Document Clustering using Python

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology in

Computer Science and Engineering/Information Technology

By

Pawan Palariya (151247)

Under the supervision of

Dr. Yugal kumar

to

Jaypee University of Information Technology
Waknaghat, Solan-173234
Himachal Pradesh, India

CERTIFICATE

I hereby declare that the work presented in this report entitled **“Document Clustering”** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2018 to December 2018 under the supervision of **Dr. Yugal Kumar**(**Assistant Professor [Senior Grade]**).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Pawan Palariya (151247)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Yugal Kumar

(Assistant Professor Senior Grade)

Dept. of CSE

Dated: 08-05-2019

ACKNOWLEDGEMENT

It's our privilege to express our sincerest regards to our project supervisor **Dr. Yugal Kumar (Assistant Professor Senior Grade)**, for their valuable inputs, able guidance, encouragement, whole-hearted cooperation and direction throughout the duration of our project.

We deeply express our sincere thanks to our Head of Department **Prof. Dr. Satya Prakash Ghrera** for encouraging and allowing us to present the project on the topic **"DOCUMENT CLUSTERING USING PYTHON"** at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

At the end we would like to express our sincere thanks to all my friends and others who helped me directly or indirectly during this project work.

Date: 08-05-2019 Pawan Palariya(151247)

TABLE OF CONTENT

LIST OF FIGURES

ABSTRACT

The grouping in classes of similar objects is called the clustering process of a set of physical or abstract objects. The most common type of supervised learning is clustering. For unattended organization of document, automatic extraction of the topic and quick information recall, clustering documents is more specific technique. Rapid, high-quality document clustering algorithms helps users navigate, summarize and organize information effectively. Every day, from the extremes of large or minor portals around the world, we find a large number of documents. The K mean algorithm is the partial clustering algorithm most commonly used. The present work is directed to cluster documents using Self-Organizing Map (SOM), FUZZY C Meaning and K means algorithms. The aim of this study is to divide and data sets into K clusters with a closest significance (Similarity).

CHAPTER 1

INTRODUCTION

1.1 Introduction

Clustering is a learning procedure which automatically gather a lot of items into subsets or groups. Since there isn't standard content order rule, it is hard for the general population to utilize the enormous content data sources successfully. In this way, the administration and analysis if text information become significant. DBMS systems offers access to information store yet this was just a little piece of what could be picked up from the information. Breaking down the information by different procedures increased further learning about the information expressly put away to infer learning about the topic. This is where data mining or information came into picture.

With the exponential development of data and furthermore a rapidly developing number of content and hypertext archive oversaw in authoritative intranets, represent the accumulated learning of association that turns out to be increasingly more accomplishment in today's data society. Since there isn't standard content characterization criterion, it is exceptionally hard for individuals to utilize the gigantic content data source effectively. Therefore, the understanding and analysis of content information become significant, these days such fields of data mining, retrieval of information and data recovering have brought incredible regard for both foreign and domestic experts.

It aims to automatically make clusters by merging document clusters, it is a standout amongst the most significant errands in AI and computerized reasoning and has gotten much consideration in ongoing years. The principle accentuation is to cluster the data by attaining the best accuracy. Document Clustering has numerous significant applications in the region of information mining and data recovery. While doing the clustering investigation, we first segment the arrangement of information into gatherings dependent on information similitude and after that dole out the marks to the groups. The various calculations are utilized for clustering the reports and to improve the quality as it were.

1.2 Problem Statement

How can you study about the internal structure of the documents in a way that it seems informative? And how can you cluster and visualize the document in 2D space?

1.3 Objective

To cluster the top 100 IMDB movies on the basis of synopses using Fuzzy C means, K-means and Agglomerative algorithm.

1.4 Methodology:

It is important to understand that from collecting the documents to the collection of bunch of document is not a single operation. It includes different stages; generally there are three main phases: document representation, document clustering and feature extraction and selection.

Feature extraction starts with resolving each of document into its component parts and describe their syntactic roles to give set of features. This set doesn't have stop words. Then from the group of extracted functionality the representative features will be selected. Selection of features is an important preprocessing method used to rule out noisy features. The measurements of the features are reduced and data is much better understanded and cluster results, efficiencies and performance are improved. It is widely used in fields such as the classification of text. It is thus used primarily to improve the efficiency and efficiency of clusters. Term frequency (TF), inverse document frequency (Tf · IDF) and their hybrids are the most commonly used function selection metrics Each document in the corpus consists of k characteristics with the highest selection of metric scales, according to the best methods of choosing, and some of the improvements are made in old methods. Documentation methods include binary (presence or absence of the document), TF (i.e. frequency of the document term), and TF.IDF. We are applying clustering algorithms in the final stage of the document clustering process, grouping the target documents on the basis of features selected into distinct clusters.

Approaches for the document clustering are:.

1.4.1 Data Preprocessing:

Data preprocessing is a method of data mining that involves changing raw data into a reasonable format. Real data in certain practices or drifts are regularly fragmented, conflicted or affected and are likely to contain numerous errors. The pre-processingof data is a shown strategy for solving these problems. Crude data for further preparation is provided by preprocessing of data. In fact, data is filthy and insufficient in relation to characteristic estimates, which are short on particular characteristics of the intrigue or which contain entire data. Data that contains mistakes or abnormalities are upsetting They are contradictory because there are incoherence in codes or names. There are no quality data, so quality mining results are not available. The choice of quality must be based on data quality. The data center needs predictable value data reconciliation.

- **Stages of preprocessing in clustering:**

It is critical to underline that getting from an accumulation of documents to a clustering, is anything but a solitary activity, yet is more a procedure in different stages.

These stages incorporate progressively conventional data recovery tasks, for example, creeping, ordering, weighting, separating and so forth. A portion of these different procedures are key to the quality and execution of most clustering calculations, and it is in this way important to consider these stages together with a given clustering calculation to outfit its actual potential.

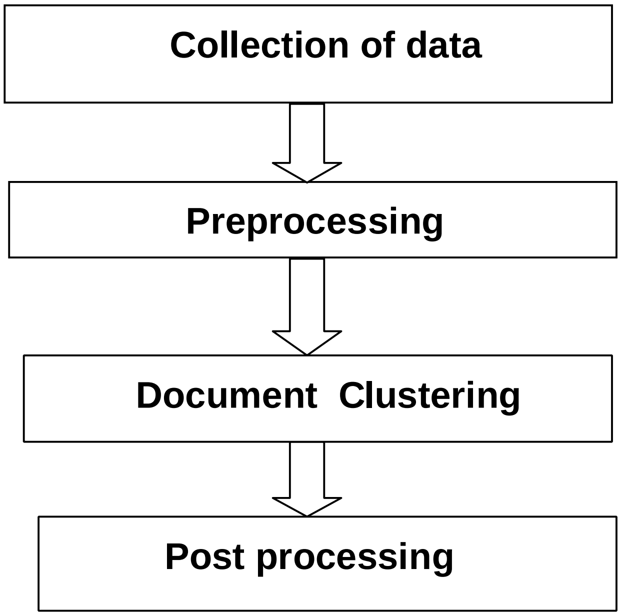


Fig 1-> stages of processing of data

- Collection of Data:

  It contains the methods such as slithering, ordering, separating and so on which document collection is used, records and recovers documents in superior ways and cans them for the expulsion of the additional information. For instance: stopwords.
- After processing:

  It is done to speak to the data in a structure that can be utilized for clustering.There are numerous methods for speaking to the documents like, Vector-Model, graphical model, and so forth. Numerous measures are additionally utilized for gauging the documents and their similitudes.

- Document Clustering:

  The cluster review is used in text documents. Clustering documents is a programmed collection of document content into clusters that allows for a higher level of resemblance between documents within a cluster, but they do not at all resemble those found in various clusters.
- Post processing:

It incorporates the significant applications in which the document clustering is utilized. For instance, the suggestion application which utilizes the consequences of clustering for prescribing news stories to the users.

- **Preprocessing techniques:**

- Stopwords:

  It is the initial stage of pre-processing that creates an attractive list of terms for the document. Word stops are the words that are sifted in the preparation of natural language information. The paper is examined to find out how many words have been described. By contrasting it and the stop word list, stop words are expelled from each document. This procedure reduces the number of words in the paper.

  This system helps in improving the adequacy and productivity of text handling as they lessen the ordering file size. There can be no final list of stop words. An example of stopword lists is given below:

  A an and are as at be by for from has he in is it it‟‟s of on that the to was were

  The phrase query "I love basketball" ,which contains one stop word(I), which will be removed from the phrase, whereas only "love" and "basketball" remains unchanged. This is the way how this method works with larger documents. These words have to real meaning in the content of the texts.

- Stemming:

  It is a process to reduce different words (common form) to their roots. Stemming is a strategy to find methods for search terms to improve the retrieval adequacy and to reduce the indexing of files. Stemming is typically achieved through expulsion from file terms of any add-ons and prefixes (appends) prior to the actual assignment of the term to the index. Since ot means the structure of the word but is different, it is important to differentiate between each structure of the word and its structure. All sorts of stemming calculations were created to do this. For example: The word “like” has its forms like likes, likely, liking, liked.

A famous algorithm for stemming is:

Porters stemming algorithm: Although the Lovins stemmer is the oldest stemmer by far , I'd say that Martin Porter's Porter Stemming algorithm is by far the most popular one (i.e., the mostly used one).It basically converts the word to its base form.

The rule of the algo:

<condition> <suffix> → <new suffix>

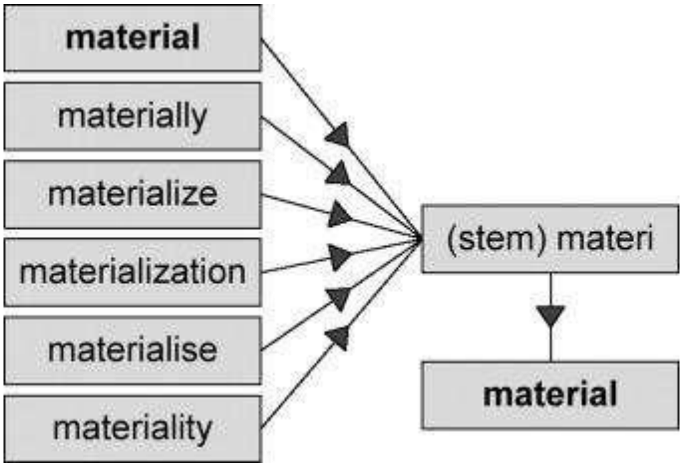


Fig 2-> example of stemming

1.4.2 Document representation:

A number of keywords / conditions taken from the document indicates a document. Documents are denoted by a model called a space vector model for clustering algorithms. Every document(d) in this model is said to be avector(s) in term-space. The vector(tf) indicates each document.

dtf = (tf1,tf2,……,tfn), where tfi is the number of times the ith term is occurred in the document.

