International Journal of Science Engineering and Advance Technology (IJSEAT)
ISSN 2321-6905, Vol. 8, Issue 11, November -2020

# Document Clustering to organize the similar documents into classes using K-Means to improve retrieval and Time complexity

Roopa Devi Chandanala[1], N.Harini[2]
[1]M.Tech Scholar(CSE) in Department of Computer Science and Engineering,
[2]Assistant Professor, Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vizianagaram (Dist), Andhra Pradesh, India.

**Abstract-**

Document clustering has been one of the quickest developing exploration field for as far back as couple of many years. It has become a significant errand in content mining on account of the gigantic expansion in records on the web. All the associations require the best possible administration of printed information. Record grouping is the unaided procedure that assists with getting sorted out the comparable archives into classes to improve recovery. The paper clarifies the periods of record bunching and the improvement in report grouping utilizing quality weighted k-intends to group the archives and to place the comparable reports in the best possible group. Test results shows that precision of proposed strategy is high contrast with the essential k-implies as far as F-Measure and time complexity. Grouping centers to coordinate an assortment of information things into bunches, with the end goal that things inside a group are more "comparative" to one another than they are to things in different groups. The k-means strategy is one of the most broadly utilized grouping procedures for different applications.

## I. Introduction

Today as we know in every industry almost all document paper has their electronic storage copies. As compare to manual storage this method is safer and occupies much smaller space and also this electronic file has very easy and quick access. But because of increase in number of electronic document, it is very hard to handle that it is hard to organize, analyze and manage the electronic document efficiently by putting some manual effort. So this comes to a challenge for efficient and effective organization of a text in document automatically [1]. And therefore this increases the demand of the tool that can be used for analyze and discover useful information from document. Solution for this problem is to use data mining technique and usage of data mining technique on text document is called as text mining or text data mining which already increase the interest of many researchers in the field of research [2]. Text mining is roughly equivalent to text analysis where it derives high quality of information from the text. This text mining actually involves analysis of text such as information retrieval, lexical analysis to study word frequency distributions, pattern recognition and information extraction. The main aim of this technique is to turn text in useful data. Text mining technique is used in various application such as security application, biomedical application, record management, software application, online media application, marketing application, sentiment analysis, academic application etc. All these application are based on one common task that is extracting high quality information from the text document. Document clustering is field of data mining which automatically arranged useful data into group where data in category are similar to each other and dissimilar to other category of document [2]. Clustering can be known as one of the most important unsupervised learning problem. In general, the clustering is defined as the process of organizing objects into groups where its members are similar in some way". Therefore, cluster is a collection of objects which are similar internally, but clearly dissimilar to the objects belonging to other clusters. In simple cluster is refer as group of similar kind of objects and it very useful for organizing document to improve the browsing retrivation and support. Clustering is also defined as a process of partitioning a set of data or objects into a set of meaningful subclasses which is called as a clusters. Clustering is very helpful to user in understand the natural grouping or structure in a data set and it is used as either standalone tool to get better insight into data distribution or in pre-processing step of other algorithm. Good clustering will produce high quality result where intra class similarity is high and inter class similarity is low. The quality of a clustering result is depending on both the

similarity measure used by the method and its implementation and also it is measured by its ability to discover some or all of the hidden patterns. Clustering has wide application in pattern recognition, image processing, document classification, economic science, marketing, insurance, land use and many more [3].

## II. Related Work

K-means clustering comes under partitioning clustering algorithm. It partitions given data into K clusters. Several other clustering algorithms are proposed for dealing with document clustering. Novel algorithm [2] for automatic clustering suggested how clustering is done automatically, improved partitioning K-means algorithm[3] presented newmethod for initializing centroids. Ontology based kmeans algorithm [4] presented how ontological domains are used in clustering documents. Improved document clustering algorithm using K-means [5] presented solution of over clustering by partitioning of documents using divide and conquer approach. Above discussed methods [2],[3],[4],[5] used 20newsgroup, Rueters-21578 and Real time data sets. All Algorithms used cosine similarity measure for finding similarity. Initial kvalue is specified in every algorithm except first and last one because first and last algorithm is automatic. They considered documents categories for automatic clustering. Zero clustering is the major advantage of first algorithm. i.e. documents with zero value in similarity matrix also get cluster. Remove over clustering is the major advantage of fourth algorithm. All Algorithms are adaptable to dynamic data except second, because second algorithm calculates average document similarity matrix. After analysis we come to know that these algorithms have limitations, i.e. random center generation, not consider the semantic analysis of documents. The improved k means algorithm has a major limitation it takes nonexclusive words also and do not match the words by semantic basis. To overcome these limitations a new algorithm is developed.

