International Journal of Future Generation Communication and Networking

Vol. 13, No.2s, (2020), pp. 1603–1608

Document Clustering in Product Development Analyzer using TFIDF and K-Means Algorithm

Mohit Murotiya[1], Madhur Mahajan[2], Ketan Laddha[3], Sourabh Rathi[4],

Prof. Shreya Ahire[5,]

[1,2,3,4]*Student, Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon Bk. , Pune, Maharashtra, India, 411041*

[5]*Professor, Dept. of Computer Engineering, NBN Sinhgad School of Engineering,Ambegaon, Pune, Maharastra, India, 411041*

Abstract

The physical constructional supervisory of documents is costly in terms of time and efforts. The pass-over sizable amount of documents to interpret labouring is additionally challenging issue. Therefore, the knowledgeable means are needed to cope up with the challenges. The Clustering is altogether the motorized outcome. It is significant tool in many approaches of Data Sciences and Business logics. Document clustering classify the records into diverse gatherings called as groups, where the record in each group have same possessions as indicated in closeness or analogy/affinity measure. This paper proposed method for clustering textual documents using Automatic text classification with TF-IDF, Word embedding algorithm and classifies data using K-means clustering machine learning algorithm.

Keywords *-Document Clustering, K-Means Algorithm, Tf-Idf, Word-Net, Stop Word Removal, String Matching*

I. INTRODUCTION

The clustering is an self-activating organization or group of data, which diminishes the time and also the complexity to the great proportion/dimensions. It is considered as most unsupervised intelligent retrieval approach that deals with detecting the hidden samples and structures of high dimensional baggy information or data which manually is impossible to be done.

Therefore, to perform these tasks we use sophisticated machine learning algorithms for better organize and getting the desired information about the data. The information retreival algorithms used for grouping together the analogous data points and represents the unsupervised learning. The information points are clustered means together that are supported by some confirmed feature similarity. Like, the distance between two co-ordinate point. It is necessary to gather or collect the information into structured or understandable form for better understanding and used for go ahead to do the further activities.

Document Clustering could be an information analytic techniques, which is used for partitions the document into groups of same objects using similarity measure specified similar objects are placed within the identical cluster, and dissimilar objects are placed in distinct or different cluster.

For the text documents, the occurance, count of words and other attribures provides a scattered feature depiction with illustrable feature labels, within the recommended network. The cluster predictions are made up of using the logistic models, and have the predictions based upon logistic or multinominal regression models. Enhancing these models we ends up in extremely self-tuned descriptive clustering approaches that spontaneously selects the quantity of clusters and thus the numbers of features for every cluster that the model made.

1.1. Aim: -This model includes clustering of technical documents and all the guidance which are required for BE projects like Synopsis, PPT, Stage 1 report, Stage 2 report, Paper publishing.

1.2. Scope:

ISSN: 2233-7857 IJFGCN

To help student for understanding of Journal (IEEE) papers.
To guide student through overall product development process (Requirement gathering, Requirement Analysis, Designing, Testing, Deployment).

1.3. Motivation:

I. It is difficult job for a student to understand and implement any technical paper due to hard languages.

II. Difficulty in implementation of SDLC for real time projects.

II. LITERATURE SURVEY

[1]. "Partition based Clustering Techniques" by Dilip Singh Sisodia and Aakansha Verma

In this paper, Partition based and Hierarchical clustering techniques are debate about the grouping of document with partitions supported by clustering algorithms like K-Medoids and K-Means algorithms and Some various other methods or technologies. The various similarity measures like Euclidean distance, Jaccard Coffeficient and other similarity distance measures techniques.

It does not use mean because of the center of cluster. Instead of that, this algorithm uses a k-medoids point by using these distance measures, and therefore the addition of lowest from all of them is picked as k-point. In each of repetition, a randomly value is picked as representative within the present set of medoids which is replaced with a randomly picked representative from the gathering, if it increases the clustering quality[8 9].

These method are used for a way that the space function is calculated using SLINK and CLINK. These methods are :-

In SLINK analysis, we take distance between two clusters into account to be same because the shortest distance between those two points are specified by one point that is in one cluster and therefore the other point is in another cluster

In CLINK analysis, we take the sufficient to space between those two elements that are best distance from each other.

[2] "Multi-Viewpoint Approach for clustering" by Anjali Gupta and Rahul Dubey

The proposed system is split into some set of modules that are used to make a Efficient System. Some pre-processing steps are applied before executing the clustering algorithms is stop-word removal, stemming, term frequency and tokenization. Then after finding the k-number of clusters, we then calculate cosine similarity for the objects that are maximum dissimilarity among documents from which it owned by.

