The data extracted from annotated documents, provided by NLP algorithms,is processed into structuring database and is finally made ready for extraction. In this context, ‘mining’ is synonymous to ‘analysis’, since the aim is to obtain useful information so as to achieve a new understanding from the text data. In view of the fact that the data is structured now, standard statistical procedures and techniques, used for text data, can be applied. Clustering is an approach in which every texture or object with similar propertiesis combined logically into one class of objects physically and makes the entire class available with one access to the disk. Many clustering methods are available each of which may provide another dataset group.

Domains of text mining applications:

Text miningemergeswith technology and it collects large unstructured documents to extract knowledge. The text mining technology is used in a number of possible applications. Following is a list of few such applications.

- *Analysis of customer profiles*, e.g. mining the incoming customer complaints and feedback emails.
- *Patent analysis* for major technical players, trends and opportunities, for example, patent database analysis
- *Dissemination of information,* e.g. organization and summary of business news and reports for personalized services.
- *Company planning of resources,* e.g. mining reports on company activities, statuses and reported problems.
- *Matters related to security,* e.g. the analysis of plain text sources like internet news. It also includes the assessement of text encryption.
- *Open-ended survey responses*, such as the analysis of certain words or terms which respondents commonly use to describe pro’s or con’s (under investigation) of any products or services with regard to the subject matter of the study.

2.0 LITERATURE SURVEY

B. MeenaPreethi et al [1] recommended various applications on digital data and its issues. As the technology is getting advanced, the volume of data generated is also getting increased. To restore the process of unstructured data, scientists have recommended compulsory identification of patterns while processing, transforming, extracting, mining and evaluating the unstructured data. Eventually this scenario improves the efficiency and the goal to generate an application on processing the unstructured text document with proper selection of techniques can be attained.

Meiping Song et al [2]implied spatial-based calculations on matrix data over non-negative factorization.Various combination of pixels are regulated through spatial resolutionthat transversely obtain the results according to object recognition and classification. Varied mixture models like linear, bi-linear and non-linear are used to analyze data in spectral unmixing approach, which tends to interpret the fractions produced through hyperspectral imagery activity. Moreover, it is observed that the NMF is carried out mostly with linear mixtual modelby considering new constraint-based extraction of features and smoothness in an effective manner. When the data is compressed, one can obtain thereduction of convergence and direct decomposition of pixels with local minimum.

DipeshShrestha[3] pointed to sentiment analysis in text mining.The research focused on removal of noise in large amounts of data using modernized techniques and different software tools. Steps like data acquisition, preprocessing, feature extraction and labelling of data were performed in this research.

Yu-Xiong et al [4]interpreted highly-accurate optimized clustering algorithms such as K-means and PSO. This resulted in the reduction of major issues relevant to unstructured data. K-means clustering algorithm found an apt solution to resolve major issues by performing the relevant procedures.The priority of this algorithm is to initialize the clusters which can again be advanced with hybridization of the algorithm (Particle Swarm Optimization). The combination of two algorithms improves its efficiency while the benchmark datasets are considered to perform cluster analysis from machine learning repository.

JiaQiao et al, [5] recommended data managing clustering algorithms in database management system. Based on data structure, three practical real-time based data mining algorithms were proposed to perform cluster analysis. All the three algorithms were compared, analyzed and performed on platforms like healthcare, business, industries., etc that deal with large amounts of unorganized data.

E. Laxmi Lydia et al, [6] proposed a method, for innovative storage and retrieval of information with large amounts of data using NMF rules,with a combination of K-means as KNMF for grouping similar documents. They performed the analysis on unstructured text documents with the most advanced methodologies. The research was conducted in the Hadoop distributed file system structure. The study also compared the analyses of various preprocessing stemming algorithms such as Lovins stemming, porters stemming, iterative Lovins stemming algorithms with respect to index compression factor, words-stemmed factor and correct stemming word factor.

SerhatSelcukBucak et al, [7] suggested NMF as a study of incremental clustering for new samples of essential batch nature.An online processing data representation tool was suggested in this study for massive data.As a solution to optimal rank selection problem, it categorized significant clustersthrough NMFand partitioned methods. To examine the results linearly, separate data was manifested. NMF using Multiplicative approach[9] on parts-based data and NMF to new matrix decomposition method reduce the actual data matrix and yield better results[10].

E. Laxmi Lydia et al, [12] suggested text mining implementation through Hadoop framework for document clustering using Euclidean distance and termed it as ‘frequency-inverse document frequency’. Here the obtained document matrix was specified using each term of the document. Term frequency- Inverse document frequency provides weights to every term and forms a matrix. To check the similarity between the documents, EuclideanDistance Measure was determined.

E. Laxmi Lydia et al, [13] compared and worked on Hadoop cluster architectures (Standalone, Pseudo, Fully-distributed) that have implemented text mining document clustering through prominent well-known big data platform in Hadoop. Hadoop clusters are directed to process the unstructured data as clusters[25]over distributed cluster nodes. The data waspartitioned as chunks to analyse and was clustered into nodes.Based on the scalability of the datasets, the performances of storage and data retrieval were determined among three Hadoop cluster nodes.

E. Laxmi Lydia Maheshwari, [14][15]implemented ‘text mining’ for document clustering that uses Euclidean distance along with K-means algorithm over unstructured data. The researchers considered three documents, calculated each term weight, measured the distance among document terms and clustered it using K-means algorithm. A Non-Negative Matrix factorization with k-means was used for clustering the newsgroup20 dataset. Whereas a key phrase extraction was used in filtering the stopwords data, Iterated Lovins stemming algorithm for extracting the data to exact root, TF_IDF for the formation of document matrix and KNMF for clustering.

A. SudhaRamkumar et al, [23] focused on the work of text document clustering. Within the process of implementation, they selected the processing area for a large amount of data by eliminating irrelevant data using dimensionality reduction. The process of dimensionality reduction was performed by two kinds of approaches in text mining such as reduction and selection of features. These approaches were implemented along with K-means algorithm and estimated InfoGain of K-means. Through overall analysis of the text data, the performance of BBC sports datasetwas identified using K-means with InfoGain (with feature selection) which was found to be higher than K-means without feature selection.

Mounika Gupta et al, [24] conducted a survey analysis in the research area of natural language text documents and identified the purpose of arranging similar text documents together as clusters with large data. Yogapreethi. N et al, [22] also concentrated on the process of data mining and text mining that involve high quality information. The authors adopted analytical methods that are generally used for the purpose of categorization, clustering, and furthermore analysis. U. S. Patki et al, [21] explored various hard clustering techniques to group the documents that exhibitsimilar features over a vast digital generation of knowledge.

M. Uma Maheshwari et al, [27] in their innovative study, processedthe clustering of documents through graph structures using graph fusion model. They constructed the graph based on the features obtained from word frequency and semantic frequency of terms that indeed develop a fusion model graph. KNN, with weighted edges, was correlated to the graph to obtain normalized mutual information. The metrics considered for the study were purity, precision, recall, and F-measure.

Avanish Singh et al, [18] worked on virtual machine that provides services to Hadoop framework from docker. They designed a method called ‘teragen’ using terasort tool that provokes indiscriminate data over Hadoop distributed file system clusters[16]. The study further performed the benchmark testing, evaluated the process of virtualization while optimizing time was also deployed. The experiment was performed by considering five virtual machines.