*Sorting the Document utilizing Multi Class Classification in Data Mining*

Finding the examples and anomalies is one of the significant issues in the field of data mining. Particularly in the field of human services examination has turned out to be hard to anticipate the examples and basic leadership. Classification strategies are utilized to recognize the exchange name. The classification procedures are utilized to gather the examples in the learning stage and recognize the anomalies in preparing stage. In social insurance investigation, just classifications are restricted with two class levels as positive and negatives. The side effects of patients are gathered and classified into patterns at that point by utilizing the examples; they identify the seriousness level of infections. The proposed framework for the most part concentrates on distinguishing the seriousness level of patients by upgrading the limit classifications. This thought can be accomplished by basic chunks which is a record or ascribe used to characterize classification where that characteristic considered as the choosing expert. The classification precision can be enhanced with basic pieces and improving to help multi class (low, medium, high and typical) and different trait situations. The basic pieces distinguishing proof and classification conspire is enhanced to help different classes. The framework can be embraced to deal with blended characteristic data esteems. The limit estimate algorithm is upgraded to decrease the location multifaceted nature. Post preparing operations are tuned to recognize classes for different classification data condition. [3]

Classification procedures are utilized to recognize the exchange name. Basic chunks are utilized to speak to the space learning of the data gathering. Classification precision is enhanced with basic chunks and class limit algorithm. The framework is upgraded to help different classes and multi characteristic condition. Classification precision is enhanced by the chunks based classification plot. The framework diminishes the Computational multifaceted nature. The framework supports blended property data for classification process. [4]

## III. Document Representation

A text document is typically exhibited as a vector of term weights (highlights) from an arrangement of terms (word reference). Each of these terms happens in any event once in a specific least number of documents. A large portion of the text classification specialists accept in their examinations the Bag-of-Words portrayal display (a vector space demonstrate). It accept the document's structure not vital, while the text (a solitary expression or an entire document) is

depicted as an unordered arrangement of articulations. The request of words in an expression or linguistic use are likewise insignificant. Highlight vectors are articulations seen in a given document. The rundown of words (word-list) W = (w1, ..., wd) in a preparation set comprises of every single unmistakable word (likewise called terms), which can be found in the preparation cases after rejection of stopwords. They are the words not bearing any basic information, similar to a few, and, additionally or uncommon words (seeming just once in the specimen). For the document D, its component vector (term) is depicted as T = (t1, ..., td), coming about because of W. The estimation of every component of T can be double (esteem 1 implies that the given word is available in the case), or whole number, showing the quantity of the word appearances in a document. As document includes entire expressions can likewise be considered.

Algorithms having a place with this field emphatically depend on both preparing and testing data. A preparation set is an arrangement of named documents communicating the theory. It is utilized to recover information about specific classes of documents. The test-ing set is utilized to check the nature of the algorithm. The classification is mapping objects into the limited arrangement of whole number numbers (classes). The learning procedure comprises in discovering traits in cases that permit the recognizing object of discrete classes. The significant issue is overfitting, that is the extreme change of the algorithm to the preparation set. The algorithm influenced by overfitting has unacceptable prescient execution, since it concentrates on immaterial points of interest in data. It is critical to choose sets dispassionately, which can be accomplished by the cross-approval.

There are two sort of mistakes happening in machine taking in: the example blunder (inspecting, estimation blunder) and the genuine mistake (outright, worldwide). The specimen blunder happens while watching an example (subjectively chose set of documents) of the entire populace. The genuine mistake is a likelihood of the inaccurate case classification arbitrarily chose from the populace with a specific likelihood conveyance. The genuine blunder is esti-mated utilizing the example mistake on different specimens.