It is popular, simple and easy algorithm, but it has some limitations also. There are various types of algorithm which are available of k-mean algorithm and they work around the drawback of k-means. It mainly relies on the inceptive cluster centre selection which has the issue for selection of correct value of k-point and cluster[10 11]. The research of the improved k-means can choose the correct value of k-point by selecting high values objects as a centre of distinct cluster so they will provide an productive and necessary clustering for k-means algorithm, due to that simplicity and easily understandable of k-means algorithm that makes it choice for the several clustering applications.

However, this algorithm suffers from many deficiency same like the matter of initialize, dead point problem, and thus the predetermined number of cluster k. We introduce a totally distinct and good method for initializating the value of K that aim to looking out the suitable or proper initialization of centre for k-means.

[3] "Web Document Clustering" by Khaled M. Hammouda and Mohamed S. Kamel

ISSN: 2233-7857 IJFGCN

The clustering of documents must not be only based on single-word analysis, but also on the phrases similarlity. As the similarity between those documents should be supported by matching of phrases instead of checking on single-words only.

The paper suggested a system that is total of four components to intensify the clustering of documents problems. These components are as follows:

• The first component is efficient of establishing the weight of miscellaneous web document phrases and inseminate the document into sentence integral for go ahead processing.

• The second component is Index Graph which is used for predicating the indexing on web documents using the phrases.

• The third component is Phrase-based similarity measure which is used for examining the degree of imbricate between the documents, pair-wise document similarity. • The fourth component is Incremental Document Clustering method which supported high cluster of solidarity by enhancing the pair-wise document similarity inside all other cluster.

The applications of framework includes automatic grouping of program results, building the taxonomy and plenty of others[12 13]. This paper mainly specialise in the usage of such model on standard corpora and see its effect on clustering compared to traditional methods.

III. EXISTING SYSTEM APPROACH

The system needs to create customer-relevant business processes is an intermittent theme in marketing – particularly those dealing with the nature of marketing, competitiveness and strategies.The system converts the technical and other papers into a simple documented file. These simple file helped the students that what exactly in that paper. So we give any paper in Pdf form by using K-Means Clustering Algorithm.

Disadvantages:

- In that it only uses K-Means which takes random values for clustering due to that System sometimes gives wrong output.

IV. PROPOSED SYSTEM APPROACH

The student gets confuse regarding IEEE paper selection due to hard language of the paper as well as difficult technical terms so it is tedious job for student to understand and implement the paper. Also, student don't have knowledge about product development life cycle (Phases of product development).

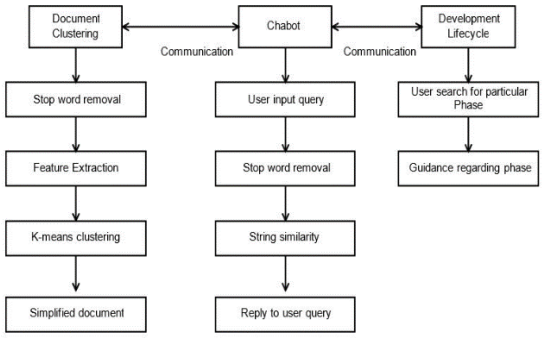

Fig 1: System Architecture

ISSN: 2233-7857 IJFGCN

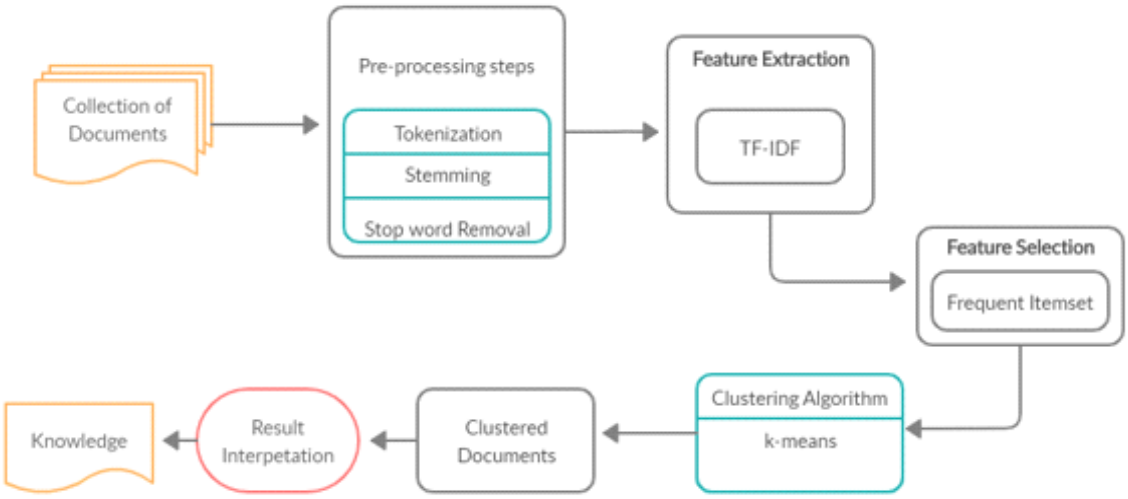

Fig2 : Proposed System Architecture

Advantage:

- Review of products sale
- Costs
- Profits projections

Disadvantage:

- Development Time & Costs
- Manufacturing Costs

V. Results

1. Stop Word Removal: -

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

What are Stop words?

Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

2. String Similarity: -

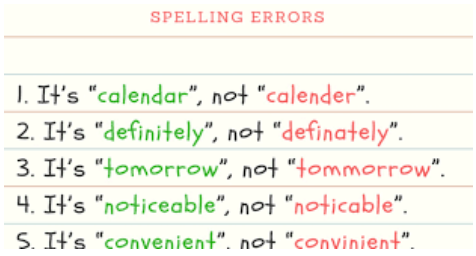

Fig 2: Example of String Similarity

Word similarity matching is an essential part for text cleaning or text analysis. Let’s say in your text there are lots of spelling mistakes for any proper nouns like name, place etc. and you need to convert all similar names or places in a standard form.

ISSN: 2233-7857 IJFGCN

3. **K-Means Clustering: -**

In pattern identification, the k-nearest neighbour algorithm is a static methods that is used for regression and classification. In both cases, the input consists of the k closest samples in the feature space. By using, we can cluster the documents.

1. Find k most similar cluster.
2. Identify set of items C, divides the documents with their heading name.
3. Clustering the top N- most frequent items in C that the active document clustered or not.

VI. Comparative Results

In our experimental setup, we are identified the stop words during matches the strings and divide the whole document content in separate clusters.

Sr. No	Clusters	Non-Clusters
1	5	2

VII. Acknowledgement

We would also like to show our gratitude to the Dr. Shwetambari Chiwhane,NBN Sinhgad College of Engineering for sharing their pearls of wisdom with us during project Clustering.

VIII. Conclusion

In the Proposed System, it deals with the problems of Clustering by applying various techniques and algorithm to make it efficient. In this paper we investigated many active algorithms and distinct approaches to improve the disadvantage of K-Means Algorithm.

References

[1] K.M., Kamel: Efficient Document Indexing of Phrase Based for Web-Clustering. IEEE Trans. on Knowledge and Data Eng. 16(10), 1279–1296 (2004)

[2] Shashank Paliwal, Vikram Pudi, "Web Investigating Usage of Text Segmentation and Inter-passage Similarities to Improve Text Document Clustering", in Proc. 8th International Conference on Machine Learning and Data Mining, Hyderabad, India, July, 2012, pp. 555–565.

[3] Hart, Susan, New product development, ch. 12, pp. 315

[4] Consumer Insight, Retrieved February 21,2010, from
http://newsroom.electrolux.com/files/2009/07/fact-sheet-consumer-insight-final.pdf
http://www.patagonia.com

[5] Philip Tavell, professional mountain biker and product developer at Craft sportswear, Borås, Sweden

[6]Udayan Birajdar, Sanket Gadhave, Shreyas Chikodikar, Shubham Dadhich, Shwetambari Chiwhane, "Detection and Classification of Diabetic Retinopathy Using AlexNet Architecture of Convolutional Neural Networks", Proceeding of International Conference on Computational Science and Application, online 05 January 2020, pp 245-253, Springer paper.

[7]Rubab Hafeez, Fahad Maqbool,Sharifullah Khan(2017). Topic Based Summarization of Multi-Document using Clustering

ISSN: 2233-7857 IJFGCN

[8] Shwetambari Kharabe, C. Nalini," Robust ROI Localization Based Finger Vein Authentication Using Adaptive Thresholding Extraction with Deep Learning Technique", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018.
[9] Shwetambari Kharabe, C. Nalini," Using Adaptive Thresholding Extraction - Robust ROI Localization Based Finger Vein Authentication", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.
[10] Shwetambari Kharabe, C. Nalini," Evaluation of Finger vein Identification Process", International Journal of Engineering and Advanced Technology (IJEAT), International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019.
[11] Udayan Birajdar, Sanket Gadhave, Shreyas Chikodikar, Shubham Dadhich, Shwetambari Chiwhane, "Detection and Classification of Diabetic Retinopathy Using AlexNet Architecture of Convolutional Neural Networks", Proceeding of International Conference on Computational Science and Application, online 05 January 2020, pp 245-253.

[12] Dr. C. Nalini, Shwetambari Kharabe, Sangeetha S," Efficient Notes Generation through Information Extraction", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S2, August 2019
[13] Shwetambari Kharabe, C. Nalini , R. Velvizhi," Application for 3D Interface using Augmented Reality", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6S2,August 2019.

ISSN: 2233-7857 IJFGCN