

HEB


The Task of Clustering In Data Mining

CASS

Manisha Soni¹, Reshu Grover²^{1,2}Computer Science & Engineering, LIET Alwar, Rajasthan, India¹manishasoni1013@gmail.com, ²reshugrover3@gmail.com**Email ID- serviceheb@gmail.com****ABSTRACT:**

As the name concern Unstructured data is the data having no pre-defined format i.e text files, documents on the website, email, pdf, tweets, customer reviews and so on. A huge amount of textual data is increasing quickly, the ability to summarize, understand and make sense of such data for making better retrieval of appropriate data remains challenging. Existing methods are appropriate for analysis of structured data but these methods are not appropriate for large volume of unstructured data in order to extract useful knowledge. The main motivation of this paper is to understand and exploring the huge collection of unstructured data and discovered the useful knowledge from massive amount of data in less time to improve performance and time by using appropriate data mining algorithm.

Keywords: Unstructured data, Text, Data Mining Techniques Clustering

Access this Article Online	Quick Response Code: 
Website: http://heb-nic.in/cass-studies	
Received on 23/05/2019	
Accepted on 29/05/2019 © HEB All rights reserved	

1. INTRODUCTION

The term unstructured is used in every form of online and offline. It is not limited to the field of computer science and information technology, now it is used in every field of real-life applications. Data explosion, and 91% of the appellant in the survey say they are aware of unstructured data files present in their business. More than one-fourth of the survey appellant now say that the majority of their business data is unstructured. Unstructured data in their business will exceed structured data within next three years by Industry group [1]. Today knowledge becomes the biggest asset of all companies so maximum of the knowledge is recorded in unstructured format. Unstructured data is a significant part of the Big Data explosion.

2. DATA MINING OF UNSTRUCTURED DATA

Unstructured data: Unstructured data is data comes from machines generated or human generated and it is classified into two types.

- i) Non-Textual unstructured data includes multimedia data like images, videos, and MP3 audio files .
- ii) Textual unstructured data includes examples like email messages, collaborative software and instant messages, memos, word processor documents, PowerPoint presentations etc. And the different standards for unstructured data are open XML, SMTP, SMS, CSV and Information and content exchange [3].

3. LITERATURE REVIEW

Dr.Goutam Chakra borty [4] et.al in 2014 proposed an view look at how to organize and analyse textual data for extracting useful and informative knowledge from a large collection of document and for using such information to improve business operations and performance.

S.Geetha [5] et.al in 2012 proposed an approach to extract the data from unstructured data by organizing into structured way in the form of Data Relations. Set of rules are used to interpret domain knowledge from the unstructured data.

By segmenting the data using part of speech and its syntactic structure present in the input unstructured text, which will help us to categorize the data into entities, actions and construct the relations among these entities and actions. This approach applied in the “News Retrieval System” which collected news from Various Pages and processed on the basis of Page ranking and displayed on the Single page in an effective way.

Yuanming Huang [6] et.al. in 2010 proposed the theory and methods on massive unstructured audio video intelligent information process in emergency system mechanisms like visual computing, cognitive modelling, cognitive science will be the latest achievement of the method of mathematical analysis for unstructured multimedia data in emergency system. This research results can build mass

information intelligent service platform technical support system framework, breaking the mass of information intelligence services in a number of technical bottleneck.

Dr.Muhammad Shahbaz [7] et.al in 2014 proposed solution in this work is the development of a System (Sentiment Miner). It will provide features to process and distinguish text files for opinion mining at sentence level using Natural language Processing techniques and Opinion Mining algorithms.

4. PROPOSED SYSTEM:

Objectives:

- a) Building a dataset for unstructured data. This dataset can adaptively adjust to the dynamic change of the data.
- b) To propose a model for discovering the useful information and knowledge from unstructured data.
- c) To clustering the data using k-mean and hierarchal data mining algorithm to measure efficiency [2].

In this paper we proposed a new approach for mining of unstructured data, our approach is divided in the seven steps, the flow diagram of the proposed approach is as shown in Figure 1, Text mining is the automatic extraction of implicit, previously unknown, and potentially useful information and patterns, from a large amount of unstructured textual data, such as natural-language texts. In text mining, each document is represented as a vector, whose dimension is approximately the number of distinct keywords in it.

