Implement k-means Clustering Algorithm for Document Data Analysis

Gutha Murali[1] Ms. S Anthony Mariya Kumari[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupati, India

Abstract— At present time enormous measure of helpful information is accessible on the web for access, and this gigantic measure of information is shared data which can be utilized by anybody expected to utilize. The accessibility of different nature of record information has lead to the assignment of grouping in the dataset. Bunching is one of the essential procedures utilized for grouping of the huge dataset and broadly pertinent numerous territories. High caliber and quick archive bunching calculations assume a huge job to effectively explore, outline and arrange the data. Late investigations have demonstrated that partitional grouping calculations are suit-capable for substantial datasets. The k-implies calculation is commonly utilized as partitional grouping calculation since it may be effortlessly actualized and it is most effective as far as execution time. The significant issue with this calculation is affectability for determination of the underlying allotment and its intermingling to nearby optima. In this examination it think about refined helpful data from record informational collection utilizing least crossing tree for archive bunching and great nature of groups have been produced on a few report datasets, and the yield demonstrates acquired shows a viable enhancement in execution.

Key words: Document Clustering, K-Means Algorithm

I. INTRODUCTION

In these scenario data mining is the most important concept in computer science. now most of the researchers focused on Pattern Reorganization, Spatial Data Analysis, Image Processing, Economic Science, Biological Data Analysis, and WWW etc. Document clump is one of the simplest familiar analysis issues within the field of knowledge mining. The aim of document clump is to divide information into subsets and contain it's unrealistic objects. There is many formulas of clump algorithmic rule on the market in literatures. The massive quantity of text documents is major downside as a result of the expansion of text document is increasing day by day. The need of effectively manage or explore the results of search engine queries, inspires the study of document clustering. The concept behind the clustering of the document is to find the hidden similarity and the discovery of good groups.

In document file we have different document data are available and N is set of data points. In the concept we are find the distance between the data. Usually the same properties of data are quantitively evaluated by some optimality measures such as minimum intra cluster length or maximum inter cluster distance, Therefore clustering analysis has become an essential and valuable tool in various fields.

Here we have to calculate the distances between data points based on the tern frequencies. If term frequencies are calculate based on the document data. Here we take the no of document data files. Data files are converted from unstructured data to structured data.

II. RELATIVE STUDY

There are many document clustering techniques based on distance like k- means, k-medoid, DBSCAN., etc. Previous clustering based Minimum Spanning Tree Algorithms works on divide and conquer scheme. Here first step of we convert the unstructured data into structured format, then in second step produce the distance matrix and to finish find the clusters. This paper analyze public networking sites data using K-mean classification algorithm. The experimental outcome of text document classification on social networking sites dataset.

Chang J.et. al. proposed in 2010 "A model of classification based on minimum spanning tree for massive data with Map Reduce implementation". In this paper they gift a classification model with tries to search out Associate in Nursing intermediate model between higher than 2 extremes aiming at taking advantage of their benefits and removing some downside.

D.S Hindustani projected in 2000 "Analysis of Social Networking Sites exploitation K- Mean cluster Algorithm" during this paper they gift build clusters for social networking sites mistreatment k-means clump algorithm.

K. Alsabti projected in 1998" [a]An economical k-means clump Algorithm" projected a high level description of the direct k-means clump rule.

III. PROPOSED SYSTEM

In this section new framework is proposed for document clustering.

1) Remove stop words.
2) Stemming and Term Selection in documents.
3) Generate the clusters based on distances.

A. Removing Stop Words:

First we remove all stop words and special symbols. Stop words are the words which don't have meaning with respect to the classification. So these words are eliminate when the term matrix is created for the categorization purpose. Stop words are 'of, 'it', 'the', 'was', 'were' etc., along with all removed prepositions, conjunction and articles from the document data.

B. Stemming and Term Selection in Document:

After eliminating the stop words, the stemming process will be applied. The stemming process is remove the prefixes and suffixes. The objective is to delete the variation that arises from the amount of verity grammatical forms of the similar word. The stemming process helps to decrease the size of the data dictionary file.

In term selection is a critical task for the classifier performance. With increasing number of documents, the number of features also increases. To reduce the size of the dictionary, the threshold term selection method is used. In this

method, the upper and lower thresholds are decided according to the number of words in the dictionary. After that the term which exceeds the upper threshold and the terms below lower threshold are extracted from the document. This helps to reduce the size of the dictionary.