Dimensionality lessening is a vital advance during the classification procedure. It permits precluding irrelevant highlights of a document, which frequently lessen the classification effectiveness, diminishing their speed and precision. Furthermore, DR decreases overfitting. The dimensionality lessening techniques are partitioned into highlight extraction and include determination strategies.

Data preprocessing is utilized to clean the text from the dialect subordinate factors and comprises in tokenization, stop-words expulsion or stemming. Highlight extraction is the initial phase in data preparing, changing a text document into less complex shape. Documents in text classification contain a lot of highlights, while a large portion of them are immaterial or commotion. Dimensionality lessening is a strategy for excluding in a factual procedure a lot of watchwords with a specific end goal to make a relate process a large amount of key words in order to create a relatively short vector. The process of DR consists of the following steps:

*Tokenization*

A document is dealt with as a chain of tokens (marks), which is isolated into sets of tokens.

*Stop-words expulsion*

The words like and, additionally, now and again are utilized to compose text pretty regularly, so they can be basically evacuated in classification process.

*Stemming*

The use of stemming algorithm, which changes over other word frame into a comparative standard shape. This progression is a procedure of consolidating tokens to their unique frame, such as relegating to allot, checking to tally, et cetera (see 1).

After FE, the following stage in preprocessing is to make the

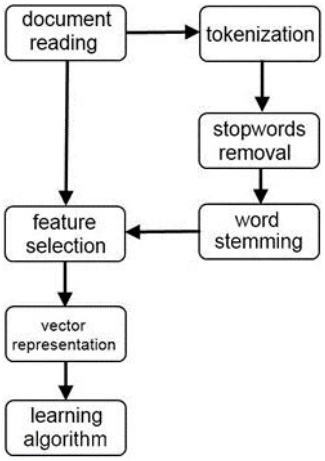



Fig. 1. Document classification process.

Vector space in light of the component determination (FS), to modify the accuracy and viability of the document classification. By and large a decent technique for include choice considers the area and algorithm attributes. In document classification FS is utilized to decrease the measure of highlight space dimensions and alter the adequacy of classification.

**IV. Documents Classification**

The documents can be grouped by three ways unsupervised, managed and semi regulated classification Many methods and algorithms are proposed as of late for the bunching and classification of electronic documents, our concentration in these determination will be on the administered classification systems, new improvement and some future research heading from the current writing. The automatic classification of documents into predefined classifications has seen as a dynamic consideration as the Internet use rate has immediately amplified. From most recent couple of years , the errand of automatic text arrangement have been broad investigation and quick advance appears around there, including the Regularly managed learning strategies are utilized for automatic text classification, where pre-characterized class names are doled out to documents in view of the probability recommended by a preparation set of named documents.

Rocchio's Algorithm assemble model vector for each class i.e. the normal vector over all preparation document vectors that have a place with class Ci ascertain likeness between test document and each of model vectors.

Dole out test document to the class with most extreme closeness this algorithm is anything but difficult to actualize, quick student and have pertinence input system yet low classification precision, straight blend is excessively basic for classification and steady are exact. This is a generally utilized pertinence criticism algorithm that works in the vector space demonstrate. The specialists have utilized a variety of Rocchio's algorithm in a machine learning context, i.e., for taking in a client profile from unstructured text. The objective in these applications is to automatically prompt a text classifier that can recognize classes of documents

The k-Nearest neighbor algorithm kNN is utilized to test the level of likeness amongst documents and k preparing data and to store a specific measure of classification data, in this way determining the class of test documents. In paired classification issues, it is useful to pick k to be an odd number as this stays away from tied votes and figure similitude between test document and each neighbor appoint test document to the class which contains the vast majority of the neighbors

This technique is powerful, non-parametric and simple to actualize, as contrast with Racchio algorithm more nearby normal for documents are considered, however the classification time is long and hard to discover ideal estimation of k. While presents the utilization of expressions as fundamental highlights in the email classification issue. They performed broad exact assessment utilizing our huge email accumulations utilizing two k-NN classifiers with TF-IDF weighting and similarity individually.