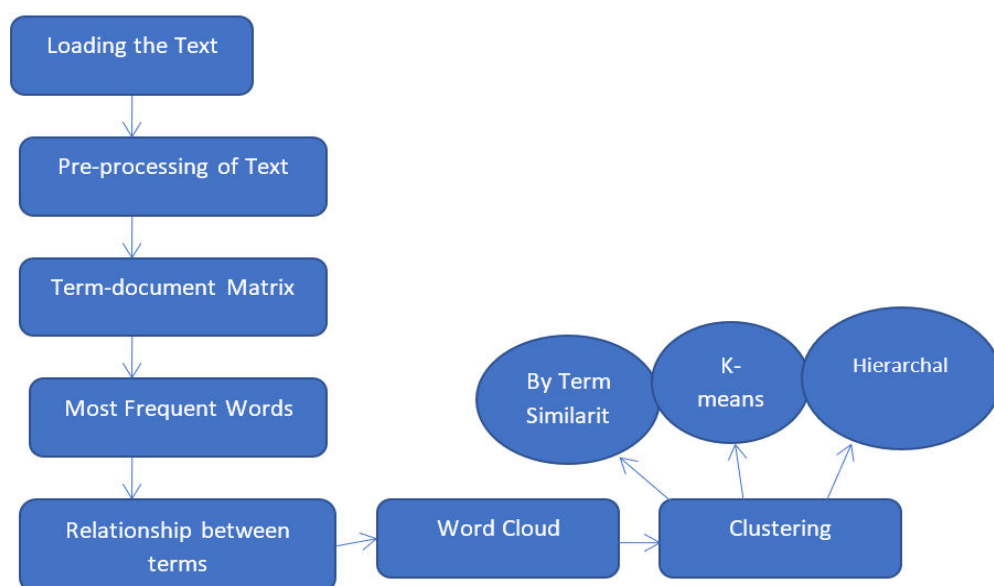


Figure 1 A Flow graph of Text Mining

4.1 Loading the text files

Large document corpus may afford a lot of useful information to people. Corpus consist but it is also a challenge to find out the useful information from huge number of documents. Especially with the explode of knowledge around the world, corporate and organizations demand efficient and effective ways to organize the large document corpus and make later navigating and browsing become more easy, friendly and efficient. In this we have to check files are loaded or not as shown below in Figure 2.

4.2 Pre-processing

Stop-words are words that do not contain any useful information. It includes numbers, special characters, converting the data from uppercase to lowercase, punctuations etc.

4.3. Term-document matrix

A document-term matrix or term-document matrix is a type of matrix that illustrate the frequency of terms that appear in a collection of documents. In a document-term matrix, rows correspond to documents appear in the collection and columns corresponds to terms and tabulate the data according to their frequency and also eliminate the common terms, this will make 10% matrix empty as shown in Figure 3.

4.4 Most Frequent Words

In this we are going to find out the word that are most frequently occurred in the documents and we also adjust the frequency of word like 3,4. Suppose we will adjust the minimum frequency 3 and then it will plot the word frequency graph as shown in Figure 4, which shows the word that occurs 3 or more times in documents.

4.5 Relationship between Terms

If we are thinking about a term that you have found to be particularly relevant to your analysis, then we may find it favourable to classify the words that most highly correlate with that term or not.

4.6. Word cloud

Humans are generally strong at visual analytics. That is part of the reason that these have become so popular. With the help of word cloud we find out the documents from which data they are related. A word cloud is shown in Figure 5 and word of frequent words as shown in Figure 6.