Yojna Arora et al, [26] shown the parallel processing of data using Hadoop frameworks such as general-purpose processing, abstraction frameworks, SQL frameworks, graph processing frameworks, machine learning frameworks and streaming frameworks. Large organizations like Google, Facebook and so on leverage Hadoop stack to attain better analytical results. K. Tamilselvi et al., [19] also focused onHadoop technology with the term ‘parallelism’ for generation of data using big data analytics. C. S. Arage et al, [20] carried out their work on Electronic Health Records (EHR) databases using Hadoop cluster technology to enhance the performance of healthcare system for delivery and processing of cost-saving and quick transfer of medication for patients. Jayalakshmi DS et al, [17] conducted a study on various research issues and various comparisons among the deployment procedures that exist in Hadoop on cloud platforms.

Bikram Keshari Mishra et al, [28] implemented previously-existed clustering algorithms and proposed the Far Enhanced Clustering Algorithm for internal cluster effectiveness of indices such as Dunn’s, Davies-Bouldin’s, C, Calinski indexes and Silhouette coefficient for significant assessment of clusters. Abhay Kumar et al, [8] determined clustering algorithms which anticipate smiliar characteristic approaches in existing data.

AmreenKausaur et al, [36] performed effective analysis over documents, identified challenges associated with natural language processing upon text mining of digital libraries, healthcare and life science, sentiment analysis, business intelligence, analysis of open-ended survey responses, automatic filtering and competitor investigation using different approaches like term-based mechanism, phase-based mechanism, concept-based mechanism and pattern taxonomy mechanism.Zainab Zaveri et al, [38] selected a method for text summarization for instant implementation using frequency-based approach, frequency recognition strategy, keyword frequency technique and k-implies grouping.SakshiBhalla et al, [32] compared the text summarization using three prominent methods (Extraction-based, Abstraction-based and aided summarization) on fuzzy logic, neural networks, deep learning and semantic analysis.

R. K. Jeyauthmigha et al, [34] recognized an efficient way to cluster the data through feature reduction and recursive feature elimination procedures using a two-phase mechanism. This research considered network anomaly systems using KDD cup 1999 dataset. The functioned techniques were K-means, Hierarchical Agglomerative and Density-Based Clustering. Chouhan. R et al., [39]recommended two prominent algorithms for document clustering. PSO algorithm attainedthe optimal stage in search space. A further process was continued using K-means algorithm by considering the state space points.

G. L. AnandBabu et al, [31] advised conceptual-based techniques for various schemes and actions performed in text mining to extract the beneficial information which will be more helpful in text miningapplications. The experimented tools on information retrieval were intelligent miner and text analyst while for extraction of information, text finder and clear forest. In case of summarization of text, tropic tracking tool and sentence ext tool were used and for categorization of text, intelligent miner was used. For clustering of text, carrot and rapid miner were used.

Swayanshu Shanti [33] reported some of the portioned clustering techniques that can easily classify the bigdata by recognizing various statistical and computational objections using FCM and K-means mechanisms.Sajitha et al, [35] proposed an approach to cluster the data using MapReduce by applying hash code and DNA encryption methods. Iva Pauletic, [11] implemented some clustering performance measures to assess the applications developed using K-means algorithm. Vivegapriya et al,[37] suggested multilingualism in Natural Language Processing by applying graph-based unsupervised algorithms which support in finding the frequency of the co-occurrence.

3.0 METHODOLOGY

3.1 Text mining system architecture

The text mining system collects documents as inputs and then processes each document by checking its sets of formats and characters. The pre-processed documents then go through a phase of text analysis and repeat the techniques at the instance till necessary information is extracted. Additional combinations could be used based on the organization's objectives. The information collected can be inserted into ainformation management system that provides the user with a wealth of knowledge. Each step in the text mining system is illustrated in the figure 2.

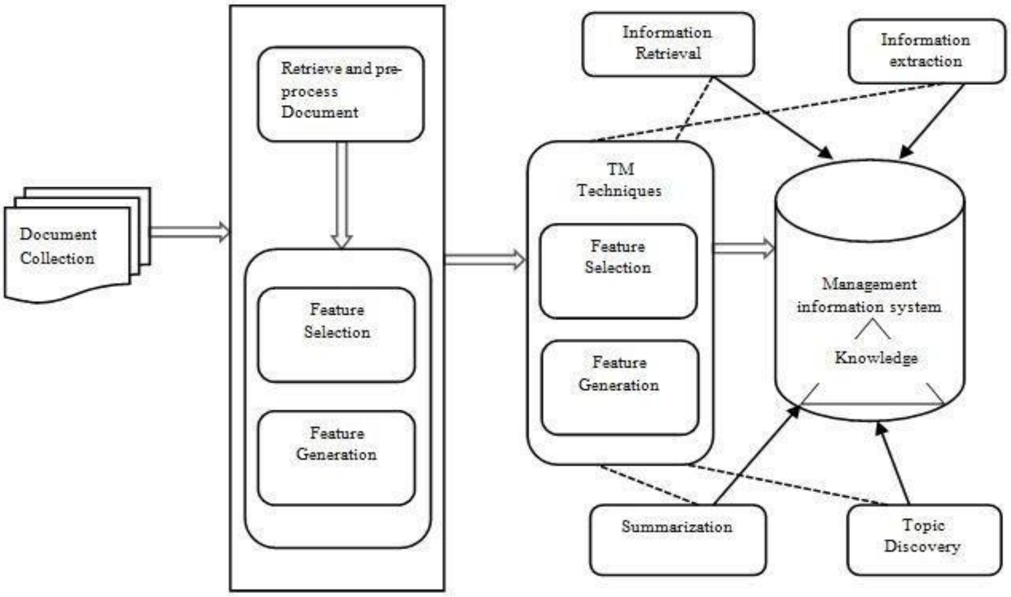


Fig.2: Text mining process

Various document format files in chat, sms, emails, boards of messages, newsgroups, blogs and wiki or websites, are collected from different sources. This unstructured document dataset is pre-processed to carry out three tasks:

- Use the space to delimit the file to individual tokens
- Remove stop words that have no meaning for clustering process
- Use the algorithm for stemming words with a common root word

In order to represent unstructured text records in a highly structured format, the functions for extraction wereprocessed. The selection of algorithms helpsin the identification of key features that require a thorough search of the entire cardinal characteristics subsets, where large numbers are available. It is impractical to find the satisfactory features instead of an ideal setting for supervised learning algorithms.

The appropriate choice includes text mining applications such as data recovery, information extraction, resuming and the discovery of topics for necessary discovery process for knowledge[30].

3.2 Unsupervised machine learning

The algorithm for the machine study investigates the data to identify patterns. There is no response key or instruction from a human operator. By analyzing the available data, the machine determines the correlations and relations. The machine learning algorithm is used in an uncontrolled learning process to interpret large sets of data and to deal with these data accordingly. In order to define its structure, the algorithm attempts to organize this information. This could lead to data, being grouped into clusters, or arranged in a highly organized way. The unsupervised machine learning task is to group the non-sorted data that has similar patterns. The evaluation of data gradually improves and refines the capacity to make decisions. Uncontrolled learning is classified into two categories such as clustering and association. A problem with clustering is, where you want the data groupings inherent, to be found. A problem with association rule of learning is that, you want to find rules that describe large parts of your information.