Now if I talk about the idf matrix We also want to use the frequency of the term in the collection for weighting and ranking. Rare terms are more informative than frequent terms. We want low positive weights for frequent terms and high weights for rare terms. Hence, Tf x idf (tfidf) product of a word gives the result that the word is so frequent in the document that it multiplies the singularity of the word w.r.t. The wording of tf-idf is as follows:

Wij = tfi, j * log (N / dfi) when ith term weight is in jth document, N is total document no, tfi, j= number of ith term occurrences in Jth document, dfi = the total number of ith term documents.

1.4.3 Dimensionality Reduction:

- Single Value Decomposition

  you see most of these answers focus on SVD as a means of reconstructing/approximating a matrix AA as the sum of outer products that are scaled. problem is there at all, but if AA can be think as linear transformation? SVD gives us understanding into the function that is Ax:Rn⇒Rm. With doing some math, we can gain some real meaning.

  Let's be accurate, Let AA called as a matrix of rank r having m rows and n number of columns. SVD tells us that:

  A=UΣVT

  where:

  •U is a square matrix m x m and is orthonormal. It have columns that are left singular vectors.

  •Σ is a matrix of m x n having only diagonal elements which starts from the top left, elements are r positive values. These values are singular values.

  •VT is a square matrix m x m and is orthonormal. It is having columns that are right singular vectors.

- Multi-dimensional scaling(MDS):

  Multi-dimensional scaling is a distance-preserving manifold method for learning. These manifold learning algorithms are assumed as their dataset lies on manifold that is non-linear and smooth and is of low dimension such that this mapping f: RD -> Rd where (D>>d) can be easily calculated by maintaining one or more than one property of some higher dimensional space. Distance preserving methods supposes that the manifold can be calculated by the distances of its points which is to be found pairwise. In these distance preserving methods of multi-dimensional scaling, a low dimensional enclosing is found from the dimension which is higher in a way that the pairwise distances between these points are not changed. Some distance preserving methods preserve spatial distances (MDS) while some preserve graph distances.

MDS is not just a single approach but more than that i.e., a house of approaches. MDS takes a dissimilarity matrix D where Dij tells the dissimilarity between two points, point i and point j and produces a mapping that is on a lower dimension, maintaining the dissimilarities as firmly as possible. The dissimilarity matrix can be calculated or observed from the dataset that is given. MDS has been extensively popular and is made in these fields of human sciences like anthropology, sociology and especially in field of psychometrics.

MDS can be partitioned into two classifications:

•Metric MDS - Metric MDS is utilized for data that is quantitative and endeavors to protect the authentic dissimilarity measurements. Given a dissimilarity framework D ,f which is a monotone function, and p which refers to dimensions in subspace metric MDS attempts to locate an ideal design $X \subset Rp$ s.t. $f(Dij) \approx d\ ij = (xi - xj)2$. Another kind of metric MDS is classic MDS (original MDS) detailing which gives shut structure arrangement. Instead of attempting to surmised the dissimilarity measurements in lower dimension, it utilizes decay of eigenvalue for in the arrangement.

•Non-Metric MDS - Non-metric MDS is utilized for ordinal data. It attempts to maintain the control of dissimilarity measurements unblemished. For instance if Pij is dissimilarity between ith and jth and P32 > P89, then non-metric MDS makes a mapping s.t.d32> d89.

- Steps for MDS:
  - Make a dissimilarity matrix.

- choose a point Xm at random and mark it as pivot.
- Generate the minimum of the stress majorizing functions.
- if (σ(Xm) − σ(Xmin)) < e, break, else set Xm=Xmin and go to step 2

1.4.4 Document Clustering:

The clustering techniques used here :

K-means algorithm

Fuzzy-C means algorithm

agglomerative hierarchical clustering

1.5 Organizations:

Chapter 1: Includes the project's brief introduction to give the basic idea what to do.

Chapter 2: Includes review of literature from numerous journals, papers from different websites and conferences.

Chapter 3: Includes system design, project implementation tools and techniques.

Chapter 4: Includes analysis of performance and results.

Chapter 5: Includes the project's conclusion, limitations and future improvements.

CHAPTER 2

LITERATURE SURVEY

2.1 Term Frequency-Inverse Document Frequency(TF-IDF):

In most of the algorithms that are generally used in clustering, the dataset in which clustering is to be performed is represented in the vectors set x= {x1, x2, …, x n}, in which the vector xi is used to denote the feature vector of single entry of the data. In VSM(Vector Space Model), data in the document can be treated as dot in the multidimensional space where it is represented by vector d, such that d={w1 ,w2 ,.....,w n}, where t(I) in one document and w(I) is the term weight of term. The weight of the term is the value that represents the role and need of this term. To calculate the weight of a given term, the occurrence or the frequency of a given term within a respective document and also in the entire set of documents is to be considered or is calculated. The most commonly used weighting scheme that is also used nowadays multiplies the Term Frequency with the Inverse Document Frequency (TF.IDF). The term frequency represents importance of the term within the particular document or how the given term is important in a given document. TF.IDF is considered as the statistical measure which defines how important a given word is for a document. It is obvious that the words that are most frequent are in a document are more important than the rest, i.e. indicates the topic of the document. Let f (I, j) = frequency of Ith term in jth document. Now we will have to normalize the term frequency (t f) in the entire corpus: tf(I,j) = f(I, j) / max{fi,j} The inverse document frequency of a document is a measure of the basic importance of the term. Terms that appear in many different documents have less chance to indicate the overall topic of the documents. Let dfi = frequency of term in the document , i = ith term presence in how many documents, idfi = inverse document frequency of ith term, = log2 (N/ dfi) Where N: documents in total, TF.IDF value: wij = tfij×log2 (N/ dfi)

Improvements in weighting of traditional terms, selection of features and reduction of dimensions. authors have explored a few generally utilized unsupervised and regulated term weighting strategies on benchmark data accumulations in mix with Support Vector

Machines (SVM) and k-Closest Neighbor (kNN) calculations. Another basic supervised term weighting strategy tf:rf, to improve the terms' segregating power for text classification task is proposed. This proposed managed term weighting strategy has reliably preferable execution over other term weighting techniques. Likewise the famously utilized tf:idf strategy has not demonstrated a consistently performance act regarding various data sets.

An entropy-based component positioning technique is proposed by Liu. and Dash In the EM (Expectation Maximization) calculation, the Base Message Length basis is inferred to choose the component subset and the quantity of clusters. They proposed a technique for filtering that is free of any clustering calculation, where feature significance is measured by the commitment to the entropy index, in view of data similarity. The errand of choosing relevant features is a difficult issue in the field of unsupervised content clustering because of the nonappearance of class names or class labels. In [8] creators have proposed another mixture model named multinomial model for mixture having a feature selection (M3FS). In M3FS, the idea of component based feature saliency is presented to the blend model. An element is significant to a specific mixture part if the feature saliency esteem is higher than a predefined edge or threshold. As the feature choice procedure is treated as a parameter estimation issue, EM calculation is utilized for assessing the model. The trial on normally utilized content datasets has demonstrated that the M3FS technique has great clustering execution and highlight determination capacity. In [7], different methodologies for selection of feature like multi type feature co-choice for clustering (MFCC), weighted semantic highlights and cluster similarity utilizing non negative grid(matrix) factorization (NMF), nearby component determination for partition of various hierarchical or leveled text clustering, approach dependent on desire boost and cluster legitimacy, in view of swarm intelligence algorithm(Ant Colony Optimization) are examined.

An entropy-based strategy for ranking the feature is proposed by Dash and Liu. In the Expectation-Maximization (EM) calculation, the Base Message Length criterion is inferred to choose the element subset and the quantity of clusters. They proposed a method for filtering that is autonomous of any clustering calculation, where include significance is measured by the commitment to an entropy list, in view of data similarity. The undertaking of choosing significant highlights is a difficult issue in the field of unsupervised content clustering because of the nonappearance of class names. In [8] creators have proposed another blend model named multinomial blend model with highlight determination (M3FS). In M3FS, the idea of component dependent include saliency to the blend model is presented. An element is pertinent to a specific blend part if the component saliency esteem is higher than a predefined edge. As the component choice procedure is treated as a parameter estimation issue, EM calculation is utilized for evaluating the model. The examination on generally utilized text datasets has demonstrated that the M3FS technique has great clustering execution and highlight choice ability. In [7], different methodologies of highlight determination like multi type feature co-choice for clustering (MFCC), weighted semantic highlights and cluster similarity utilizing non negative grid(matrix) factorization (NMF), neighborhood highlight choice for partition of various leveled content clustering, approach dependent on desire augmentation and cluster legitimacy, in light of Subterranean swarm intelligence algorithm (Ant Colony Optimization) are examined.

2.2 Dimensional Reduction:

The reduction in the dimensions for large text data is now very attractive because the high dimensionality of most algorithms poses serious problems with the efficiency[7]. These algorithms are of two kinds: extraction of functions and selection of functions. New features in the feature removal are combined by algebraic transformation from their original features. Although effective, these algorithms introduce high overall computations, which make real-world text information difficult. Feature subsets are chosen directly in the feature selectionThese algorithms are widely used in real world tasks because of their efficiency, but are based on greedy strategies rather than optimum solutions. A unified optimization framework is therefore proposed, a combined approach that incorporates the benefits of both methods. The trace-oriented feature analysis (TOFA) is the selection algorithm for this new feature. TOFA's proposed goal function integrates the objective functions of many of its prominent extraction algorithms such as unattended Maximum Margin Criterion (MMC) for primary component analysis and, thus, for both supervised and unattended issues. TOFA is an objective function. TOFA can also be used by tuning a weight value to solve semi-supervised learning problems. Experimental results on multiple realworld data sets validate for dimensionality reduction the efficacy and efficiency of TOFA in text data. Reduction of the dimensions for large text data is now attractive because the high dimensionality of most algorithms causes major problems in their efficiency [7].

2.3 Similarity measures and document clustering:

Cluster analysis methods are based on similarity measurements between a couple of objects. The identification of a pair of objects involves 3 main steps:

- Choose the variables for differentiating objects,
- Choice of a weighting system for these variables and
- Selection of a coefficient of similarity to determine the degree of similarity between the two vectors.