The coefficient theme TF-IDF (Term Frequency, Inverse Document Frequency) is employed to assign higher weights to differentiate terms during a document, and it's the foremost wide used constant theme that's made public as and it is the foremost wide used constant theme that's made public as Once text pre-processing is applied over the raw document datasets, it's going to be reborn into form of binary matrix.

The essence behind our algorithmic rule, K-means (with the K from the K means that algorithmic law and therefore the means that from k-means since we have a propensity to use an equivalent distance and same functions), is to facilitate the calculation of the initial centroids employing a greedy approach; we have an inclination to believe that the k-means' original algorithmic program desires improvement with the initial random alternative of the centroids array. Hence, our initial step is to calculate all of the present parts that have the very best degree within the space; from there we will have AN initial configuration of what the clusters ought to seem like. On the second run, we have a tendency to eliminate all the centroids that area unit in a very single cluster and choose k clusters with the very best results of the similarity perform to be taken because the real cluster centroids. This being done, we have a tendency to restate on the remainder of the info parts to check if the centroids area unit aiming to amendment.

IV. PROPOSED ALGORITHM

The following algorithm steps are used to generate a clusters for document data.

Steps:

1) Input: Let us take a document dataset D=(D1, D2, ...Dn)
2) Eliminate stop words from document data.
3) After eliminating stop words to apply the stemming process.
4) Term selection process.

Algorithmic steps for k-means clustering

Let us take $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of centers

1) Randomly choose *'c'* cluster centers
2) Find the distance between each data point and cluster centers.
3) Allocate the data point to the cluster center whose distance from the cluster center is minimum of all cluster centers.
4) Recompute the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

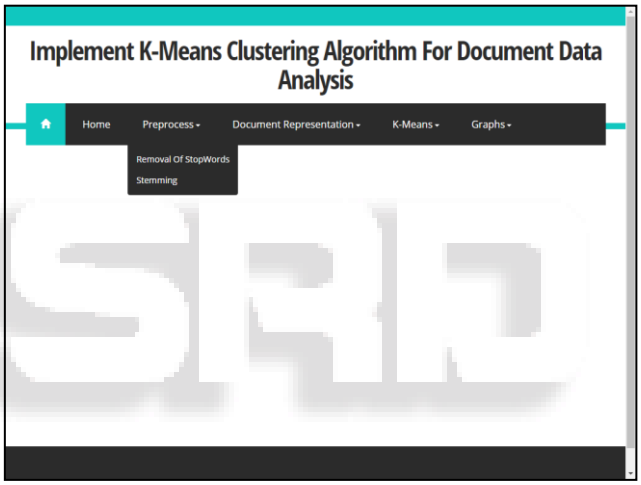
Where, *'c_i'* indicates the number of data points in i^{th} cluster.

5) Recompute the distance between each data point and new obtained cluster centers.
6) If no data point was reallocate then stop, otherwise repeat from step 3.

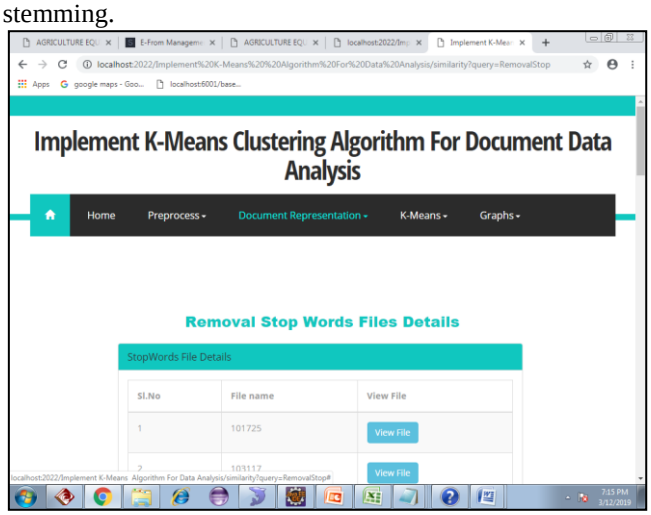
k-means is one all told the simplest unattended learning algorithms that solve the well proverbial bunch balk. The procedure follows a lenient and simple because of classify a given data set through a precise vary of clusters (assume k clusters) mounted apriority. The foremost arrange is to stipulate k centers, one for each cluster. These centers got to be placed throughout a cunning manner due to all totally different location causes different result. So, the upper various is to position them the most quantity as possible distant from each other. Succeeding step is to need each reason happiness to given information set and associate it to the nearest center. Once no purpose is incomplete, the first step is completed Associate early cluster age is completed.. A loop has been generated. As a results of this loop we have a tendency to tend to may notice that the k centers modification their location step by step until no tons of changes unit of measurement done or in numerous words centers do not move any longer.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