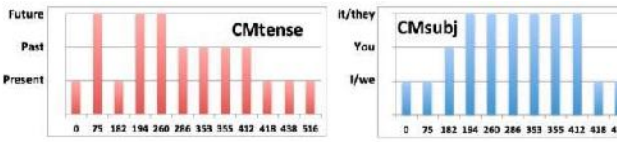
**V. Proposed System**

**SEGMENTATION OF POSTS**

For an archive d, there are 2jdj 1 potential divisions. Among them, we are keen on the one that is all the more precisely lined up with the various expectations of the content. Finding the correct division is a difficult assignment for which there is as of now a huge assortment of work, from division of questions to segmentation of documents, In these examinations, a decent division is one where each portion is (I) cognizant and (ii) to a great extent separated from its adjoining sections. Since our basis for division is the expectation based, these two properties mean a division where each section: (I) passes on a solitary

clear goal; and (ii) this aim is exceptionally not quite the same as those passed on by the nearby fragments. Comparably, the above models call for division with profound fringes are three difficulties: distinguish the highlights to use for recognizing the goals, measure the intelligibility inside a fragment close by the profundity of the outskirts of a competitor division, and, select the best segmentation among the candidates.

.




| | Possible Segmentations |
|---|---|
| Boxes | 75, 182, 201, 259, 285, 338, 355, 371, 418, 436, 488, 535 |
| (a) CMtense-Based | ([0,75],[76,182],[183,201],[202-285],[286-418],[419-535]) |
| (b) CMsubj-Based | ([0,182],[183,201],[202,418],[419,488],[489,535]) |
| (c) CMqneg Shift | ([0,182],[183,201],[202,438],[439,535]) |
| (d) Intention-Based | ([0-182],[183,418],[419-535]) |
| (e) Thematic | ([0-49],[50-535]) |

**Fig. . CMs and Segmentations**

**Coherence and Depth Computation**

Intelligibility and Depth Computation

Instinctively, as implied prior, to assess the nature of a division we need to gauge what variety is seen inside a section as far as the client goals and how the expectations of a portion vary from those of the contiguous fragments (which would legitimize why the nearby bits of text have been set in various fragments). In this way, given a bunch of highlights, we should have the option to quantify the soundness of a portion and the profundity of a fringe.

To gauge the "profundity" of a fringe, one can abuse the idea of cognizance. An outskirt is "profound" if the CMs in the two fragments it isolates are fundamentally unique. To quantify this distinction, we eliminate the outskirt, which by and by would imply that the sections to its left side and right would turn into a solitary enormous portion, and we measure the lucidness of this fragment. That huge "speculative" section will have either a lower rationality than the two people (showing a profound outskirt) or a higher lucidness, demonstrating a shallow fringe. Along these lines, the profundity of a fringe bi between fragments si and si+1 is: the place where the portion s is the section coming about because of the connection of si and si+1.

In past work, the distance measurements of cosine dissimi-larity, Euclidean distance, and Manhattan distance on term-based portrayals, have been utilized to choose whether two fragments ought to stay isolated or should be better converged as one. Nonetheless, in the test segment, we delineate that term-based portrayals and distance met-rics are not exceptionally compelling for expectation based division.

**Border Selection**

To locate the best division we need to choose the best fringe positions in the archive. With the capacity to mea-sure rationality of a portion and the profundity of an outskirt, we can characterize a measure to decide how solid or feeble a fringe position is. A potential outskirt bi in position I is a decent decision if every one of the two fragments si and si+1 that bi isolates has a solid rationality and bi has high profundity. In light of this, we dole out a score to a potential fringe position. The score can be processed utilizing a weighted amount of intelligence and profundity, the f-insights [19], or some other measurement as long as it is steady with the above standard. We are really processing it as the normal of the three boundaries, i.e.,

There are two expansive ways to deal with recognize the fringes that characterize an expectation based division in an archive. One is a top-down methodology that at first considers the entire record as one section and checks for potential positions a fringe can be submitted in request to part the fragment into two. The position is chosen so the subsequent two sections have a normal score that is superior to the score of the

fringes before the split. The methodology recursively parts portions insofar as such fringes can be found. Its principle restriction is that the examination of the profundity and cognizance in sections that vary essentially regarding length may deceive the calculation. For comparable reasons, com-paring two long sections may prompt mistaken choices.