Precise clustering requires precise determination of the proximity of a pair of objects, as proposed and widely used by a variety of similarity or distance measurements including cosine similarity, Jacksar correlation coefficient, euclidean distance and relative entropy[5], either in terms of pairlion similarities or distance. In[5], few similarities are commonly used: the euclidean distance: it is a standard metric for geometric problems. It is the usual distance of two points and can easily be measured with a ruler. The algorithm K-means also contains the default distance measurement.

- Cosine Similarity: the vector correlation corresponds to the similarity of two documents. The angle between vectors, i.e. similarity between cosines, is quantified. The Jaccard coefficient compares the sum of the terms compared to the sum of the terms in each paper, but not the shared terms. Another measure of the extent to which two vectors are related is the Pearson Correlation Coefficient. Average Kullback-Leibler DivergenceThe divergence of KullbackLeibler (KL divergence) is a common measures for assessing differences between two probability distributions, also known as relative entropy. Since many types of similarity are available to identify the degree of similarity among a pair of objects the coefficient of distance, association coefficients, probabilistic coefficients, and correlation coefficients are described by Sneath and Sokal.

 Distance coefficients: Due to their simple geometric interpretation, e.g. Euclidean distance was used extensively in cluster analysis. However, in the context of information retrieval, a major limitation of the Euclidean distance is that it can lead to two documents being considered highly similar to each other; although they do not share any common terms at all. Therefore, the Euclidean distance is not widely used for clustering documents.

• Association coefficients: they were widely used for clustering documents. It is, respectively, the number of terms common to a pair of documents with a and b terms. Standardization is essential here to handle diff size docs. The Dice coefficient and the Jaccard coefficient are two commonly used standardized association coefficients. · Probabilistic coefficients: the main criterion for the formation of a cluster here is that the documents contained therein have the maximum likelihood of being corelevant to a query together. Correlation coefficients: the use of correlation coefficients for document clustering does not appear to have been reported. The cosine function is usually used to measure the similarity between two documents, but if the clusters are not well separated, it may not work well. authors applied neighborhood concepts and links to solve this problem. If there are two similar documents, they are called each other's neighbors. The number of their common neighbors is the link between two documents. The concept of neighbors and links involves global information in measuring the proximity of two documents. This concept is proposed in conjunction with the k-means algorithms family as:

i)A new method for selecting initial cluster centroids using candidate ranks of documents;

ii) A new measure of similarity by combining cosine and link functions; and,

iii)A new heuristic function to select a cluster to be divided using cluster centroid neighbours.

The experimental results on real-life data sets showed that this approach can significantly improve the accuracy performance of document clustering without significantly increasing the execution time.

2.4 Evaluation of document clustering algorithm :

The judgement of the clustering results is one of the most valuable issues in the analysis of clusters. Evaluation is the analysis of the output to understand how good the original data structure is reproduced[6]. The evaluation methods are divided into two parts: an internal quality measure: based on pairwise document similarity the overall similarity measure is used here and no external knowledge is used here. Cluster cohesiveness can be used as a cluster similarity measure. One method to calculate the cohesiveness of the cluster is by using the weighted similarity of similarity of the internal cluster[6]. Measurement of external quality: some external data knowledge is required. Entropy is an external measure. It provides a measure of goodness at one level of a hierarchical clustering for unnamed clusters or clusters. One other external measure is called as the F-measure that defines a hierarchical clustering's effectiveness[6].

Shanon’s Entropy: Entropy could be think as a cluster quality measure[6]. The data distribution category is first calculated for each cluster, i.e. let pij denotes the probability that a cluster j member is part of category i. Then each cluster j entropy is can be find out as [6]: $Ej = - \sum pij \log (pij)$ .The Total entropy can be find out by summing each cluster entropies weighted by size of the cluster:

$m\ Een = \sum ((nj*Ej) / n)\ j=1$

Where m is clusters in total,

nj is jth cluster size and n is documents in total.

F-measure: It is a collection of the concept of information retrieval accuracy and recall. Precision in question in total by documents retrieved for a query in total. Recall can be calculated by dividing relevant documents collected for a query in total by the total number of relevant documents in the collection as a whole[6]. For jth cluster and ith class :

$Recall (i, j) = nij / ni$

$Precision (i, j) = nij / nj$

where nij is the members in total of ith class in jth cluster, nj is the members in total of jth cluster and ni is the no of members of ith class. The F-measure of jth cluster and ith class can be found with the help of recall and precision where:

$F(i, j) = (2 * Recall(i, j) * Precision(i, j)) / ((Precision(i, j) + Recall(i, j))$

For an entire hierarchical clustering, the F-measure of any class is the most extreme value it accomplishes in the tree at any node and below the average value of F-measure is taken to find its overall value.

$F = \sum ni / n \max \{F (i, j)\}\ i$

Where the max of all clusters is taken over at all levels, and n is no of documents in total. Higher F-measure value signifies improved clustering [6]. Under [9], biasness of F-measure towards hierarchical clustering algorithms is described. such that, the patterned version of the F-measure i.e. F norm, is expected to solve the problem of validation of cluster in case of hierarchical clustering. Different results of the experiments show that in checking hierarchical clustering results across different datasets with different distribution of data. un normalized F-measure less suitable than Fnorm.

Conclusions from some research papers:

[1]. Michael Steinbach of the University of New York presents the results of an experimental study of some common document clustering techniques in his paper. In particular, the two main clustering algorithms to document clustering are compared, agglomerative i.e. K-means and hierarchical clustering. Hierarchical clustering is often represented as the clustering approach of better quality, but is limited because of the complexity of its in quadratic time. K-means and their variants, on the other hand, have a time complexity that is linear in the number of documents, but is thought to produce lower clusters.

[2]. Chengxiang Zhai and Charu C. Aggarwal has given a particularized survey of text clustering issues in their paper. Clustering in the text domains is a widely studied problem of data mining. Numerous applications are found in classification, customer segmentation, visualization, collaborative filtering, organization of documents and indexing.

[3]. T.Raju and Ashish Moon used cosine similarity and correlation similarity in their paper to measure the similarity between the objects which belongs to the same cluster and the dissimilarity in those which belongs to the different clusters.

[4]. In their paper, S.Bruce and Anton V.Leouski Croft did search results clustering by comparing classification methods. Problems related to clustering algorithms for representation of document and representation of cluster are discussed. At the end paper says that maintaining sufficient terms of document clustering representation for 50-100 frequencies.

CHAPTER 3

SYSTEM DEVELOPMENT

3.1 System Architecture:

3.1.1 Hardware Specification:

Document clustering systems may sound too complex, but very few hardware cost is required. All you need is a good computer with a good editor and that's it, Not much additional hardware specifications requirement.

3.1.2 Sofware Specification:

- Python Idle 3.6.5

  It is a python editor where we will actually implement the algorithm to extract data sets, manipulate them, and accurately predict the expected results.
- NumPy

  It is mainly useful for python library to add support for a large collection of high-level mathematical functions for these arrays to operate. Along with large, multidimensional arrays and matrices.
- Matplotlib

  Matplotlib is a Python programming language plotting library and its NumPy numerical math extension. It provides an object-oriented API to embed plots for general purpose applications.

- Pandas

  It is also a library for manipulation and analysis of data by Python. It provides data structures and operations to manipulate numerical table and time series data in particular.
- Pandas-DataReader

  Used to extract data from a large number of web sources.
- BeautifulSoup4

BeautifulSoup4 is an HTML and XML document parsing package of python. It generates a parse tree with the help of parsed pages that are used to extract data from web pages, which is also helpful in the case of web scraping.

- Scikit-Learn/SkLearn

  Machine learning python library consists supporting vector machines or SVM, gradient boosting, random forests, and k-NearestNeighbour which are algorithms of classification, clustering and regression.

- Word Cloud

  Word Cloud is a technique for visualization of data used for representing the text data in which the size or dimension of each word shows its importance or frequency. Important textual data points could easily be represented with the use of a word cloud. Word clouds are extensively used for inspecting data from social network websites.
  For the generation of the word cloud in Python, modules that are needed
  – matplotlib, pandas and wordcloud.
  For the installation of these packages, you will have to run these commands :

```
pip install matplotlib
pip install pandas
pip install wordcloud
```

Advantages of Word Clouds :

1. Analyzing the feedback of employee and customer.
2. Used to Identify the newly SEO keywords used for target.

Drawbacks of Word Clouds :

1. Word Clouds are not always perfect for each and every condition.
2. optimized data should be used for context.

3.2 Software Process Model:

3.2.1 Waterfall Model:

The model we used here is the model of the waterfall. Classical model of waterfall is the basic model of life cycle development of software. It's simple but it's idealistic. This model was very popular earlier but is not used nowadays. But it is very important because the classical waterfall model is based on all other software development life cycle models.

Classical model of waterfall divides the cycle of life into a set of phases. This model believes that after completion of the previous phase, one phase can be started. That is the one-phase output will be the input to the next phase. Thus the process of development can be considered in the waterfall as a sequential flow. The phases are not overlapping here.

Advantages of Waterfall Model

The advantage of the development of waterfalls is that it enables control and departmentalization. A schedule can be set with deadlines for each stage of development and a product can proceed one by one through the phases of the development process model.

The waterfall model advances through states that are easy to use. These states can easily be understood and explained too.

Due to the model's severity, it can be managed easily – each phase has a review process and outcomes that are specific.

Phases are firstly processed and then finished independently in this model and do not overlap. The waterfall model works well for smaller projects in which requirements are very well understood.

Disadvantages of Waterfall Model

Time and costs for each of the stage in the development process are difficult to estimate.

Once an application has been tested, Going back and changing something that wasn't well thought out at the concept stage is very difficult.

This model is not good for complex projects that are also object oriented.

If there is any high or moderate risk of changing requirements in the project, this model is not suitable.

3.3 Technologies used:

3.3.1 Python

Python is a language of open source programming that is both good and easy to read. It was made in 1991 by a programmer named Guido van Rossum. Python is named after the Flying Circus show by Monty Python. The show's jokes include many examples and tutorials.

Python is an interpreting language. For running, there is no need to compile interpreted languages. On any computer that it can run on itself, a program called an interpreter runs python code. This means that if the programmer has to change the code, they can quickly see the results. This also means that Python is slower than a compiled language such as C because it does not run machine code directly.

Python is a good programming language for beginners. It is a high-level language, which means a programmer can concentrate on what to do instead of how to do it. Writing programs in Python takes less time than in another language.

Programming languages other than python such as C, C++, Java, Perl, and Lisp inspired Python.

Some things Python frequently uses are

- Web development
- Programming for games
- Building GUIs for desktop
- Scientific and Network programming

Advantages of Python:

•Python is the most widely used developers recently compared to other programming languages. In the next paragraphs, in contrast to other languages, we will look at the advantages of Python programming language for developers.