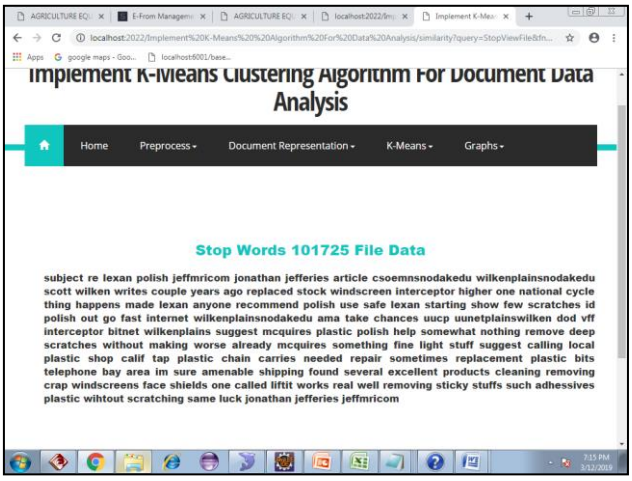
V. RESULT



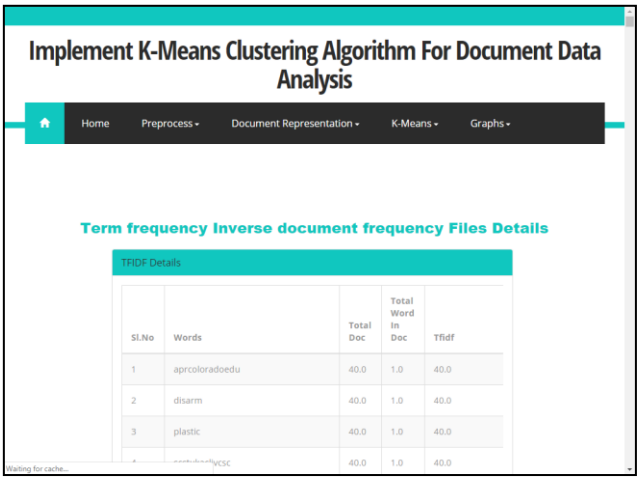
Selecting the pre-process for removal of stop words and stemming.



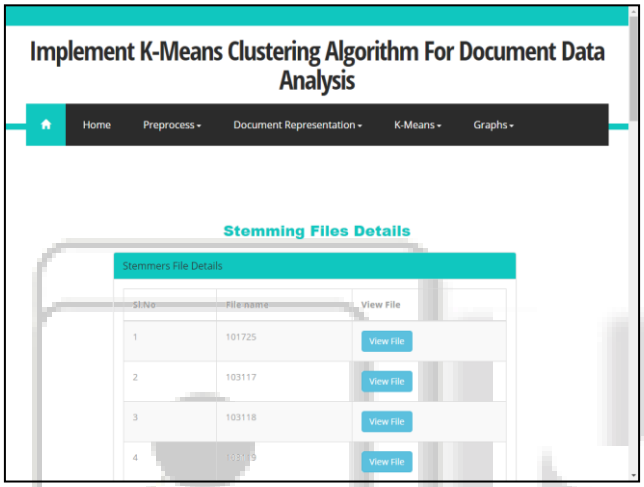
Here we are selecting the stop word files from different document files.



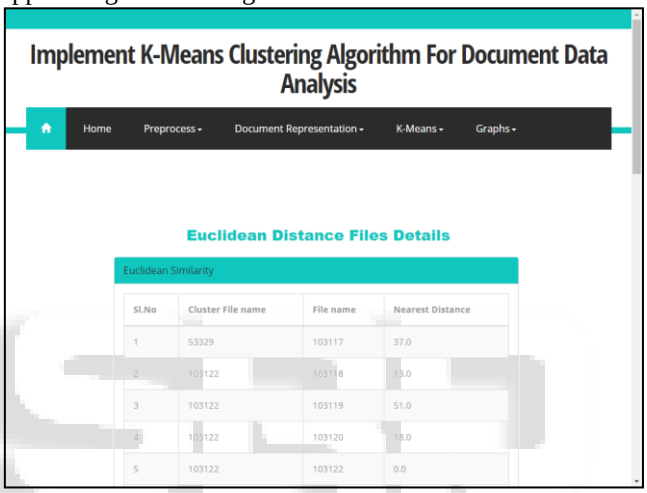
Here we are removing the meaning less data from selected files.