The other methodology is base up. It at first considers each text unit as a section and iteratively combines consecu-tive fragments to shape longer portions. The converging of two continuous portions is performed by essentially eliminating the fringe that isolates them. We propose various procedures to execute the base up methodology. Every technique utilizes some various models for concluding if to blend seg-ments.

**SEGMENT GROUPING**

The subsequent stage in expectation based post coordinating is to rec-ognize portions that are planned for a similar objective (or reason). We really need to make gatherings with the end goal that seg-ments with comparable expectations end up in similar gathering and sections with various aims in various gatherings. Since the real goal isn't known yet we have displayed it through a vector of highlights, a characteristic decision for making the ideal gatherings is to perform bunching on the component vectors comparing to the expectations of the portions. Each group would then be able to be viewed as an agent of some correspondence objective. We use I to mean a bunch, and C to indicate the arrangement of the created groups.

We have discovered that utilizing the component vector with no guarantees (mean-ing with the supreme estimations of the highlights) isn't extremely effec-tive. All things considered, we need to catch the overall commitment of each component, in this way we have made a vector of loads that depend on the element esteems. We signify this vector with the letter F . We consider two kinds of loads that catch the strength of the utilization of every CM clear cut worth, i.e., of each component. The main kind estimates the strength of the utilization of every CM esteem inside the portion, i.e., in contrast with the recurrence of the other downright estimations of a similar correspondence mean showing up in the section. Utilizing the thought of the circulation table DSbCMr of a correspondence mean CMr presented in Section we characterize the vector Fs of loads, one load for each component.

.

| | | |
|---|---|---|
| Doc A, Seg 1: | . .IhaveanHPsystem | . from every disk |
| Doc B, Seg 1: | My boss gave me . . | . Linux pre-installed |
| Doc A, Seg 2: | Do you know whether | . . . have 1TB disks |
| Doc B, Seg 2: | I am thinking to . . | the entire system? |
| Doc A, Seg 3: | I am asking because . . . | the right one |
| Doc B, Seg 3: | I have already looked . | . . related to it. |

Fig Segments of discussion posts An and B. Fragments found to have a place with a similar goal group show up together

**Segmentation Refinement:**

It is conceivable that more than one fragment from a similar record end up in a similar bunch, in the event that they have a similar expectation however are not consec section variety and outskirt profundity. We make one more ignore the groups and if such cases are discovered, all the portions that have a place with a similar archive in a bunch are linked into one. At the end of the day, accepting the bunching C of the sections of an assortment of reports D, for each group I2C, another arrangement of fragments is considered rather that is developed as: fsj 9d2D: Ss0 2I ^ s0 2Sd, s0gwhere the image [ on portions shows connection. Because of this progression, each report may have all things considered one fragment in each bunch.

**MATCHING**

To play out the record coordinating, i.e., to distinguish the archives in an assortment that are identified with a reference report dq, one path is to see the report dq as an inquiry and afterward measure the relatedness of one another record d0 to that question in a manner like

how IR procedures work. As of now referenced, our position is that such an assignment ought not think about each record all in all however should be specific on every aim separately, and afterward consolidate the outcomes.

Coordinating concerning a particular Intention: . Each group is the projection of each report on the particular expectation that the bunch speaks to. Along these lines, to quantify the relatedness of an archive d0 to the reference report dq as for a particular aim I, it is sufficient to gauge the relatedness For figuring this relatedness any content correlation, for example summarizing language models or IR strategies might be utilized One of the most popular IR methods is the TF/IDF. The center of the first TF/IDF technique and its probabilistic change BM25 comprises of a term weighting plan that gauges a term in a record thinking about the quantity of its appearances in relationship to the quantity of its appearances in the wide range of various archives. We devise a form that is somewhere close to the first and the BM25, and mulls over expectations. Specifically we start with a change of TF/IDF that approaches BM25 and has been actualized in MySQL 5.5.3 for full content looking through That difference processes the heaviness of a term t in an archive d0 as The relatedness of a document $d^0$ to a reference docu-ment $d_q$ with respect to an intention I, can now be computed based on the term weights. If $s_q$ and $s^0$ are the segments of the documents $d_q$ and $d^0$, respectively, in the intention cluster. where $f_{sq}(t)$ denotes the frequency of the term t in the segment sq, jIj the cardinality of the intention cluster, and $jI^tj$ the number of segments in the intention cluster.