4.7 Clustering

Clustering is the process of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (cluster)[8]

- Clustering by Term Similarity
- Hierarchical Clustering
- K-mean clustering.
- Clustering by Term Similarity
- For making clustering similarity, firstly we have to eliminate a lot of uninteresting or infrequent words. This makes 15% empty space.
- Hierarchical Clustering
- In this we have to cluster the data based on euclidian distance.
- Euclidian distance measure the distance between words by applying the formula. Euclidian distance is the square root of sum of the difference between two words as shown in Figure 7. Cluster can be view as a dendrogram. Does not need the number of clusters as input possible to view clusters at different levels of granularity.
- Then two clusters will combine so that the similarity between the cluster is closest until the number of clusters becomes 1 or similar to the no. of cluster as specified by the user[9].
- Start with n clusters, and a single sample indicates one cluster. Cluster can be visualized using a tree structure (a dendrogram). In this we does not need the number of clusters as input possible to view clusters at different levels of granularity.
- Find the most similar clusters C_i and C_j then merge them into one cluster.
- Repeat step 2 until the number of cluster becomes one or as specified by the user. The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge. There are several ways to calculate the distances between the clusters C_i and C_j .

4.7.1 K-mean clustering[8]

A view of K-mean clustering as shown in Figure 8.

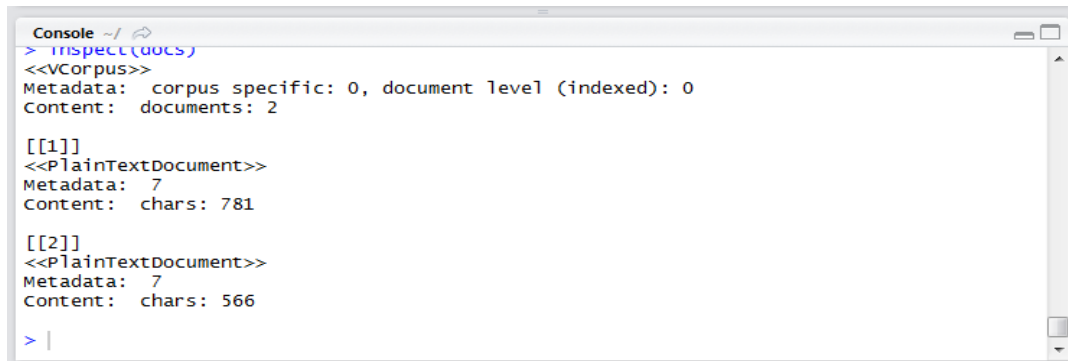
- Initialization, The first step in k-mean is to define the number of clusters and the centroid that is elementary to defined for each cluster.
- Classification, The distance is estimated for each data point from the centroid and the data point having least distance from the centroid of a cluster is assigned to that particular cluster.
- Centroid Recalculation Clusters generated previously, the centroid is again repeatedly calculated means recalculation of the centroid.
- Convergence Condition Some convergence conditions are given as below:
 - Stopping when reaching a given or defined number of iterations.
 - Stopping when there is no exchange of data points between the clusters.
 - Stopping when a threshold value is achieved.
- If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied Agglomerative Hierarchical.

5. IMPLEMENTATION AND RESULT ANALYSIS:

For the Implementation and results analysis of Unstructured data, we have used two datasets file which are analyse using R programming on windows 7 operating system and i3 processor with 1 GB Ram .

Result analysis is shown in the n Figure 2 to Figure 8:

5.1 Loading Files



```
Console ~\ |
> inspect(docs)
<<VCorpus>>
Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 2

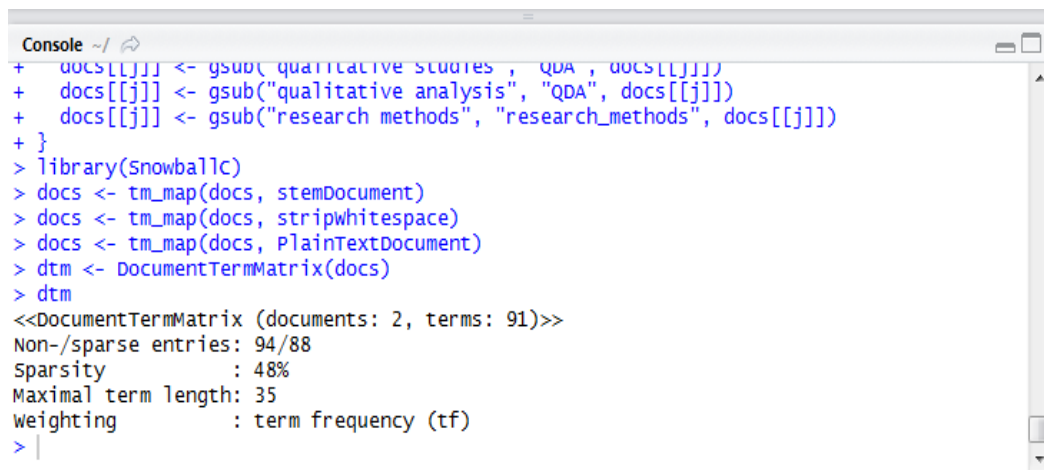
[[1]]
<<PlainTextDocument>>
Metadata: 7
Content: chars: 781

[[2]]
<<PlainTextDocument>>
Metadata: 7
Content: chars: 566

> |
```