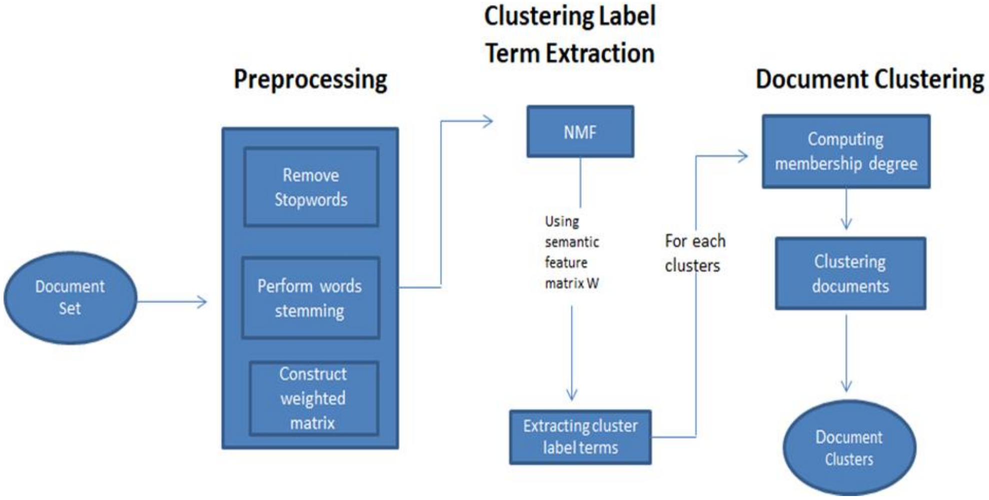

Preprocessing
Clustering Label Term Extraction
Document Clustering
Document Set
Remove Stopwords
Perform words stemming
Construct weighted matrix
NMF
Using semantic feature matrix W
Extracting cluster label terms
For each clusters
Computing membership degree
Clustering documents
Document Clusters

Fig.3: Proposed procedure for document cluster

Clustering is a mechanism by which the points are split into similarity-based classes. The process for K means is as follows, if the value of K is given as an input.

- Divide the objects or data into K-subsets in which the data is not null
- Recognize the mean point of the clusters for current split
- Allocate each point to a particular cluster
- Calculate the distances from each point and then allocate the points to the cluster based on the minimum distance from the center
- After rearranging the points, calculate the new mean based on the newly-assigned points.
- To run the K-means parallel, Hadoop is executed in local and pseudomodes

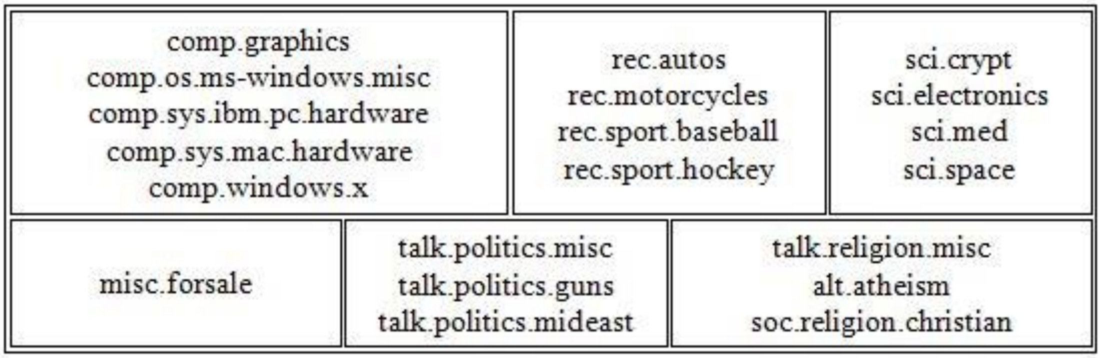
Hadoop is purely a free source programming framework based on java that promotes the dispensation of huge database in a shared computer environment. For processing and sharing information, Hadoop leverages low costcustomary servers in the business. The important characteristics of Hadoop are expense powerful model, adjustability, simultaneous processing of shared information, optimization of local information mechanical fail over the organization and subside more clusters of nodes. Two important components present in the Hadoop structure are HDFS and MapReduce framework. The structure Hadoop separates the data into small chunks and keeps all the data in a specific cluster node. This reduces the time required to store the data in the frame. Hadoop duplicates each part of the information onto another gadget within the cluster in order to provide the data during any circumstances. The number of duplicates depends on the factor of reproduction.

4.0 RESULT ANALYSIS

Considering newsgroup20 text documents as input datasets, the experimentation was carried out using Hadoop and MATLAB. Hadoop software performed the preprocessing of document clustering using stopwords, stemming, TF-IDF and KNMF algorithms. Through parallel processing, the documents were grouped into similar document folders. Apache Lucene Software performed easy way of indexing the documents and a quick way to search any document. MATLAB software is generally used to plot the graph accurately for searching and indexing the documents.

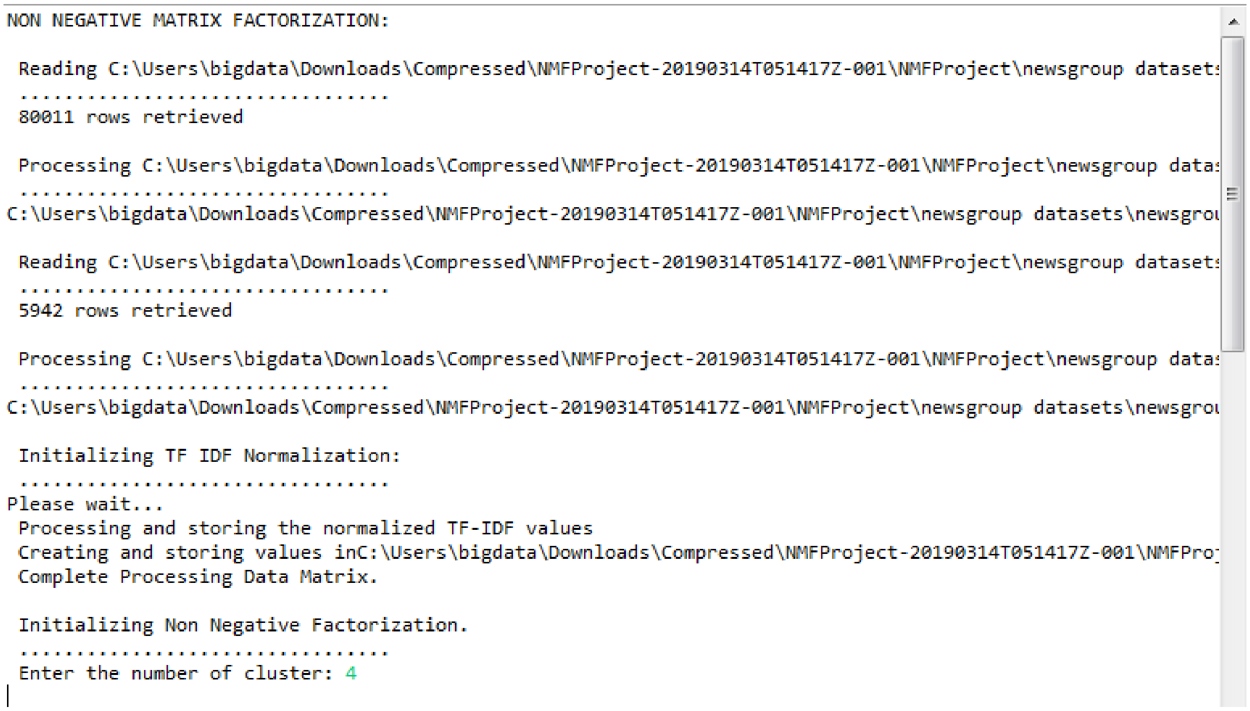
The following table1 has Newsgroup20 dataset articles. Each contains separate files individually, which are pre-processed, extractedand clustered.