**Algorithm 1** Single Intention Matching

**Input**: Cluster I, Doc. Collection D, Document $d_q$2D
**Output**: List of n documents and their intention matching

$M_I$ ;
**for each** $s_q$2$S^{dq}$
**if** $s_q$62I**continue**; // See footnote[1] scr
0
**for each** $s^0$2I
$d^0$.fd j $s^0$2$S^d$g // See footnote[2] **for**
**each** t2$s_q$
scr scr+$f_{sq}$ (t) w(t; $s^0$) log(jIj j $I^t$i)=j$I^t$j $M_I$
M[$hd^0$; scri
Return f$hd^0$; scrij $hd^0$; scri2$M_I$ ^ scr2 top-n scores

**Matching with respect to All the Intentions**.

The top-n records created across the various expectations, i.e., the set M referenced above, are utilized to produce the k most related reports to the reference archive dq. Another rundown R is made that contains each archive that shows up at any rate in one of the rundowns in M. A score is related to each such record that is the amount of the scores with which this report shows up in the different records in M. The k components in R with the most elevated score are returned as answer to the solicitation of the coordinating reports to the reference archive dq. These means are demonstrated in Algorithm 2.

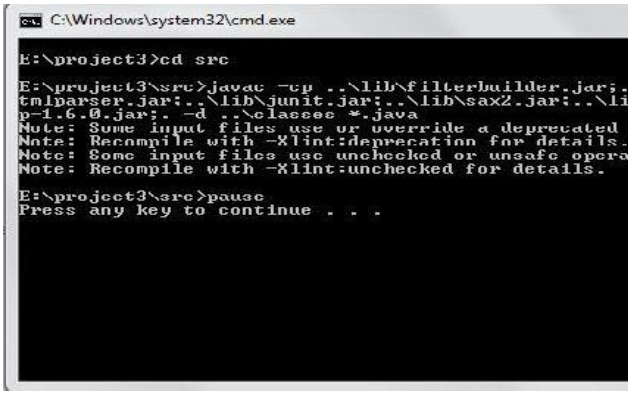
Note that a generally little incentive for n (contrasted with the estimation of k) will support archives that have high score in one rundown in M regardless of whether they don't show up in others, punishing simultaneously records that may show up in numerous rundowns however with lower scores. A moderately high incentive for n contrasted with the estimation of k, then again, will support archives that show up in numerous rundowns even with not exceptionally high scores. We have exactly discovered that a decent decision is a n equivalent.

Indexing. In contrast to segmentation and segment group-ing that are performed offline (preprocessing steps of the document collection), document matching, i.e., the retrieval of the top-k documents for a document query $d_q$, can be performed online due to its low response time (less than 3 millisecs for a collection with more than 1.5M posts, ref. Sec. In practice, in order for Algorithms 1 and 2 to be able to generate fast the (initial) top lists in each cluster I and subsequently generate the final list, we built a full-text index on the terms of all the segments of each segment group (cluster) I. Therefore, we are building jCj fulltext indices. In addition, we are building an index on the ids of the documents where the segments belong so as to be able to access faster the segments of a document query $d_q$. Fig. 6 graphically illustrates the two clusters ($I_0$, $I_1$) and the corresponding indices ($I_0$ indx, $I_1$ index) that have been formed after the segmentation and segment grouping of a small document collection ($d_1$, $d_2$).