Figure 2 A view of loading the files

5.2 Term-document Matrix



```
Console ~\ |
+ docs[[j]] <- gsub("quantitative studies", "QDA", docs[[j]])
+ docs[[j]] <- gsub("qualitative analysis", "QDA", docs[[j]])
+ docs[[j]] <- gsub("research methods", "research_methods", docs[[j]])
+ }
> library(SnowballC)
> docs <- tm_map(docs, stemDocument)
> docs <- tm_map(docs, stripwhitespace)
> docs <- tm_map(docs, PlainTextDocument)
> dtm <- DocumentTermMatrix(docs)
> dtm
<<DocumentTermMatrix (documents: 2, terms: 91)>>
Non-/sparse entries: 94/88
Sparsity           : 48%
Maximal term length: 35
weighting          : term frequency (tf)

> |
```

Figure 3 A view of Term document matrix

5.3 Word Frequency Graph

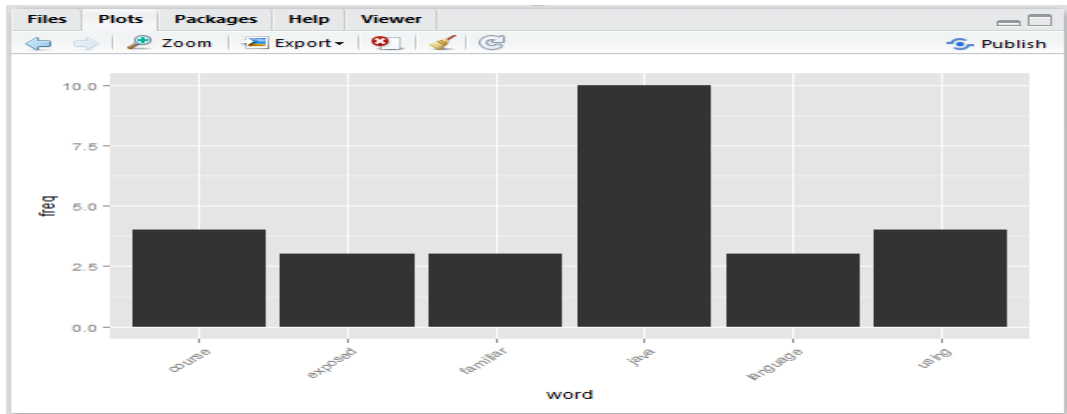


Figure 4 A view of word frequency Graph

5.4. Word Cloud

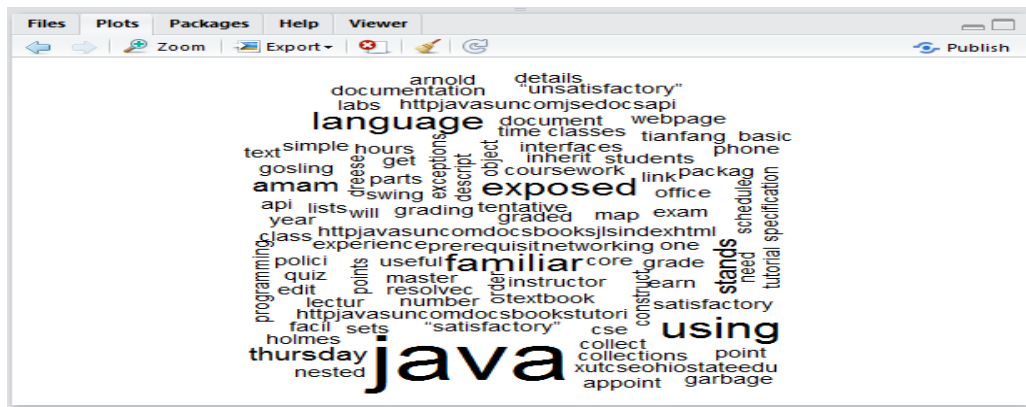


Figure 5 A View of word cloud

5.5. Word Cloud of Frequent Words

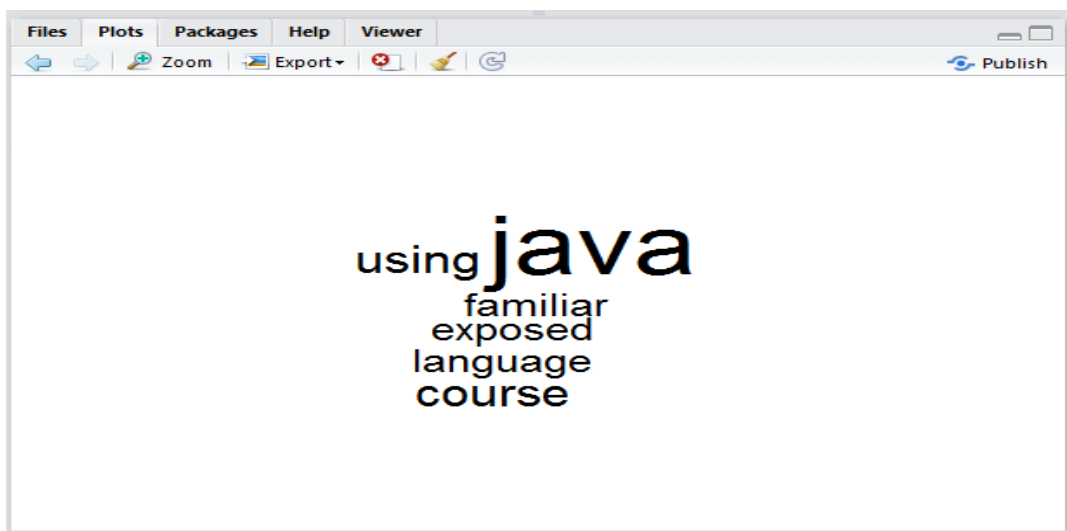


Figure 6 A View of frequent words word cloud

5.6 Dendrogram of Hierarchical Clustering

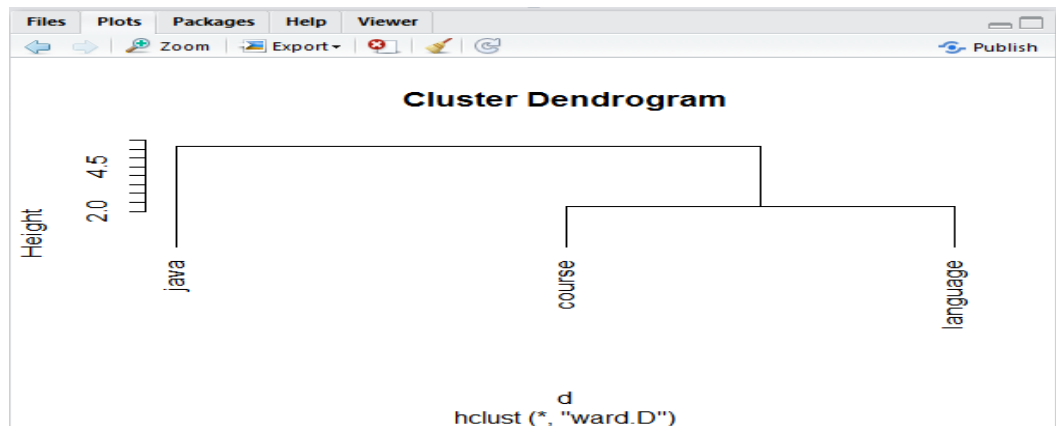


Figure 7 A view of Hierarchical Clustering

5.7 K-mean clustering

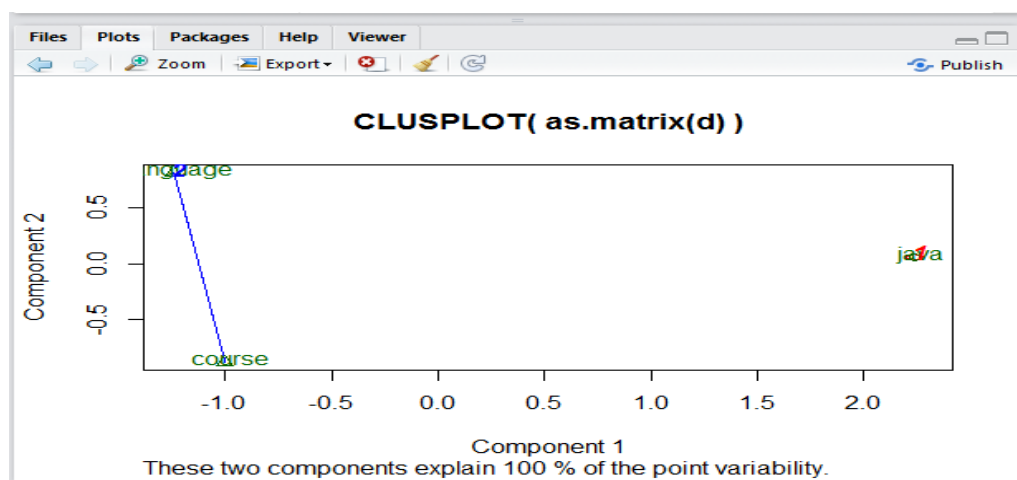


Figure 8 A view of K-mean clustering