Table 1: Different articles containing text documents in Newsgroup20 dataset



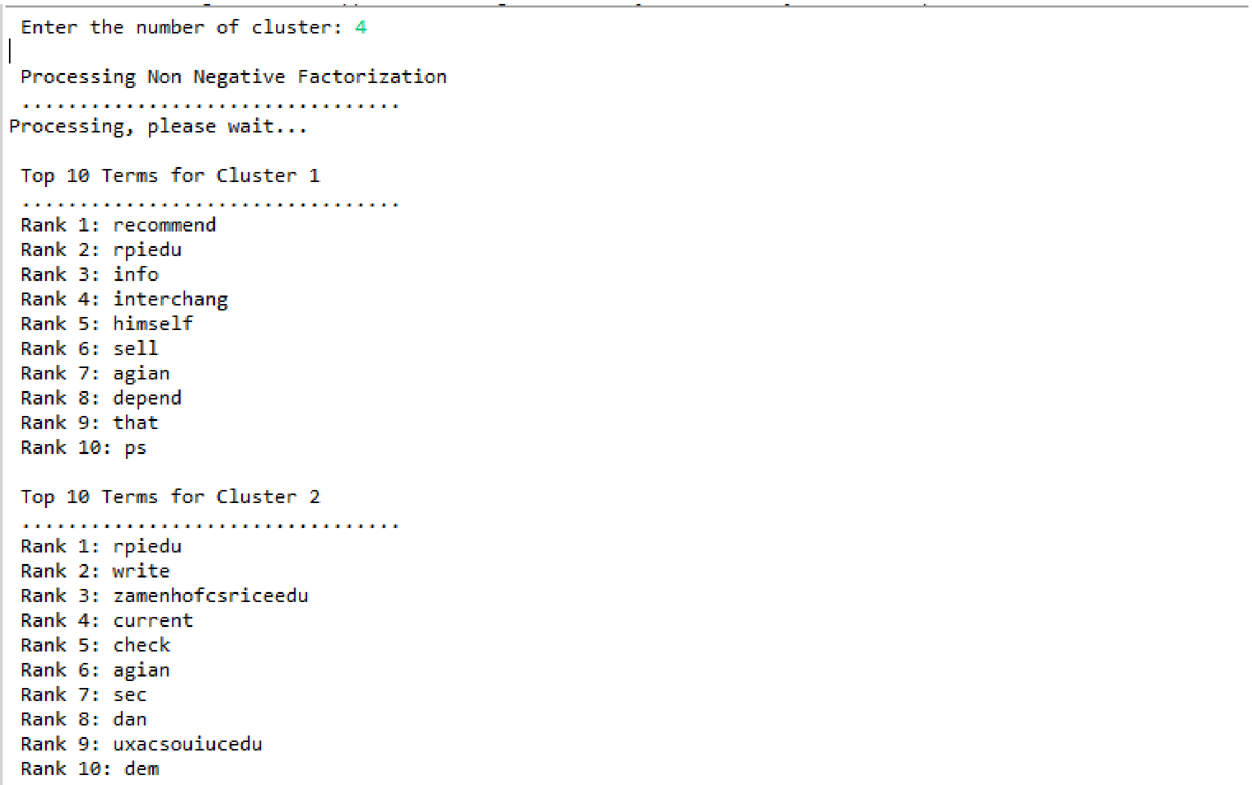
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Figure5 describes the implementation of NMF algorithm by initializing TF_IDF values. Here the values were read by rows and columns. After the initialization of TF_IDF values, K-means clustering algorithm was followed along with NMF, therefore the number of clusters need to be declared. Figure6 describes the Output for Processing Non-Negative Matrix Factorization of four clusters by presenting top 10 terms for cluster1 and cluster2.Figure7 describes the Output for Processing Non-Negative Matrix Factorization of four clusters by presenting top 10 terms for cluster3 and cluster4.Figure8 describes the GUI for automatic indexing and searching text documents using Apache Lucene and text from the selected documents by Click File menu and open File.Figure9 shows the selection of text documents from the data folder through open dialogfor automatic indexing and searching text documents using an interface. Figure 10 shows theselection of single or multiple documents for automatic indexing and searching text documents using a GUI interface.Figure 11 loads the input text documents from the data folder.Figure 12 shows the selection of single or multiple words which need to be searched from the indexed data.Figure 13 shows the final output for indexing and searching.



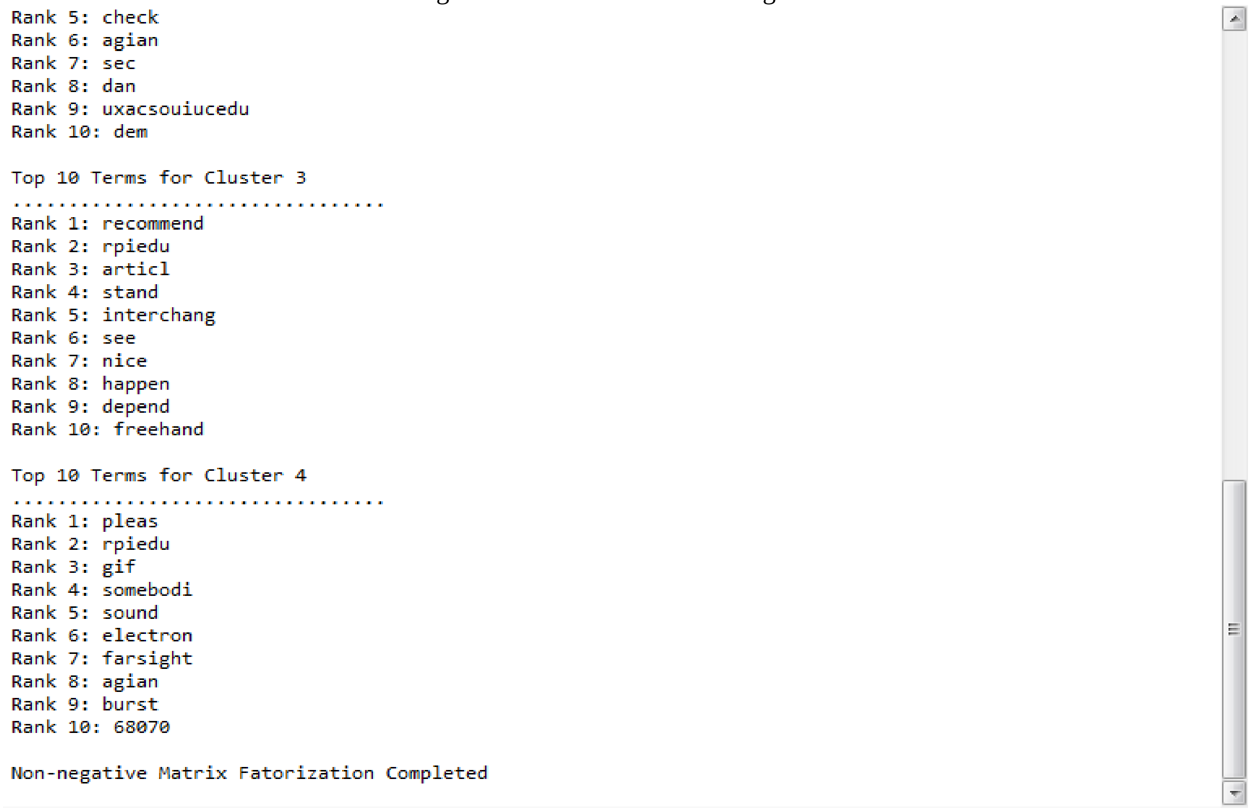
```
NON NEGATIVE MATRIX FACTORIZATION:

 Reading C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datasets
 ................................
 80011 rows retrieved

 Processing C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datas
 ................................
C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datasets\newsgrou

 Reading C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datasets
 ................................
 5942 rows retrieved

 Processing C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datas
 ................................
C:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFProject\newsgroup datasets\newsgrou

 Initializing TF IDF Normalization:
 ................................
Please wait...
 Processing and storing the normalized TF-IDF values
 Creating and storing values inC:\Users\bigdata\Downloads\Compressed\NMFProject-20190314T051417Z-001\NMFPro
 Complete Processing Data Matrix.

 Initializing Non Negative Factorization.
 ................................
 Enter the number of cluster: 4
```