•The main advantages of the Python language are that reading is easy and that learning is easy. If we compare writing in python and C, C++, It is much difficult in C and C++. You gain the opportunity to think fresh while coding with this language, maintaining the code is also very easy. Which reduces program maintenance costs and is seen as one of the benefits of Python's programming.

•An important advantage of Python language is that it is widely used by scientists, engineers, and mathematicians. That's why Python is so useful for prototyping and experiments of all kinds. In many groundbreaking fields, it is used. It is also used in film animation and machine learning.

Disadvantages of Python:

•Python acquires poor speed of execution as an interpreted language. C and C++ are much faster than python. it do not work as a compiler but as interpreter.

•The language is considered less suitable for development of mobile devices and games. It is often used on desktop and server, but Python has only developed several mobile applications. Another drawback Python has is the error of runtime. The language has many design limitations and requires more time for testing. The programmer can only see bugs during runtime.

•Python consumes a high amount of memory and is not used in web browsers because it is not safe.

3.4 Algorithm Design

Numerous methods, for example, Naïve bayes, Nearest Neighbor, Support Vector Machines, Regression, Decision Trees, TFIDF Style Classifiers and Association classifiers have been created for text classification.

In this project, we use an association rule mining algorithm (Apriori calculation) to recognize suspicious email and the further order into the classification into the alert and informative or normal emails.

It is developed explicitly for identifying irregular and beguiling correspondence in email. The proposed technique is executed utilizing the Java language. In this algorithm, there are three sections: Email Preprocessing, Building the associative classifier and validation.

3.4.1 K-means:

K-means is one of the simplest and basic unsupervised learning algorithms which solves the well-known problem of clustering. A cluster's centroid is formed in such a way that it is closely related to all cluster objects:

- Algorithmic steps for K-means clustering:

The algorithm follows a simple and easy way through a certain number of clusters to classify a given set of data.

STEP 1:

We assign ‘K’ cluster centers randomly(centroids). Let’s say these are c1,c2,…,ck, where: ‘C’ is the set containing all of the centroids.

$$\boxed{C= c_1,c_2,....\ c_k}$$

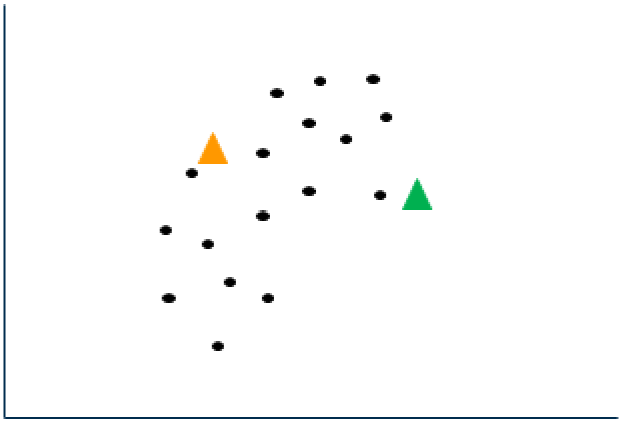


Fig 3-> shows randomly assigned cluster centers

STEP 2:

We assign each data point to the nearest center in this step, this is done by calculating the distance from Euclidean. Where, dist() is the distance from Euclidean.

$$\boxed{\underset{c_i \in C}{\arg\min}\ dist\,(c_i\ ,x\ )^2}$$

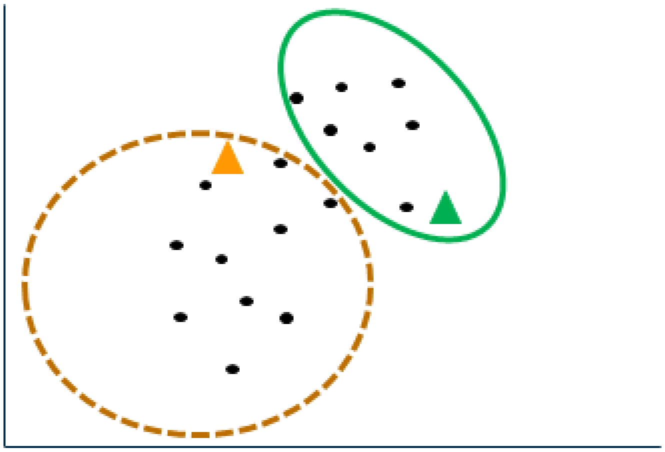


Fig ->4 shows clusters formation for first time

STEP 3:

We find the new centroid in this step by taking the mean of all the points assigned to a particular cluster. 'Si ' is the group of all the points in the ith cluster.

$$c_i = \frac{1}{|Si|} \sum_{xi \in Si} x_i$$

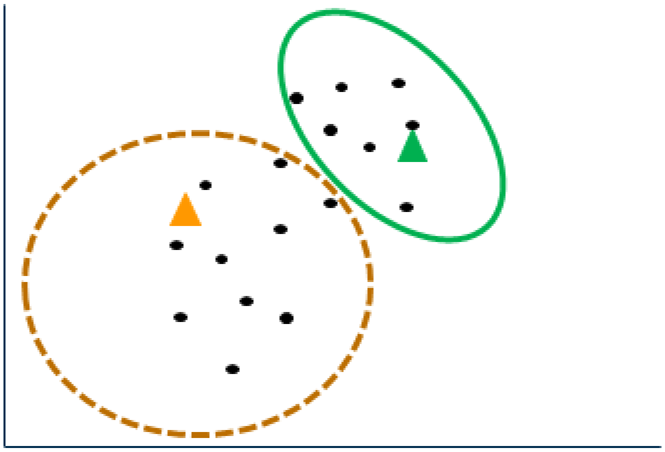


Fig->5 newly formed clusters

STEP 4:

We repeat step 2 and 3 in this step until there is no change in the cluster assignments. That means we repeat the algorithm until our clusters remain stable

- Pros:

1. Facilitate to understand, fast and strong,.

2. Relatively efficient: O (tkn), where k is the number of clusters, t is the total iterations and n is the no of objects, and generally : k, t << n.

3. Gives best result when data set are distinct or well separated from each other.

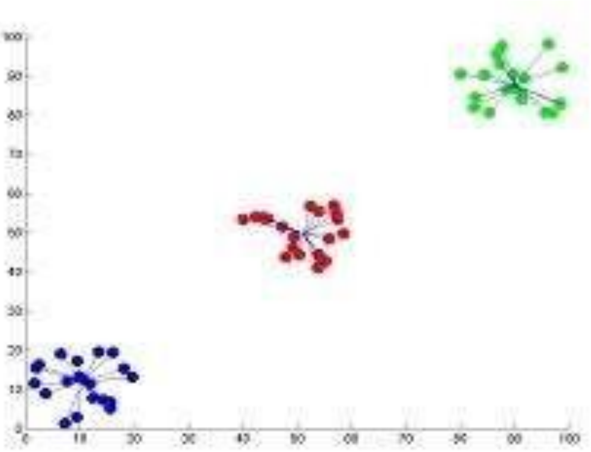


Fig->6 this graph shows the clusters when 'N' = 60 and 'k' = 3 by applying k-means.

- <u>Limitations:</u>

1.This algorithm requires theoretical details for defining number of cluster centers.

2.This algorithm does not invariate non-linear transformations, i.e. we obtain different results with different data representation (data is shown in the form of cartesian coordinates and polar coordinates will yield dissimilar results).

3. It takes lot of time.

4. It's not going to scale well.

5. The outcome is delicate to the initial allocation.

6. As a metric, variance and euclidean distance is used as a cluster scatter measure.

7.The learning algorithm provides the squared error function with the local optimum.

8. Choosing the cluster center at random cannot lead us to the desired outcome.

9. Relevant only when the mean for categorical data is defined, or fails.

10.The noisy data and outliers cannot be handled

11.For nonlinear data set, the algorithm fails.

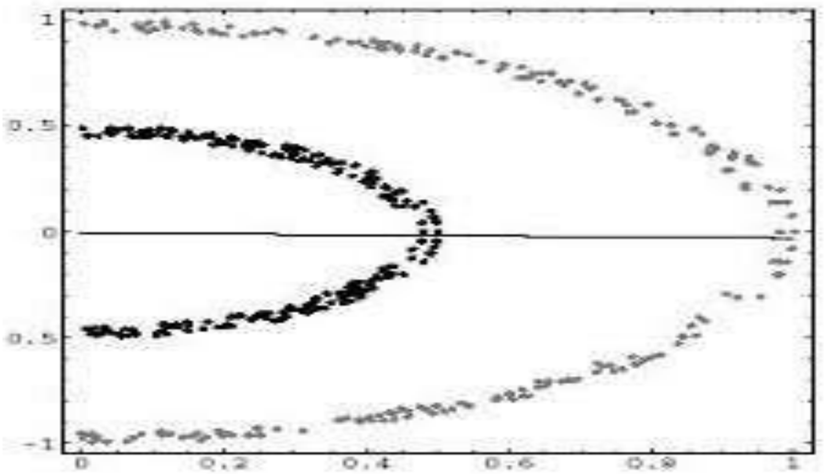


Fig 7 : showing the dataset which is non-linear(k-means fails here)

3.4.2 Agglomerative hierarchical clustering:

Hierarchical algorithms of clustering are either bottom-up or top-down.

1. At the beginning, bottom-up algorithms treat each document as a singleton cluster and then merge (or agglomerate) pairs of clusters successively until all clusters are merged into a single cluster containing all documents. Hence, hierarchical bottom-up clustering is called hierarchical agglomerative clustering(HAC).

2. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering(HAC).

- Algorithms steps for AHC:

•Assign a separate cluster to each object.

•Assess all pair distances between clusters.

•Use distance values to build a distance matrix.

•Search the shortest distance from the cluster pair.

•Remove the pair from the matrix and merge it(pair).

•Assess all distances and update the matrix with the help of all other clusters also this new cluster.

•Repeat the process until single element is taken out by the distance matrix from reducing it.

- Advantages of HAC:

  It can produce an order for objects that can be informative to display data.

  Smaller clusters are generated, which may be helpful for discovery.

- Limitations of HAC:

  No provision can be made for relocating objects which may have been grouped ' incorrectly ' at an early stage. To make sure it makes sense, the result should be closely examined.

  Different distance metrics can be used to measure distances between clusters. To support the veracity of the original results, it is recommended to perform multiple experiments and compare the results.

3.4.3 Fuzzy c-means:

In fuzzy clustering, each point is likely to belong to each cluster instead of belonging entirely to one cluster, as it happens in the case traditional k-means algorithm. Fuzzy k-means clearly attempts to solve the problem where points are somewhere close to the centers or otherwise vague by replacing separation with probability, which could be a distance function.