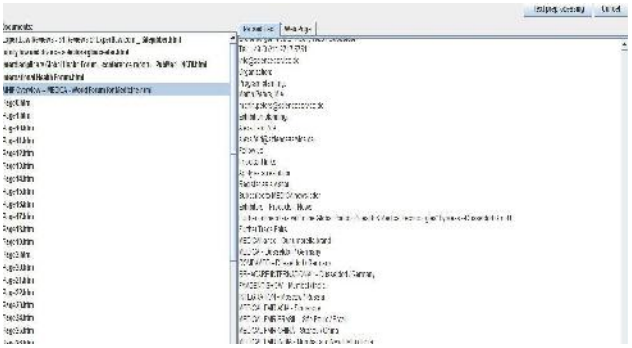
**Algorithm 2** All Intentions Matching

**Input:** Document Collection D, Document $d_q$ 2D, Int
Intention Clusters C
**Output:** List of documents
L ; M ;
**for each** l2C
**for each** $s_q$ 2 $S^{dq}$
**if** $s_q$ 62l **continue**
$M_l$ SingleIntentionMatching(l,D,$d_q$,n)
L L[f$M_l$ g
**for each** $M_l$ 2L
**for each** $hd^U$; scri2$M_l$
**if exists** $hd^0$; xi2M, with x2R
M M[$hd^0$; scri
**else** $hd^U$; xi $hd^0$; x + scri
Return $fd^0$ j $hd^0$; scri2M ^ scr2 top-k scores in Mg

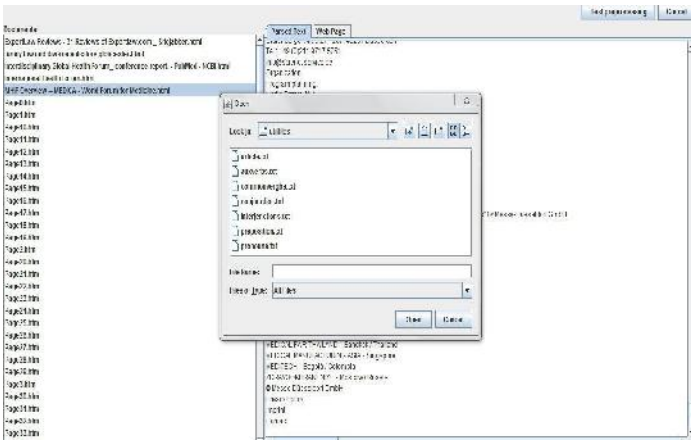
## VI. Results and Discussions



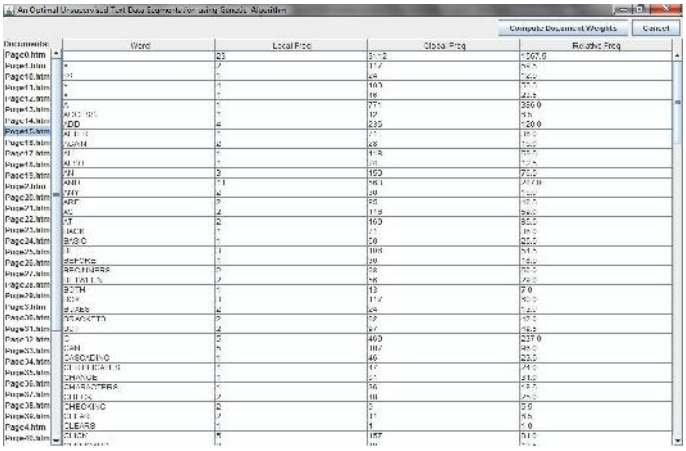

In the below page left part consists of the documents and the right part consists of parsed test and web page.



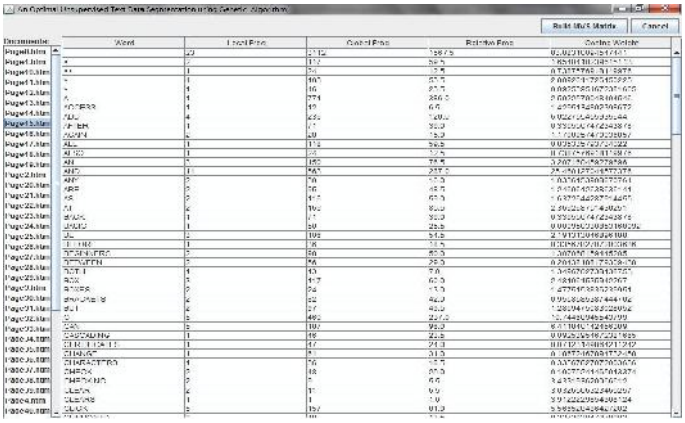
Now we have to perform pre processing to avoid gramatical words.