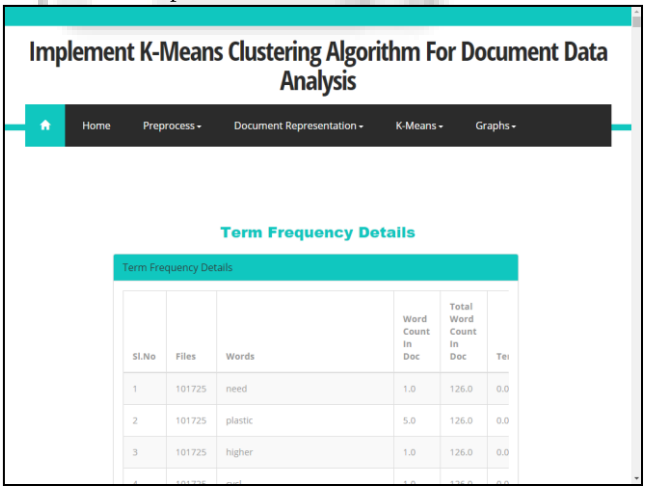
The weighting scheme TF-IDF is used to assign upper weights to distinguish terms in a document.



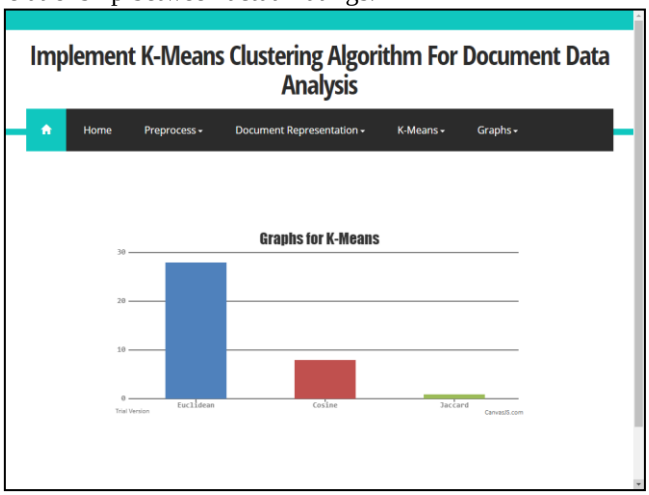
Here we are applying stemming process that is used to eliminate the prefixes and suffixes.



Euclidean distance is reportedly faster than most other means of determining correlation and it compares the relationship between actual ratings.



The term which exceeds the upper threshold and the terms below lower threshold are extracted from the document.



After applying the k-means algorithm on the document data files and finally display the graph.

VI. Conclusion and Future Scope

In this paper, we have to find the clusters for document data. The document data must contain in the form of unstructured format. In first step of the frame work convert the unstructured data into structured format, then in second step

produce the distance matrix and finally find the clusters. In the future we will explore and test our proposed document clustering algorithm in various domains and also reduce the complexity.

REFERENCES

[1] Chang, J., Luo, J., Huang, J.Z., Feng, S., Fan, J.: Minimum spanning tree based classification model for massive data with mapreduce implementation. In: Fan, W., Hsu, W., Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) ICDM Wosrkshops,. IEEE Computer Society pp. 129–137, 2010.

[2] Sakshi Saxena, Priyanka Verma, Dharmveer Singh Rajpoot " Clustering Based Minimum Spanning Tree Algorithm" Computer Science and Engineering Jayapee Institute of Information Technology Noida, India.

[3] Andreas C. Muller, S. Nowozin, christoph H. Lampert, "Information theoretic clustering using International Journal of Computer Sciences and Engineering Vol.-1(1), pp. (6-13) Sept. 2013 © 2013, IJCSE All Rights Reserved 13 minimum spanning tree"Pattern Recognition, pp. 205-215, 2012.

[4] D. S. Rajput, R. S. Thakur, G. S. Thakur ,NeerajSahu, " Analysis of Social Networking Sites Using K- Mean Clustering Algorithm", International Journal of Computer & Communication Technology (IJCCT) ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 – 7449 Vol-3, Iss-3, pp. 88-92, 2012.

[5] Han I and Kamber M, "Data Mining concepts and Techniques," M. K. Publishers, pp.335–389, 2000.

[6] D.S Rajput*, R.S Thankur, G.S. Thakur "Clustering Approach Based on Efficient Coverage With Minimum Weight for Document Data" Department of Computer Application, MANIT, Bhopal(MP).

[7] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998.