Fuzzy k-means, based on those probabilities, uses a weighted centroid. Initialization processes, termination and iteration processes are the same as from those used in algorithm k-means. The clusters at the end are best considered as probabilistic distributions instead of a hard label assignment. It should be realized k-means is a particular case of fuzzy k-means. If probability function value is 1 then data print is very close to the centroid and 0 in rest of the cases.

FCM minimizes the objective function as follows:

$$J_{FCM} = \sum_{i=1}^{c} \sum_{j=1}^{n} {\mu_{ij}}^{m} \left\| x_j - v_i \right\|^2$$

FCM is extensively used in the fields of agricultural engineering, chemistry, target recognition, image analysis, medical diagnosis, geology, astronomy, and form analysis.

.

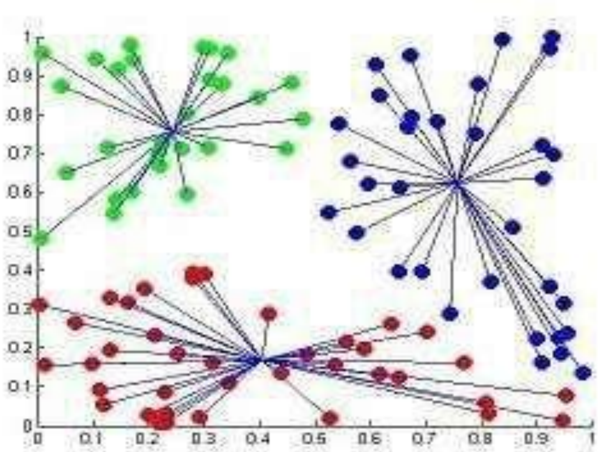


Fig 8: clusters formed by fuzzy c-means

- <u>Algorithm of fuzzy c-means:</u>

  x: dataset (unlabeled), k= the total number of cluster to be made, m= objective function parameter, e= A convergence criteria threshold

  1. Set the initial value of prototype V= {v1,v2, …, vk} , where the number of clusters k is fixed
  2. Initialize the k-means µk randomly. µk associated with the clusters and find out the probability of each data point xi. xi belongs to the cluster k,

     P(pointxihaslabelk|xi,k)P(pointxihaslabelk|xi,k).

     Vprevious □ V
  3. Compute membership functions

$$\mu_{ij} = \frac{1}{\sum_{j=1}^{k} \left(\frac{\|x - v_i\|^2}{\left\|x - v_j\right\|^2} \right)^{\frac{1}{m-1}}} \quad 1 \le i \le k$$

4.Update the prototype vi in V

$$v_i = \frac{\sum_{i=1}^{n} \mu_{ij}{}^{m} x_j}{\sum_{i=1}^{n} \mu_{ij}{}^{m}} \quad , \qquad i = 1, \dots \dots \dots, k$$

5.As far as

$$\sum_{i=1}^{k} \left\| v_i{}^{Previous} - v_i \right\| \leq e$$

- Pros:

1.Gives the top results in the case of overlapping data sets, also improved than the algorithm for K-means.

2.As a result, at the expense of more iteration numbers, according to the cluster center data points are assigned.

3.When the parameter fuzzifier (member matrix exponent U) is m=1. While it would yield results identical to K-means, it would run significantly slower as it would still go through full membership matrix estimation U.

- Limitations:

1.Apriority of the number of clusters

2.Euclidean distance measurements can be equally important

3.takes long time for calculation

4.Noise sensitivity: Low (or even no) membership degree is expected for outliers (noisy points

3.4.4 Dummy Classifier:

The dummy classifier gives you a performance measurement of "baseline." One should expect to achieve the success rate even if it's just guessing.

Suppose you want to determine whether an object has a certain property or not. If you have analyzed a large number of these objects and found that 90% contain the target property, then guessing that each future instance of the object has the target property gives you a 90% likelihood of correctly guessing. Structuring your assumptions this way is tantamount to using the most common method in the documentation you quote.

Because many machine learning tasks attempt to increase the success rate of classification tasks (e.g.), evaluating the success rate of the baseline can afford a floor value for the minimum value that the classifier should over-perform. You'd want your classifier to get more than 90 percent accuracy in the hypothetical discussed above, because 90 percent is the success rate even for "dummy" classifiers.

If one uses the data discussed above to train a dummy classifier with the stratified parameter, that classifier will predict that there is a 90% likelihood that each object he encounters will have the target property. This is different from training a dummy classifier with the most common parameter, since the latter would guess that all future objects have the target property.

3.4.5 LDA (latent dirichlet allocation):

Topic modeling refers to the identification task of topics that best describe a set of documents. Only during the topic modeling process (hence called latent) will these topics emerge. and the popular topic modeling technique is the Latent Dirichlet Allocation (LDA). Although the name is mouthful, it's very simple the concept behind it.

In short, LDA imagines a set of fixed topics. Each topic is a set of words. And LDA's goal is to map all the documents to the topics in a way that those imaginary topics capture the words in each document mostly. We'll be going through this method systematically to the end that you'll be comfortable enough to use this method yourself.

- Why topic modelling?

  What is some of the topic modeling used in the real world? Historians can use LDA to identify significant historical events by analyzing year-based text. Based on your past readings, web based libraries can use LDA to recommend books. News providers can use topic modelling to quickly understand articles or group similar articles. Another interesting application is unsupervised image clustering, where each image is treated like a document.

  A distribution of topics can describe each document and a word distribution can describe each topic.

  LDA is not concerned with the word order in the document. Usually, LDA uses the representation of a document by the bag-of-words feature. It makes sense because you can still guess what kind of topics are discussed in the document if I take a document, jumble the words and give it to you.

  k— Number of topics belonging to a document (fixed number)

  V—vocabulary size

  M—No of documents

  N— Word number in each document

  w— A word in a file(document). This is represented as a one hot encoded size V (i.e. V — vocabulary size) vector.

  w (bold w): Represents a N-word document (i.e. "w" vector)

  D— a M document collection

  z— A subject from a series of k topics. A topic is a word for the distribution. Animal = (0.3 Cats, 0.4 Dogs, 0 AI, 0.2 Loyal, 0.1 Evil), for instance.

  First let's give a proper mathematical drawing the ground-based example of generating documents above.

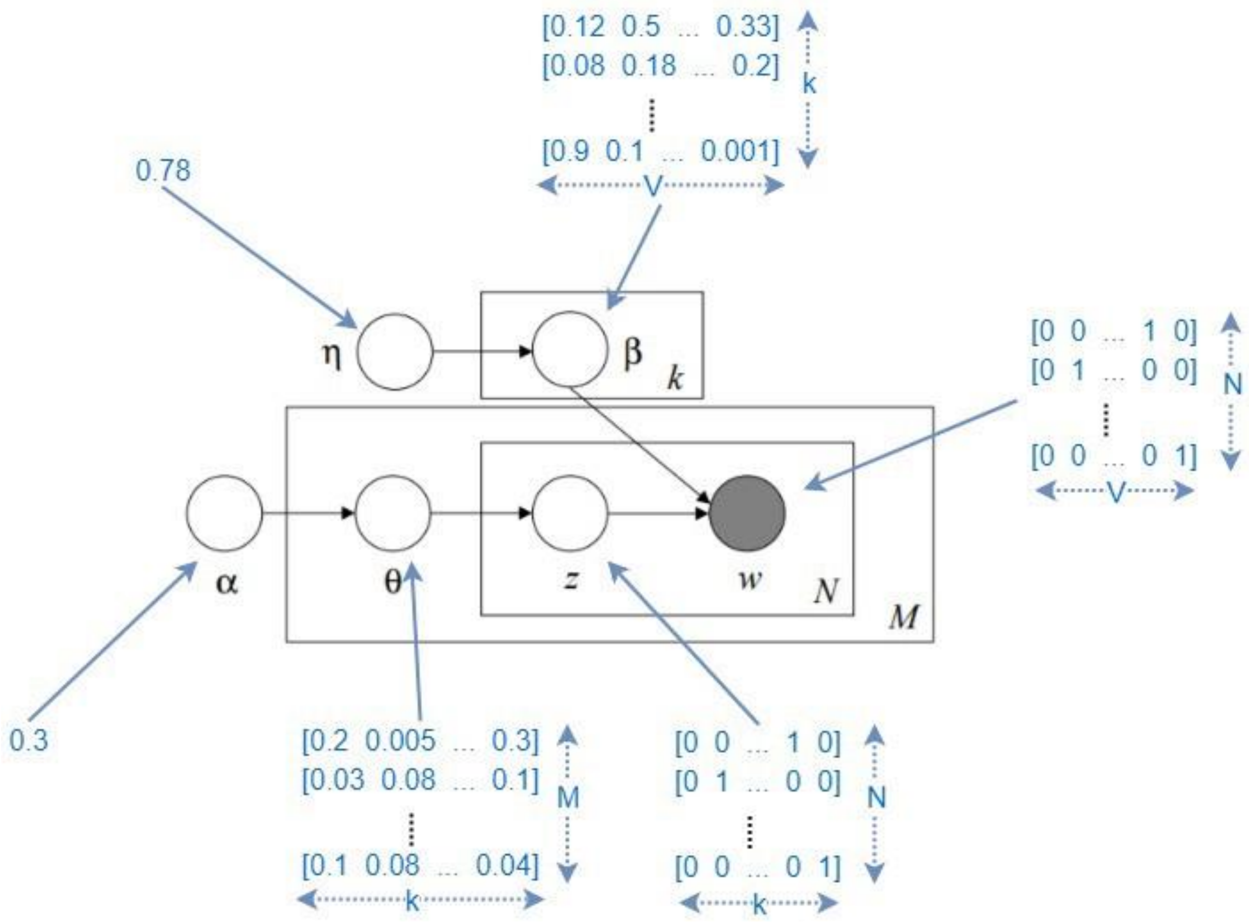


Fig 9: example of generating documents by LDA

Let's decipher what it says. We have a single α value (i.e. ground organizer θ) that defines somewhere ; the distribution of the topic for documents will be the same. For each of these documents, we have M documents and have some distribution.

Now that single document has N words and a topic generates each word. You generate N topics for words to fill in. Still, these N words are placeholders.

Now kicks in the top plate. Based on η, β has some distribution (i.e. to be precise a Dirichlet distribution— discussed soon) and according to that distribution, β generates individual words for each subject. Now you fill in a word for each placeholder (in the N placeholders set), depending on the topic it represents.

- Why are α and η constant?