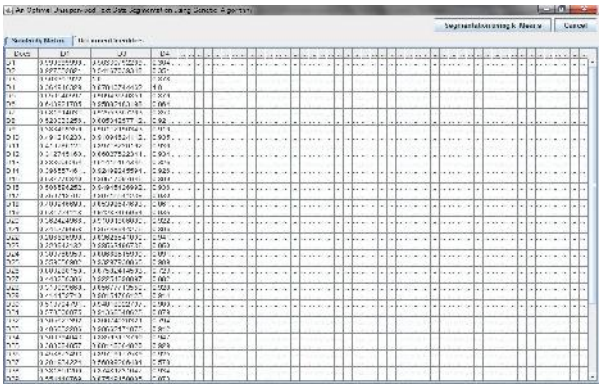
After completion of the pre processing if we click on open button the below will be appeared.



Here we can calculate the document weights.



After clicking on the button Built VSM matrix the below screen will appear.



## VII. CONCLUSION

This project presented an approach to calculate the semantic similarity between two words, sentences or paragraphs. We proposed a novel approach for matching a reference post to the k most related posts in a collection. Our method identifies and exploits post segments that convey similar author intentions. We presented several experiments regarding the right segmentation criteria, the effectiveness of the segmentation algorithms and the formation of intention clusters that prove that a rather intuitive concept, that of the author intentions to communicate a certain message, can be effectively captured by an automated process.

Moreover, due to the nature of the posts, measuring the relatedness score after having distinguished the different segments/messages that the authors intend to communicate has been proved more effective than the direct comparison of the whole posts. Specifically, our approach, according to an evaluation by real users and in comparison with direct full text comparison, increased mean precision by 10%, 12% and 10.1% considering posts in a product support, a travel, and a programming forum.

This venture introduced a way to deal with ascertain the semantic closeness between two words, sentences or sections. We proposed a novel methodology for coordinating a reference post to the k most related posts in an assortment. Our technique distinguishes and abuses post sections that pass on comparative creator aims. We introduced a few analyses with respect to the correct division measures, the adequacy of the division calculations and the development of goal bunches that demonstrate that a somewhat instinctive idea, that of the creator aims to impart a specific message, can be successfully caught by a robotized cycle.

Also, because of the idea of the posts, estimating the relatedness score in the wake of having recognized the various sections/messages that the creators mean to convey has been demonstrated more viable than the immediate examination of the entire posts. In particular, our methodology, as per an assessment by genuine clients and in examination with direct full content correlation, expanded mean exactness by 10%, 12% and 10.1% considering posts in an item uphold, a movement, and a programming forum.

## REFERENCES

**1.** M. Chen, X. Jin, and D. Shen, “Short text classification improved by learning multi-granularity topics,” in IJCAI, 2011, pp. 1776–1781.

**2.** J. Jeon, W. B. Croft, and J. H. Lee, “Finding semantically similar questions based on their answers,” in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR ’05. New York, NY, USA: ACM, 2005, pp. 617–618

**3.** S. Robertson, S.Walker, and M. Hancock-Beaulieu, “Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track,” TREC ’98, pp. 199–210, 1998

**4.** Jiawei Han, Micheline Kamber, Data Mining and Concepts.

**5.** Maqgret H. Dunham, Data Mining and Introductory to Advanced Topics.

**6.** Cluster Analysis:http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

**7.** **7**..Data Mining: http://www.webdocs.cs.ualberta.ca/~zaiane/courses/comput690

ROOPA DEVI CHANDANALA Holds a MCA Post Graduation Degree from Avanthi Institute of Engineering & Technology. She is presently Pursuing M.Tech (CSE) in Department of Computer Science and Engineering from Avanthi Institute of Engineering & Technology Cherukupalli(V), Near Tagarapuvalsa Bridge,Vizianagaram (Dist),Andhra Pradesh. Area of interest include Web Technologies, web programming and C programming, Machine learning and Data science.

N.Harini Working as Asst. Professor,DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING in Avanthi Institute of Engineering & Technology Cherukupalli(V), Near Tagarapuvalsa Bridge,Vizianagaram (Dist),Andhra Pradesh. Her Area of interests include Machine Learning,Classification Data Mining and Knowledge Discovery,Sentiment Analysis.