6. CONCLUSION

We present a framework for text mining. One of the main aspect of this framework is to identify the documents and the data they contained and evaluate the feasibility to apply text mining which may achieve good performance by using data mining technique when dealing with thousands of documents, by separating the data contained by documents into bag of words. From our experiment we analyze, pre-processing does play an important role. In document-term weighting factor is to take care of the effect of document length and make each document have the same significance.

REFERENCE

- [1] K.V. Kanimozhi, Dr. M. Venkatesan, “Unstructured Data Analysis- A Survey”, International Journal of Advanced Research in Computer and communication engineering, Volume 4, Issue 12, pp. 223-225, December 2015.
- [2] Prakash R.Andhale,S.M Rokade, “A Decisive Mining of Heterogeneous Data”,International Journal of Advanced Research in Computer and Communication Engineering, Volume 4,Issue 12,pp. 436-437,December 2015.
- [3] Luis Flipe Da Cruz Nassif,Equardo Raul Hruschka, “ Document Clustering For Forensic Analysis : An Approach for Improving Computer Inspection”,IEEE Transaction on Information Forensics and Security,Issue 1,pp. 46-54,Jan 8,2013.
- [4] Dr. Goutam Chakra Borty,Murali Krishna Pagolu, Analysis of Unstrucured Data: Application of Text Analytics and Sentiment Mining,2014.
- [5] S.Geetha,Dr. G.S Anandha Mala, “Effectual Extraction of Data Relations from Unstructured Data”, Third International Conference on Sustainable Energy and Intelligent System, Tiruchengode,Tamilnadu,India,pp 1-4,December 2012.
- [6] YuanMing Huang,Yujie Zheng, Research on Theory and Methods on Massive Audio, Video Unstructured Information Intelligent Process in Emergency System, 978-1-4244-6928-4/10/\$26.00©2010 IEEE.
- [7] Dr. Muhammad Shahbaz, Dr. Aziz Guergachi, Rana Tanzeel ur Rehman.Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text. 978-1-4799-3010-9/14/\$31.00 ©2014 IEEE.
- [8] Manpreet Kaur,Usvir Kaur, “Comparison of K-mean and Hierarical Algorithm using query redirection”, International Journal of Advanced Research in Computer Sciene and Software Engineering,Volume 3,Issue 7,pp.1454-1459,july 2013.
- [9] Sung Young Jung, and Taek-Soo Kim, “An Agglomerative Hierarchical Clustering Using Partial Maximum Array and Incremental Similarity Computation Method”, Proceedings of the 2001 IEEE International Conference on Data Mining, pp.265-272, November 29-December 02, 2001.