  In the above picture, α and η are shown as constants. But it's more complex than that in fact. For example, for each document, α has a topic distribution (either ground for each document). Ideally, a matrix of shape (M x K). And for each topic, η has a vector parameter. It's going to be shaped (k x V). The constants in the above drawing actually represent matrices and are formed to each single cell by replicating the single value in the matrix

θ is a random matrix in which θ (i, j) represents the likelihood that the i th document contains words from the jth topic. If you look at what the ground θ looks like in the above example, you can see that the balls are beautifully laid out in the middle corners The advantage of having such a property is that, as usual with real-world documents, the words we produce are likely to belong to a single topic. This is a property created by modeling it as a distribution by Dirichlet θ. Likewise, β(i, j) represents the likelihood of the I th topic containing the j th word. And β is also a distribution for Dirichlet. Belowis the quick detour to understand the distribution of Dirichlet.

Large α values push the distribution to the triangle center where lower α values push the distribution to the corners.

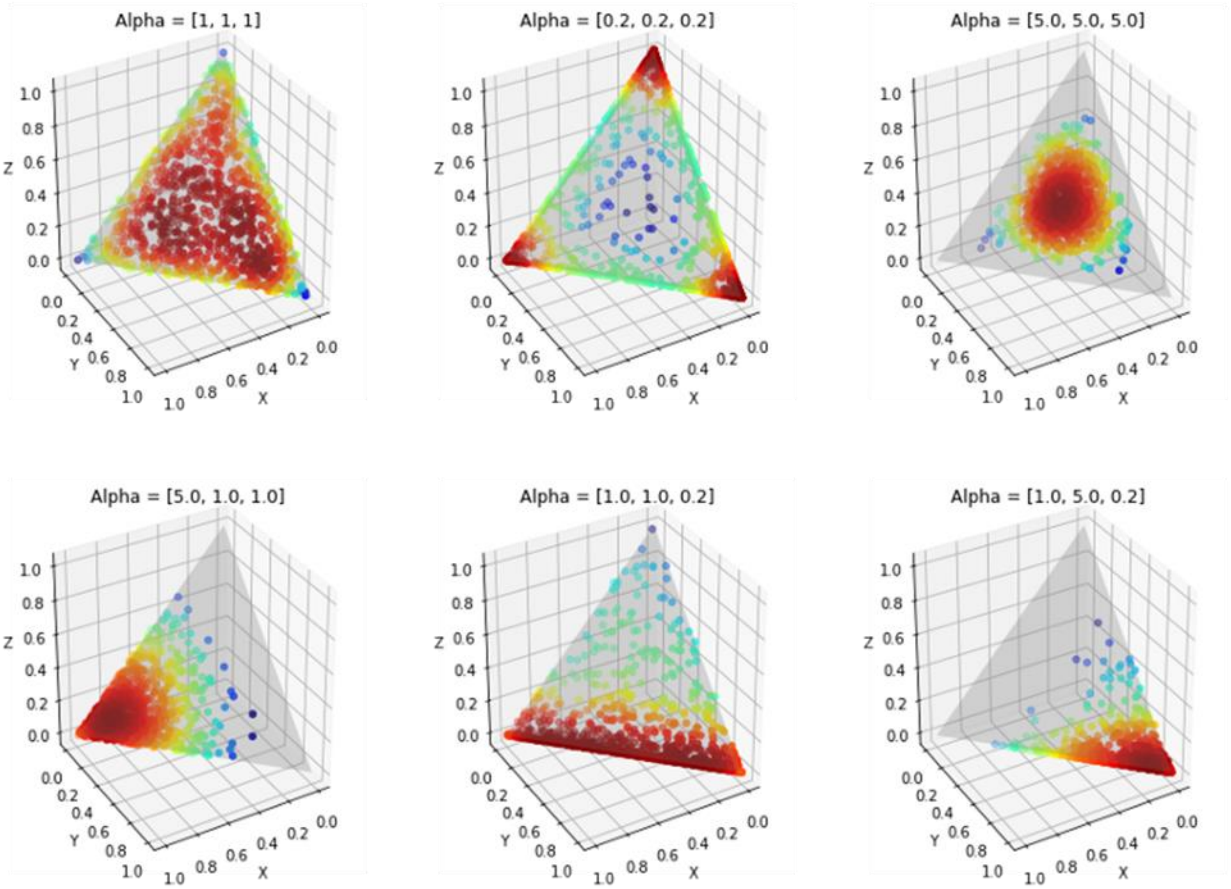


Fig 10: shows how the shape of θ changes with different α values

To know the exact values of α and η. Lets se what are the variables that we will have to find first:

α— Parameter related to the distribution that regulates the distribution of topics for all the documents in the corpus

θ— Random matrix where θ((i, j) is the likelihood that the I th document contains the j th subject

η— Parameter related to the distribution that governs how words are distributed in each topic

β— A random matrix where β(i, j) represents the likelihood that the subject will contain the word j.

we will have to find value of this term:

$$P(\theta_{1:M}, \mathbf{z}_{1:M}, \beta_{1:k} | \mathcal{D}; \alpha_{1:M}, \eta_{1:k})$$

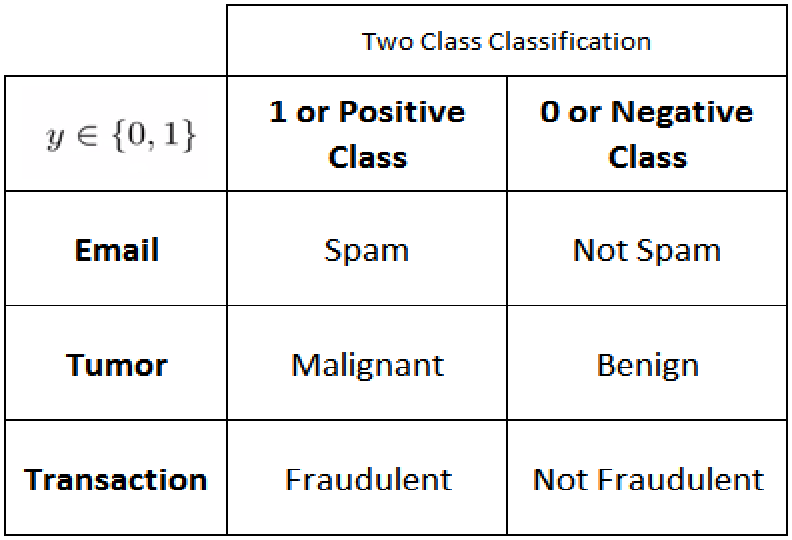
Dirichlet distribution is the Beta distribution's multivariate generalization. Here is an example of a 3-dimensional problem, where in α we have 3 parameters which affect the shape of θ (i.e. distribution). You have an N-length vector as α for a N-dimensional Dirichlet distribution.

3.4.6 Logistic regression:

It is one of the popular and basic algorithms for solving a classification problem. It is called' Logistic Regression' because the technique behind it is quite the same as Linear Regression. This term "Logistic" derives from the function called ” Logit” used in this method of classification.

What is a classification problem:

When independent variables are continuous in nature and dependent variable is in categorical form i.e. in classes such as positive class and negative class, we identify problem as classification problem. The real life example of an example of classification would be to classify mail as spam or not spam, categorize the tumor as malignant or benign, and categorize the transaction as fraudulent or genuine. All the answers to these problems are categorical, i.e. Yes or no, and that's why the classification issues are two.



	Two Class Classification	
$y \in \{0, 1\}$	**1 or Positive Class**	**0 or Negative Class**
Email	Spam	Not Spam
Tumor	Malignant	Benign
Transaction	Fraudulent	Not Fraudulent

Fig 11: example of classification

Even though, sometimes we find over 2 classes and it's still a classification issue. These kinds of issues are also called problems with multi-classification.

When not to use linear regression?

Suppose we have tumor size vs. malignancy data. Since it's a problem of classification, if we plot, we can see that all values are on 0 and 1. And if we fit the best found regression line, we can do a fairly reasonable job by assuming the threshold at 0.5.

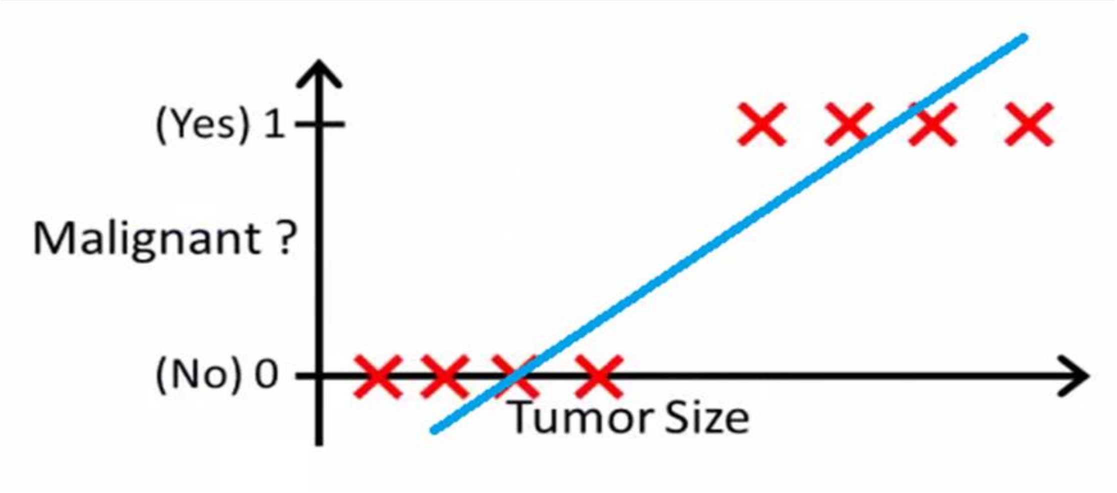


Fig 12: shows problem with linear regression

We can determine the point on the x axis from where all the values on the left side are considered to be negative and all the values on the right side are positive.