Fig. 5: Initialization of TF-IDF to cluster the data using NMF



```
 Enter the number of cluster: 4

 Processing Non Negative Factorization
 .................................
Processing, please wait...

 Top 10 Terms for Cluster 1
 .................................
 Rank 1: recommend
 Rank 2: rpiedu
 Rank 3: info
 Rank 4: interchang
 Rank 5: himself
 Rank 6: sell
 Rank 7: agian
 Rank 8: depend
 Rank 9: that
 Rank 10: ps

 Top 10 Terms for Cluster 2
 .................................
 Rank 1: rpiedu
 Rank 2: write
 Rank 3: zamenhofcsriceedu
 Rank 4: current
 Rank 5: check
 Rank 6: agian
 Rank 7: sec
 Rank 8: dan
 Rank 9: uxacsouiucedu
 Rank 10: dem
```

Fig. 6: Formationof clusters using NMF



```
 Rank 5: check
 Rank 6: agian
 Rank 7: sec
 Rank 8: dan
 Rank 9: uxacsouiucedu
 Rank 10: dem

 Top 10 Terms for Cluster 3
 .................................
 Rank 1: recommend
 Rank 2: rpiedu
 Rank 3: articl
 Rank 4: stand
 Rank 5: interchang
 Rank 6: see
 Rank 7: nice
 Rank 8: happen
 Rank 9: depend
 Rank 10: freehand

 Top 10 Terms for Cluster 4
 .................................
 Rank 1: pleas
 Rank 2: rpiedu
 Rank 3: gif
 Rank 4: somebodi
 Rank 5: sound
 Rank 6: electron
 Rank 7: farsight
 Rank 8: agian
 Rank 9: burst
 Rank 10: 68070

 Non-negative Matrix Fatorization Completed
```

Fig. 7:Formation of clusters using NMF

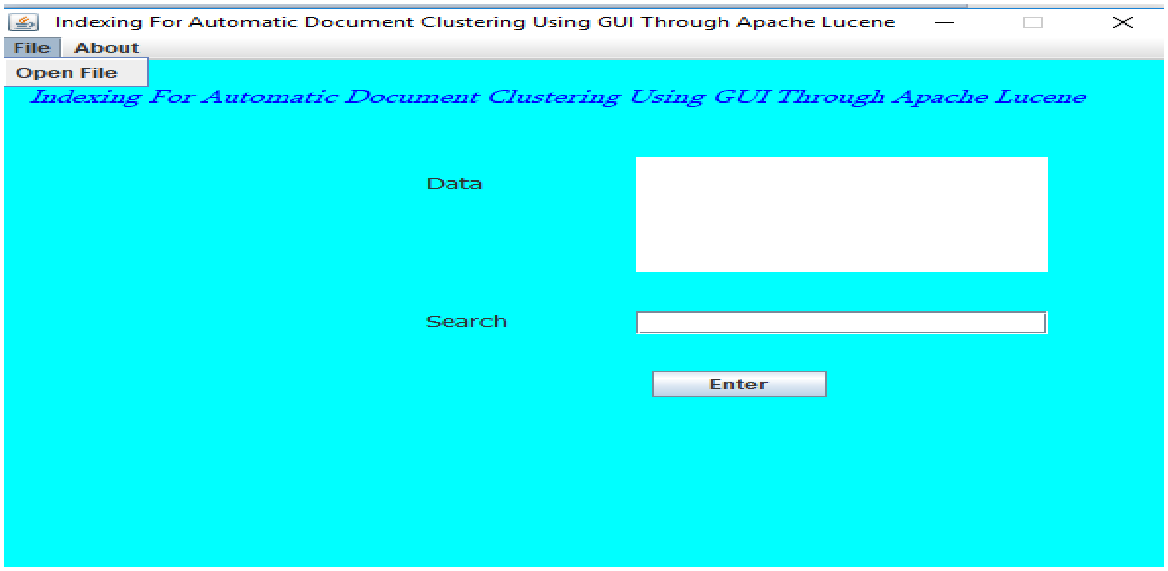
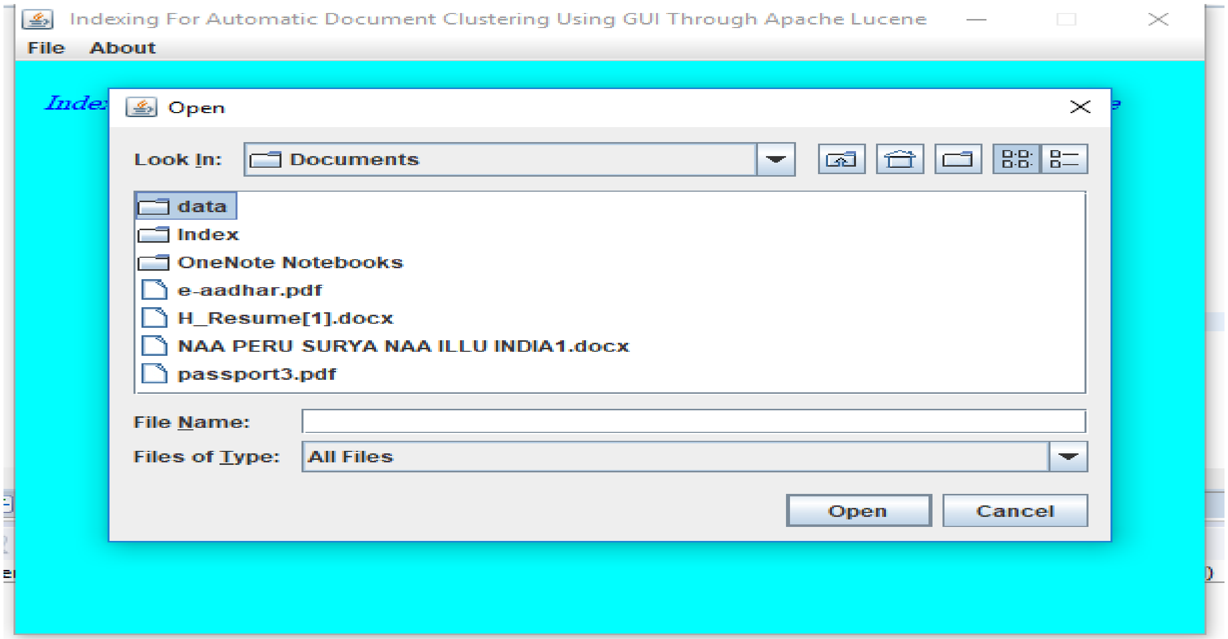

Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
File
About
Open File
Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
Data
Search
Enter

Fig. 8:

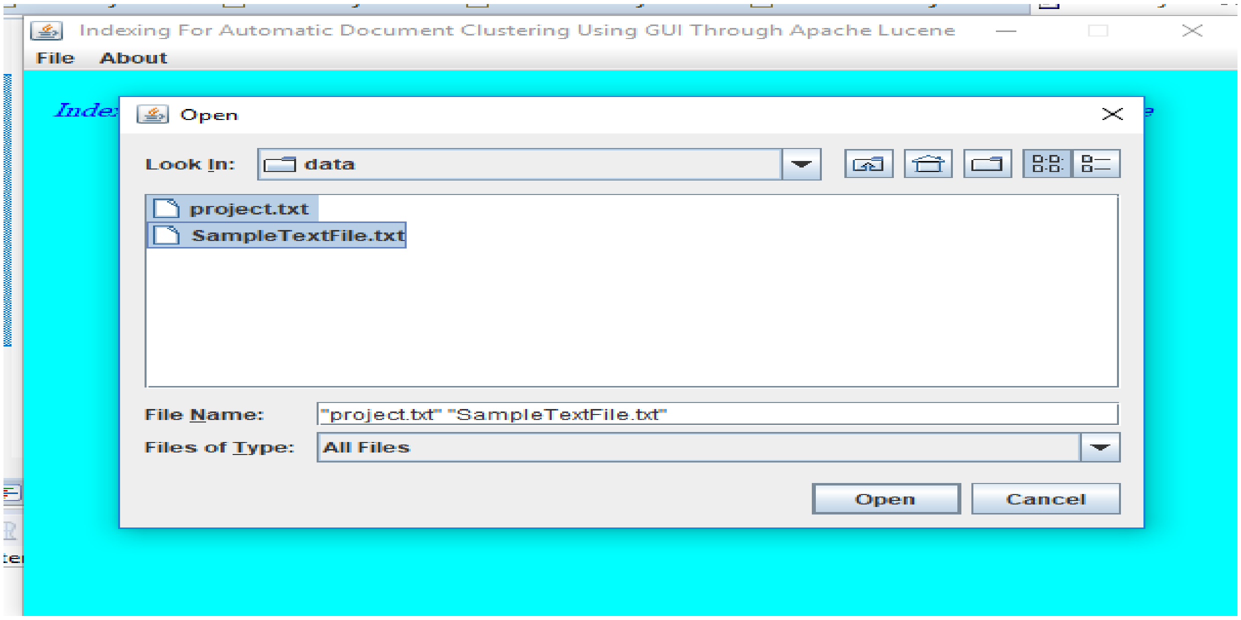
Interface to load and search data in documents


Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
File
About
Open
data
Index
OneNote Notebooks
e-aadhar.pdf
H_Resume[1].docx
NAA PERU SURYA NAA ILLU INDIA1.docx
passport3.pdf
File Name:
Open
Cancel

Look In: Documents

Files of Type: All Files

Fig. 9: Choose the data folder from open dialog


Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
File
About
Open
Open
Cancel

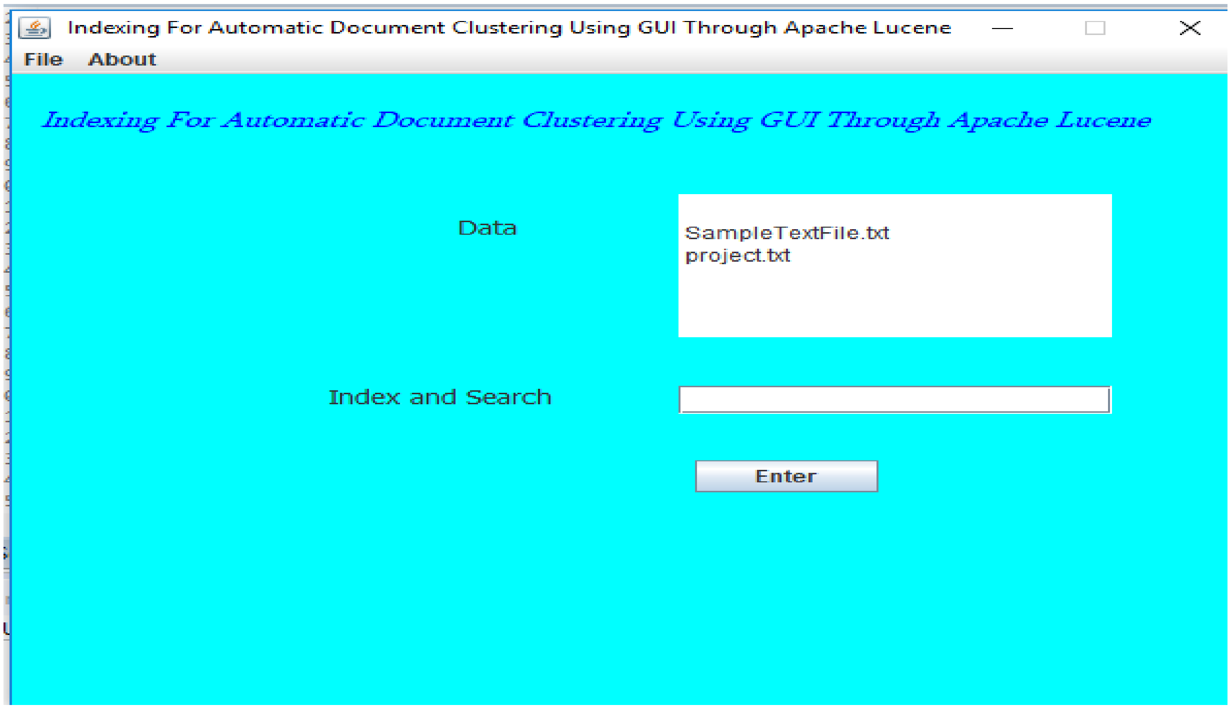
Look In: data

project.txt
SampleTextFile.txt

File Name: "project.txt" "SampleTextFile.txt"

Files of Type: All Files

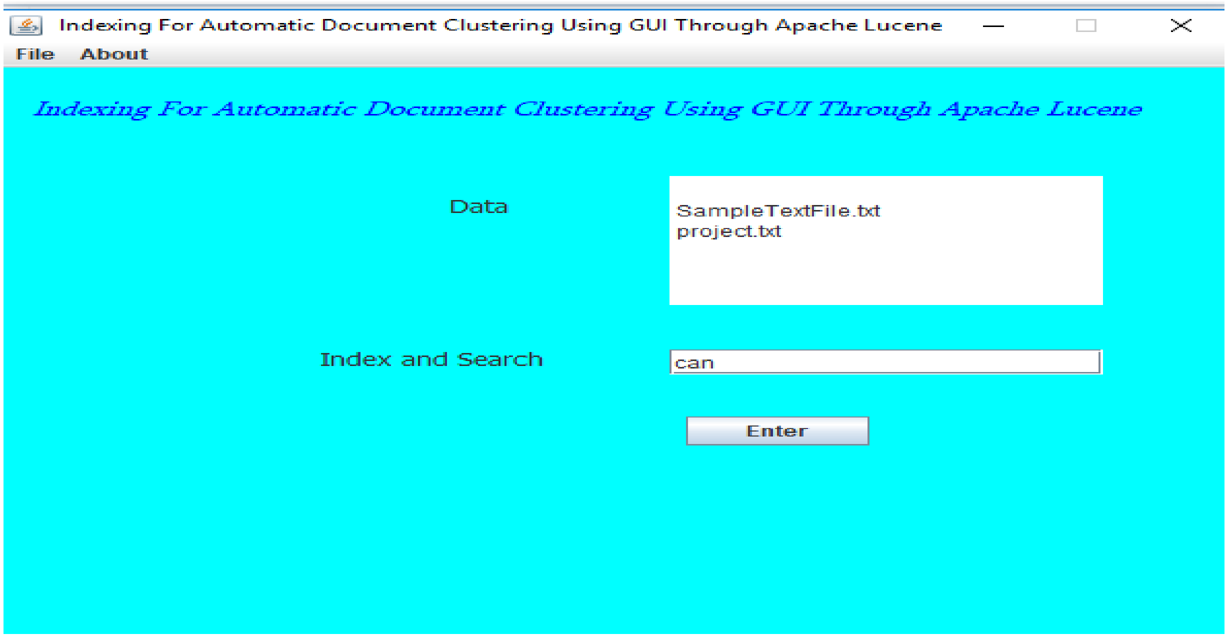
Fig.10: Selection of documents


Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
File
About
Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
Enter

Data: SampleTextFile.txt
project.txt

Index and Search:

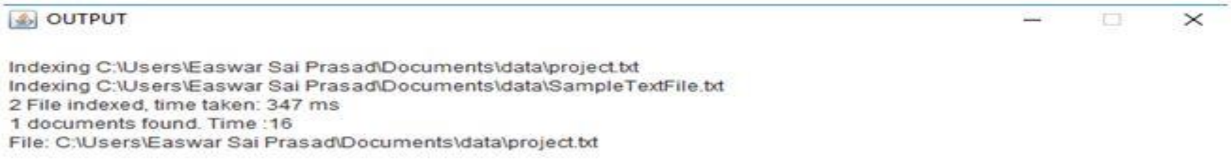
Fig.11: Selection of data from folders


Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
File
About
Indexing For Automatic Document Clustering Using GUI Through Apache Lucene
Enter

Data: SampleTextFile.txt
project.txt

Index and Search: can

Fig.12: Selection of single or multiple words



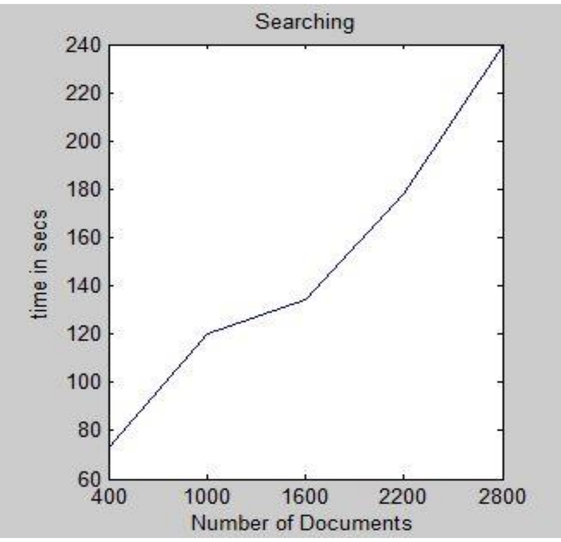
```
Indexing C:\Users\Easwar Sai Prasad\Documents\data\project.txt
Indexing C:\Users\Easwar Sai Prasad\Documents\data\SampleTextFile.txt
2 File indexed, time taken: 347 ms
1 documents found. Time :16
File: C:\Users\Easwar Sai Prasad\Documents\data\project.txt
```

Fig.13: Output for indexing and searching

Figure 14 represents the graph analysis on-time performance for minimum to maximum number of documents (i.e, 400 to 2800 documents) to search the keyword from multiple number of documents. X-axis describes the number of documents and y-axis describes the search time in seconds. The graph analyzes the increase in time with increase inthe number of documents.

Figure15 represents the graph analysis on-time performance for minimum to maximum number of documents (i.e, 400 to 2800 documents) to index all the input documents. X-axis describes the number of documents and y-axis describes the search time in seconds. The graph analyzes the increase in time with increase inthe number of documents.

This identifies the quick execution of search and indexing within seconds for a large number of text documents.



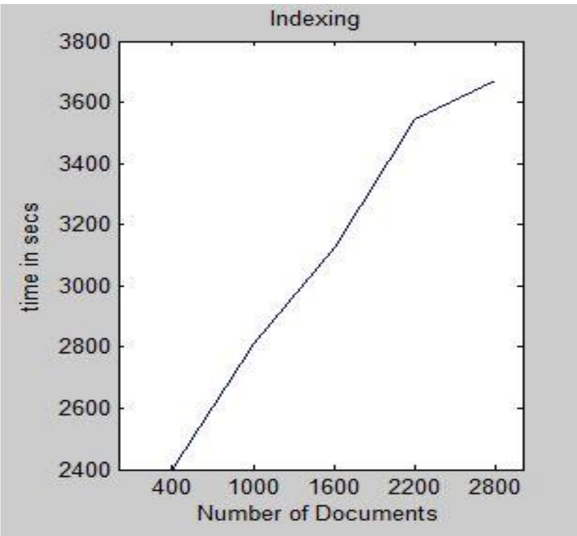


Fig.14:MATLAB graph for searchingFig.15:MATLAB graph for indexing

5.0 CONCLUSION

As history and trends move ahead, innovations leave enormous amounts of data behind. This massive ‘big data’ approach to process and analyzethe data has helped the businesses with database administration tools and traditional data processing applications. The results of the proposed method are shown in the paper by considering Newsgroup20 dataset which was used for clustering the documents and to receive top words of aparticular cluster. Initially, pre-processing techniques were used on a dataset to obtain the matrix format. The NMF can process through only matrix format documents and it can read the dataset which is in the form of a matrix that retrieves the rows of the document. The matrix is then read after processing through NMF. To obtain the TF-IDF values, the terms of the dataset should also be read and processed. This paper proposes a new model with updated rules called NMF, with a combination of K-means for document clustering and application development based on this medium. The developed model can be used to organize folders in such a way that the documentation can be divided into subfolders without any knowledge of content. Therefore, the performance in the retrieval of documents, in any situation, really increases. The accuracy of the proposed model was tested using map-reduce implementation of K-means from Apache Hadoop project.

ACKNOWLEDGMENT

The Department of Science and Technology (DST), Science and Engineering Research Board (SERB) financially support this work under the scheme of ECR. We thank DST-SERB for the financial support to carry the research work.

REFERENCES

[1] Mrs.B. MeenaPreethi, Dr. P. Radha, *“A Survey Paper on Text Mining - Techniques, Applications And Issues”,* IOSR Journal of Computer Engineering(IOSR-JCE), e-ISSN: 2278-0661, p-ISSN:2278-8727, pp 46-51.

[2] Meiping Song, Qiaoli Ma, Jubai An & Chein Chang, *“An Improved NMF Algorithm Based on Spatial and Abundance Constraints”,* 2016 progress in electromagnetic research symposium(PIERS), Shanghai, China, 8-11 August, pp 4532-4537.

[3] Dipesh Shrestha,*”Text Mining with Lucene and Hadoop:Document Clustering with feature extraction”,* thesis - WakhokUniversity, 2009.

[4] Yu-XiongWang,Yu-Jin Zhang, *"Nonnegative Matrix Factorization: A comprehensive review*", IEEE Transactions on knowledge and data engineering, Vol.25, No.6, June 2013,pp 1336-1353.

[5] Jia Qiao& Yong Zhang,*"Study of K-means Method Based on Data-Mining",*2015 Chinese Automation Congress(CAC). DoI: 10.1109/CAC. 2015.7382468.

[6] E. Laxmi Lydia and D. Ramya,*"Text Mining With Lucene And Hadoop: Document Clustering With Updated Rules Of NMF Non Negative Matrix Factorization",* International Journal of Pure and Applied Mathematics, Volume 118, No.7 2018, pp 191-198.

[7] Serhat Selcuk Bucak and Bilge Gunsel,*"Incremental Clustering via Nonnegative Matrix Factorization",*2008 19th International Conference on Pattern Recognition. DoI: 10.1109/icpr.2008.4761104.

[8] Abhay Kumar, Ramnish Sinha, Daya Shankar Verma and Vandana Bhattacherjee Satendra Singh, "*Modeling using K-Means Clustering Algorithm",*2012 1st International Conference on recent Advances in Information Technology.

[9] Chengbin Peng, Ka-Chun Wong, Alyn Rockwood, Xiangliang Zhang and Jinling Jiang, David Keyes,*"Multiplication Algorithms for Constrained Nonnegative Matrix Factorization",* IEEE computer society, 2012 IEEE 12th International Conference on data mining. DoI- 10.1109/ICDM.2012.106.