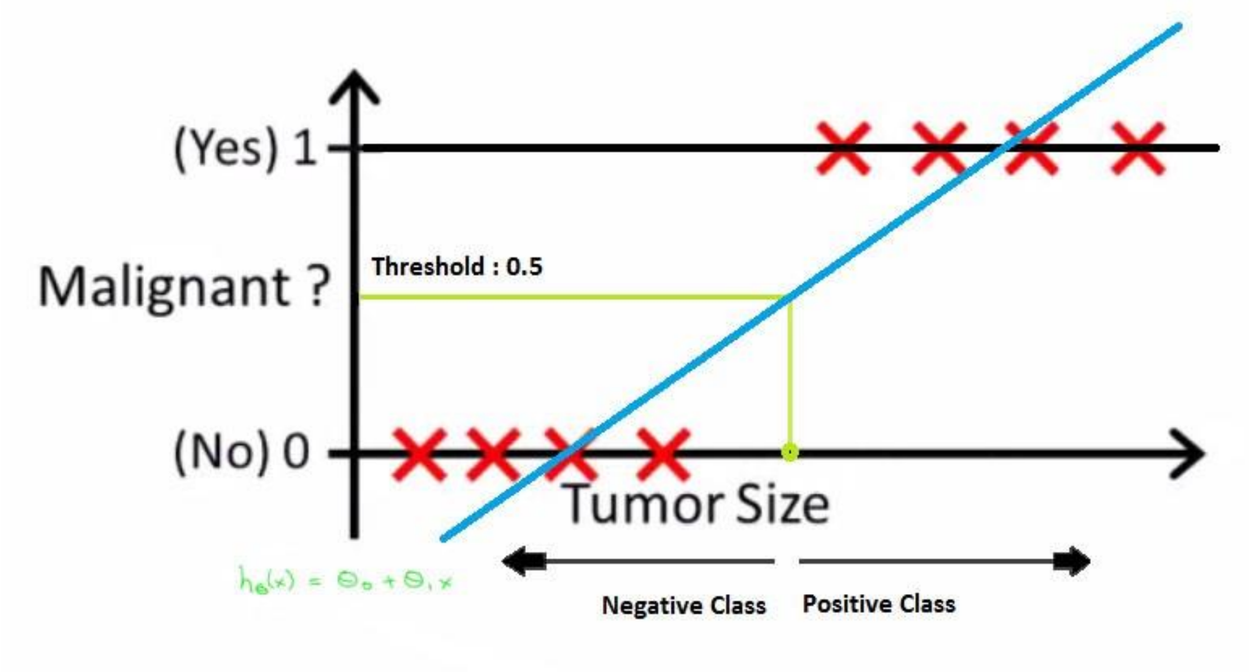


Fig 13: shows the point on x-axis on the left of which are negative values

But what if the data has an outlier. Things would be quite messy. For example, for a t threshold of 0.5,

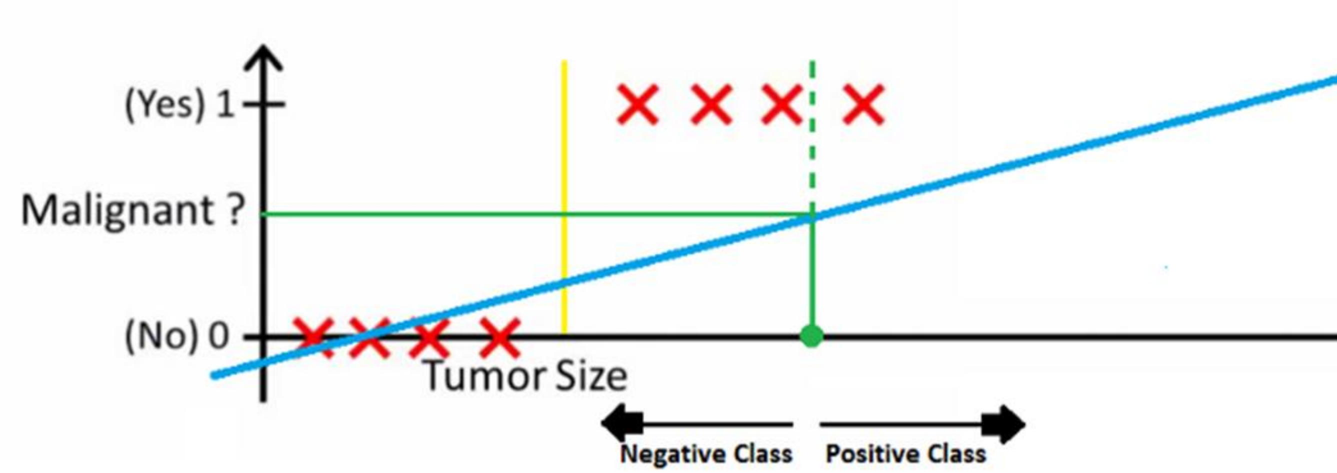


Fig 14: show if the data has an outlier

If we fit the best found regression line, it will still not be sufficient to decide any point by which classes can be differentiated. It will put in a negative class some positive examples of classes. The green dotted line (Decision Boundary) divides malignant tumors from benign tumors, but the line should be on a yellow line that clearly divides positive and negative examples. So the entire linear regression predictions are disturbed by just one outlier. And that's the picture of logistical regression.

Algorithm:

As discussed above, Logistic Regression uses Sigmoid function to deal with outliers. A logistical regression explanation can start with an explanation of the standard logistics function. The logistics function is a Sigmoid function that takes any real value from zero to zero. This is defined as

And if we plot it, the graph will be S curve,

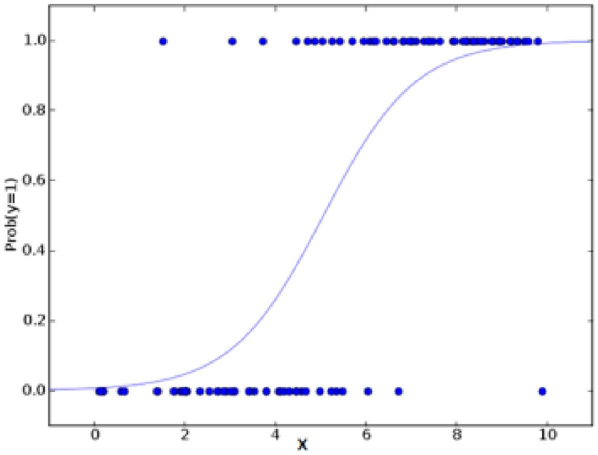


Fig 15: shows S curve of sigmoid function

Consider t in a univariate regression model as a linear function.

$$t = \beta_0 + \beta_1 x$$

So now the Logistic Equation becomes

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Now, it will take care of it when the logistic regression model comes across an outlier.

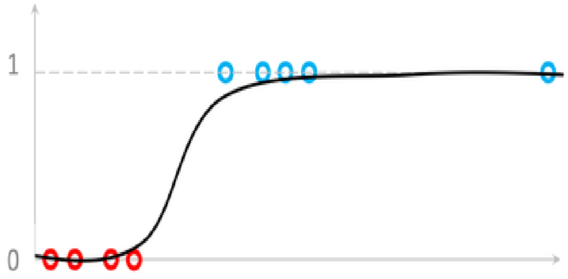


Fig 16: shift in y axis according to the outlier

But sometime, depending on the outliers positions. The y axis will be shifted to the left or the right.

What is the use of decision boundary in Logistic Regression?

The decision boundary helps you to distinguish probabilities to negative class and positive class.

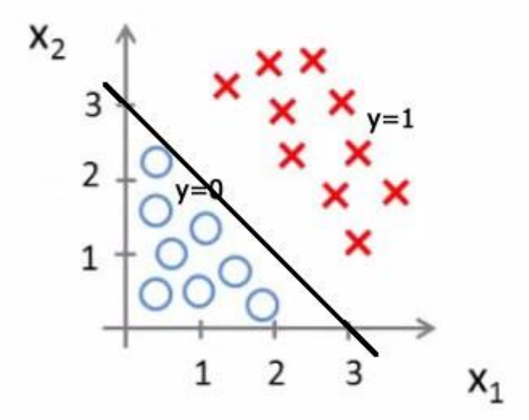


Fig 17: shows linear decision boundary

For y = 1, Equation of line would be $x_1 + x_2 >= 3$

For y = 0, Equation of line would be $x_1 + x_2 < 3$

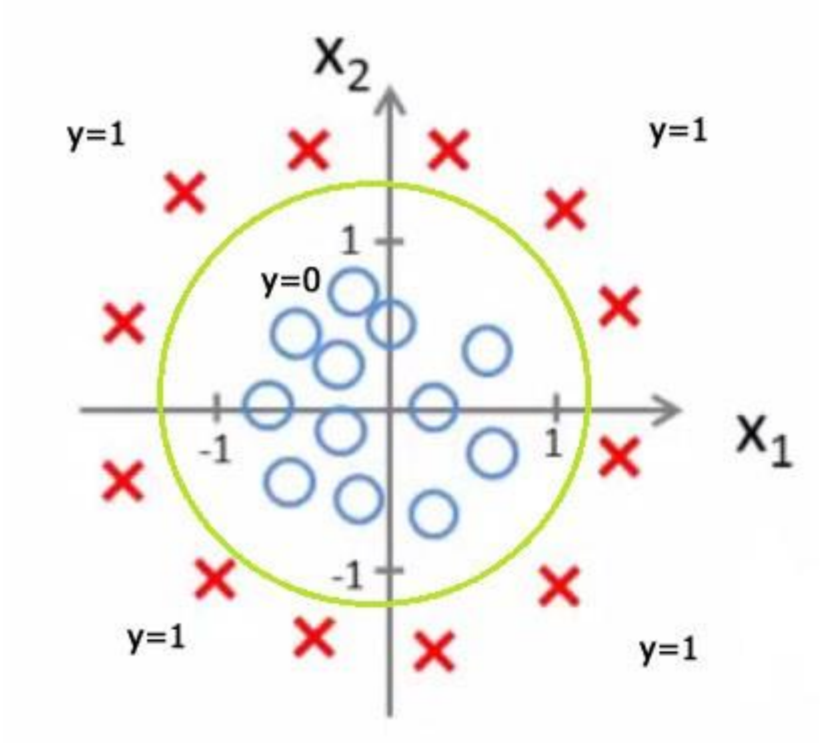


Fig 18: shows non linear boundary

For checking performance:

Confusion matrix can be used and AUC-ROC Curve can also used to check the performance.

Pros:

Linear regression is an approach that is very simple. Understanding and using is very intuitive and easy. A person can understand and use it only with the knowledge of high school mathematics. Moreover, in most cases it works. Even though it does not exactly fit the data, it could be used to determine the relationship between the two variables.

Cons:

- Linear regression by its definition only models relationships that are linear between dependent and independent variables. It assumes that there is sometimes an incorrect straight-line relationship between them. Linear regression is highly sensitive to data (or outliers) anomalies.

- Take most of your data in the range 0-10, for example. If, for any reason, only one of the data items exceeds the range, say 15, this affects the regression coefficients significantly.
- Another drawback is that if we have a number of samples available less than number of parameters, the model starts to design the noise instead of the relationship among the variables.