[10] Jie Tang, Xinyu Geng and Bo Peng*"New methods of Data Clustering and Classification based on NMF"* , 2011 International conference on business computing and Global informatization. DoI:10.1109/bcgin.2011.114.

[11] Iva Pauletic, Lucia NacinovicPrskalo, and MarijaBrkicBakaric, "*An Overview of Clustering Models with an Application to Document Clustering*", 2019 42nd International Convention on Information and Communication Technology, Electronics, and Microelectronics. DOI:10.23919/mipro.2019.8756868.

[12] E. Laxmi Lydia, K. Vijaya Kumar, P. Amaranatha Reddy, D. Ramya, "*Text mining with Hadoop: Document Clustering with TF_IDF and Measuring Distance using Euclidean*", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10,14-Special Issue, 2018.

[13] E. Laxmi Lydia, Gorapalli Chandra Sekhar, MadhuBabuChevuru, DasariRamya, K. Vijaya Kumar, "*Text Mining with Apache Hadoop over different Hadoop Clusters Architectures*", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Vol.8 Issue-2, July 2019.

[14] E. Laxmi Lydia, P. Govindaswamy, SK. Lakshmanaprabu, D. Ramya, "*Document Clustering based on text mining K-Means Algorithm using Euclidean Distance Similarity*", Journal of Advanced Research in Dynamical & Control Systems, Vol-10, 02-Special Issue, 2018.

[15] E. Laxmi Lydia, P. Krishna Kumar, K. Shankar, S. K. Lakshmanaprabu, R.M. Vidhyavathi, AndinoMaseleno, "*Charismatic Document Clustering through novel K-Means Non-Negative Matrix Factorization (KNMF) Algorithm using Key Phrase Extraction*", International Journal of Parallel Programming, Springer 2018, https://doi.org/10.1007/s10766-018-0591-9.

[16] Sujit Roy, Subrata Kumar Das, IndraniMandal, "*Hadoop Periodic Jobs using data blocks to achieve efficiency*", International Journal of Scientific Research in Computer Science Engineering and Information Technology, ISSN: 2456-3307, Vol.3, Issue3, 2018.

[17] JayaLakshmi D S, SyedaRabiyaAlam, R. Srinivasam, "*Approaches to Deployment of Hadoop on Cloud Platforms: Analysis and Research Issues*", IEEE International Conference on Recent Trends in Electronics Information & Communication (RTEICT), May 2017.

[18] Avanish Singh, P. Gouthaman, ShivankitBagla, and AbhishekDey, "*Comparative study of Hadoop over containers and Hadoop over Virtual Machine*", International Journal of Applied Engineering Research, ISSN 0973-4562, Vol.6 Issue 6, 2018, 4373-4378.

[19] K. Tamilselvi, V. Sumithra, K. Dhanapriyadharsini, "*Big Data Analytics using Hadoop Technology*", International Research Journal of Engineering Technology (IRJET), e-ISSN: 2395-0056, Vol.05 Issue 01, Jan 2018.

[20] Mr. C.S.Arage, M. P. Gaikwad, RohitTadasare, RonakBhutra, "*Analyse Big Data Electronic Healthcare Records Database using Hadoop Cluster*", International Research Journal of Engineering and Technology(IRJET), e-ISSN: 2395-0056, Vol.05 Issue 03, Mar 2018.

[21] U.S. Patki, Dr. P.G. khot, "*A Literature Review on text Document Clustering Algorithms used in text mining*", Journal of Engineering Computers & Applied Sciences (JECAS), ISSN: 2319-5606, Vol.6 No.10, October 2017.

[22] YogaPreethi. N, Maheshwari. S, "*A Review on Text Mining in Data Mining*", International Journal on Soft Computing(IJSC), Vol.7 No.2/3, August 2016.

[23] A. SudhaRamkumar, Dr. B. Poorna, "*Text Document Clustering using Dimension Reduction Technique*", International Journal of Applied Engineering Research, ISSN 0973-4562, Vol.11 No.11, 2016, 4770-4774.

[24] Monika Gupta, Kanwal Garg, "*A Review on Document Clustering*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.6 Issue 5, May2016.

[25] E. Laxmi Lydia, M. Ben Swarup, C. Narsimham, "*A Disparateness-aware Scheduling using K- Centroids Clustering and PSO techniques in Hadoop Cluster*", International Journal of Advanced Foundation Research Computation, 2(12), 2015.

[26] Yojna Arora, Dr. Dinesh Goyal, "*Hadoop: A framework for BigData processing 7 Storage*", International Journal of Application or Innovation in Engineering of Management(IJAIEM), ISSN 2319-4847, Vol.6 Issue 7, July 2017.

[27] M. Uma Maheshwari, J.G.R. Sathiaseelan, " *Text Graph- An enhanced Graph Fusion Model for Document Clustering*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol.8 Issue.782, May 2019.

[28] Bikram Keshari Mishra & Amiya Kumar Rath, 2018."*Improving the efficacy of clustering by using far enhanced clustering algorithm*", International Journal of Data mining, Modelling, and Management, Vol. 10(3), 269-292.

[29] Poonam Goyal, N. Mehala, Divyansh Bhatia, Navneet Goyal, 2018. "*Topical document clustering: two-stage post-processing technique*", International Journal of Data Mining, Modelling, and Management, Vol. 10(2), 127-170.

[30] C. Uma, S. Krithika, C. Kalaivani, "*A survey paper on text mining techniques*", International Journal of Engineering Trends and Technology, Vol.40(4), 225-229, 2016.

[31] G. L. AnandBabu, B. Srinivas, "*A Conceptual Based Approach in Text Mining: Techniques and Applications*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol.8(7), 2019.

[32] SakshiBhalla, Roma Verma, KusumMadaan, "*Comparative Analysis of Text Summarization Techniques*", International Journal of Engineering Research &Technology(IJERT), ISSN: 2278-0181, Vol. 5(10), 2017.

[33] Swayanshu Shanti Pragnya, "*Clustering: Review on Partitioned Clustering Algorithms*", International Journal of Engineering Research &Technology(IJERT), ISSN:2278-0181, Vol.6 (06), 2017.

[34] R. K. Jeyauthmigha, R. C. Suganthe, "*Efficient Clustering technique with Feature Reduction Mechanism for Network Anomaly Detection*", International Journal of Engineering Research & Technology(IJERT), ISSN: 2278-0181, Vol.6(08), 2018.

[35] Sajitha N, Bhagyalakshmi KC, Bhagyashree, Chaitra C M, "*MapReduce based K-Means Clustering over large-Scale Dataset*", International Journal of Engineering Research & Technology(IJERT), ISSN: 2278-0181Vol.7(06), 2018.

[36] AmreenKausarGorvankolla, Rekha B. S, "*Application of Text mining in Effective Document Analysis: Advantages, Challenges, Techniques and Tools*",International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol.6(04), 2017.

[37] N. Vivegapriya, A. Monika, "*Unsupervised method for processing Unstructured Dataset for Multilingual*", International Journal of Engineering Research &Technology(IJERT), ISSN: 2278-0181, Vol.5(15), 2017.

[38] Zainab Zaveri, DhruvGosain, "*Automatic text Summarization*", International Journal of Engineering Research &TechnologyIJERT), ISSN: 2278-0181, Vol.5(19), 2017.

[39] Chouhan, R., Purohit, A. "*An approach for document clustering using PSO and K-Means algorithm*", International Conference on Inventive Systems and Control (ICISC).1380-1384, 2018.DOI:10.1109/icisc.2018.8034.