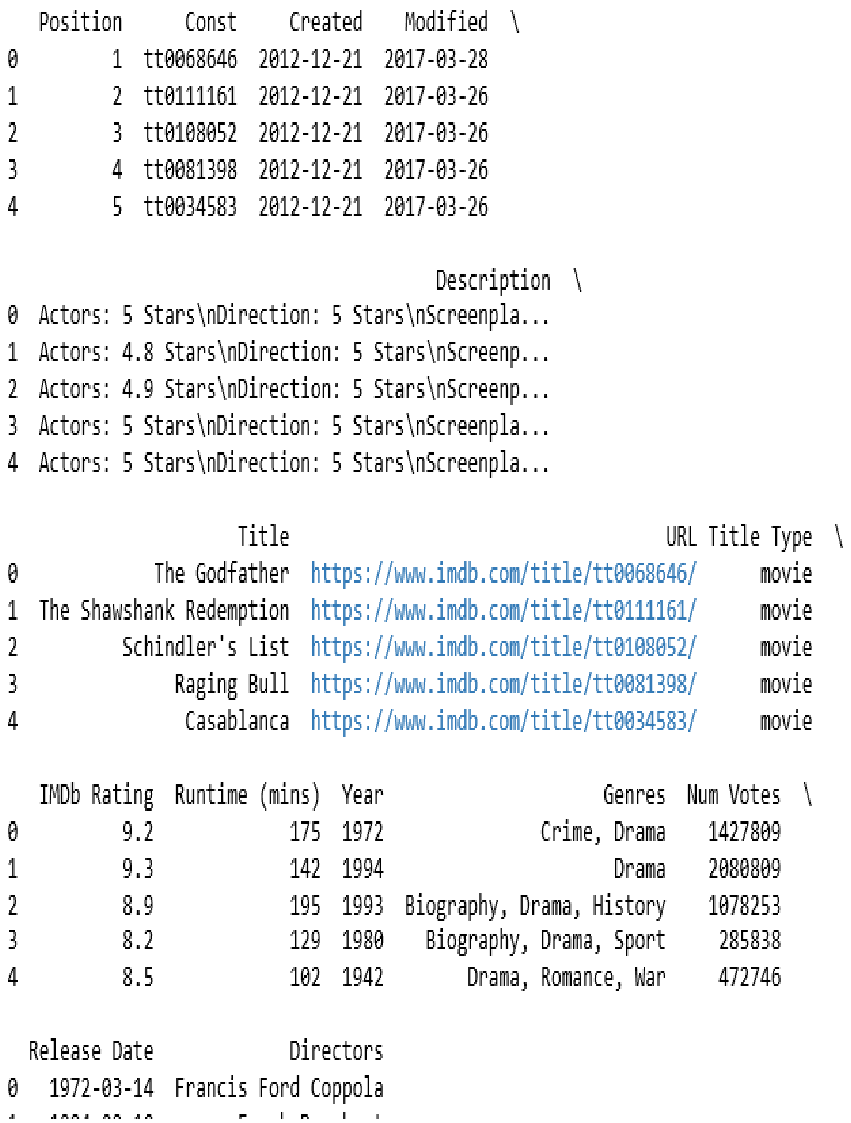
CHAPTER 4

PERFORMANCE ANALYSIS

4.1 Data Set Description

Data sets contains information about the top 100 movies which can be downloaded from https://www.imdb.com/list/ls055592025/ and separate text document file containing all 100 movies imdb synopses to which clustering is to be done.

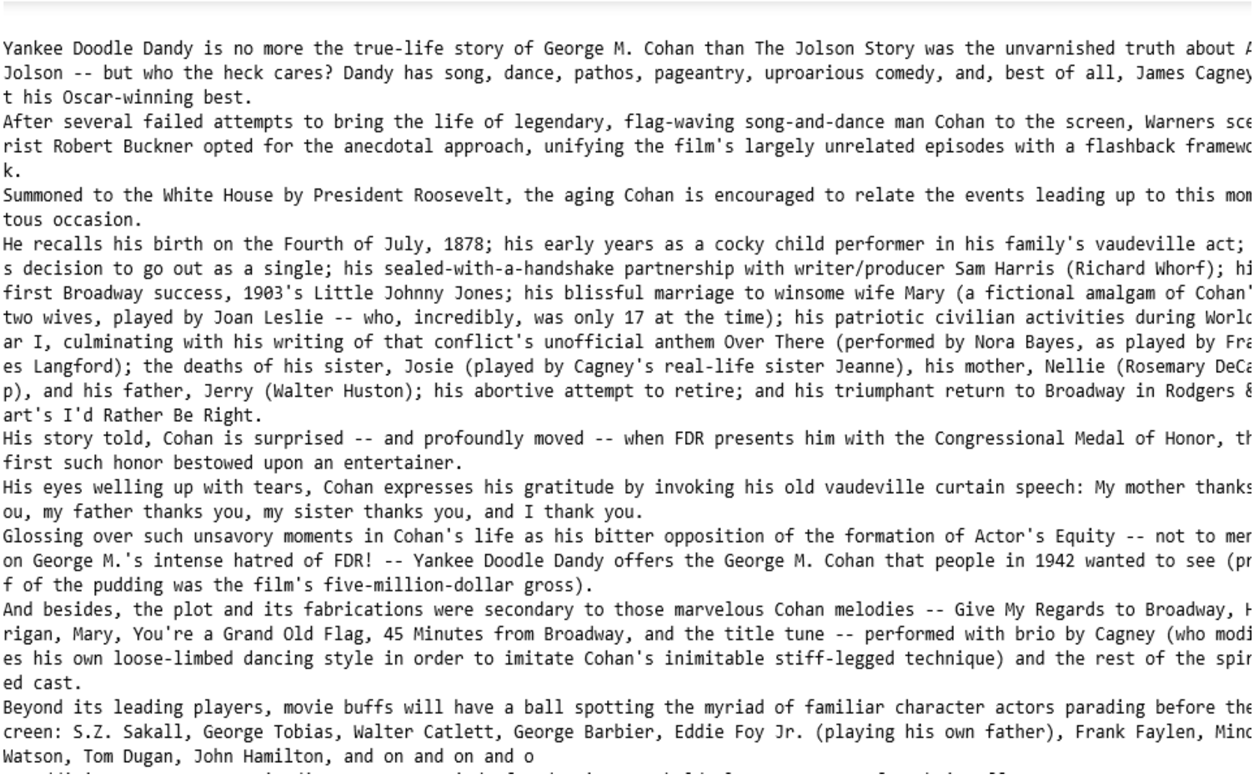
This dataset contains columns: Position, Const, Created, Modified, Description, Title, URL Title Type, IMDb Rating, Runtime, Year, Genres, Num Votes, Release Date and Directors.



```
   Position      Const     Created    Modified  \
0         1  tt0068646  2012-12-21  2017-03-28
1         2  tt0111161  2012-12-21  2017-03-26
2         3  tt0108052  2012-12-21  2017-03-26
3         4  tt0081398  2012-12-21  2017-03-26
4         5  tt0034583  2012-12-21  2017-03-26

                                         Description  \
0  Actors: 5 Stars\nDirection: 5 Stars\nScreenpla...
1  Actors: 4.8 Stars\nDirection: 5 Stars\nScreenp...
2  Actors: 4.9 Stars\nDirection: 5 Stars\nScreenp...
3  Actors: 5 Stars\nDirection: 5 Stars\nScreenpla...
4  Actors: 5 Stars\nDirection: 5 Stars\nScreenpla...

                      Title                                    URL Title Type  \
0             The Godfather  https://www.imdb.com/title/tt0068646/      movie
1  The Shawshank Redemption  https://www.imdb.com/title/tt0111161/      movie
2          Schindler's List  https://www.imdb.com/title/tt0108052/      movie
3               Raging Bull  https://www.imdb.com/title/tt0081398/      movie
4                Casablanca  https://www.imdb.com/title/tt0034583/      movie

   IMDb Rating  Runtime (mins)  Year                     Genres  Num Votes  \
0          9.2             175  1972               Crime, Drama    1427809
1          9.3             142  1994                      Drama    2080809
2          8.9             195  1993  Biography, Drama, History    1078253
3          8.2             129  1980    Biography, Drama, Sport     285838
4          8.5             102  1942        Drama, Romance, War     472746

  Release Date             Directors
0   1972-03-14  Francis Ford Coppola
```

Fig 19 : shows the dataset got from imdb website

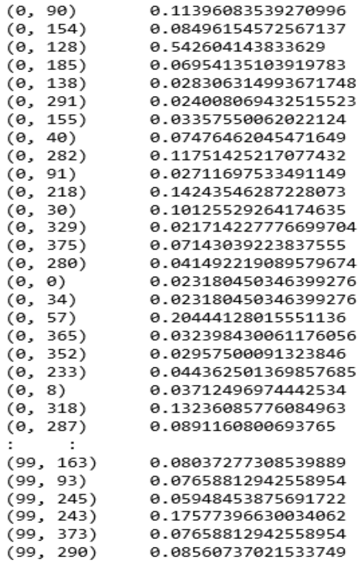
This second screenshot shows the synopses of the 100th movie in the list:



```
Yankee Doodle Dandy is no more the true-life story of George M. Cohan than The Jolson Story was the unvarnished truth about A
Jolson -- but who the heck cares? Dandy has song, dance, pathos, pageantry, uproarious comedy, and, best of all, James Cagney
t his Oscar-winning best.
After several failed attempts to bring the life of legendary, flag-waving song-and-dance man Cohan to the screen, Warners sce
rist Robert Buckner opted for the anecdotal approach, unifying the film's largely unrelated episodes with a flashback framewo
k.
Summoned to the White House by President Roosevelt, the aging Cohan is encouraged to relate the events leading up to this mom
tous occasion.
He recalls his birth on the Fourth of July, 1878; his early years as a cocky child performer in his family's vaudeville act;
s decision to go out as a single; his sealed-with-a-handshake partnership with writer/producer Sam Harris (Richard Whorf); hi
first Broadway success, 1903's Little Johnny Jones; his blissful marriage to winsome wife Mary (a fictional amalgam of Cohan'
two wives, played by Joan Leslie -- who, incredibly, was only 17 at the time); his patriotic civilian activities during World
ar I, culminating with his writing of that conflict's unofficial anthem Over There (performed by Nora Bayes, as played by Fra
es Langford); the deaths of his sister, Josie (played by Cagney's real-life sister Jeanne), his mother, Nellie (Rosemary DeCa
p), and his father, Jerry (Walter Huston); his abortive attempt to retire; and his triumphant return to Broadway in Rodgers &
art's I'd Rather Be Right.
His story told, Cohan is surprised -- and profoundly moved -- when FDR presents him with the Congressional Medal of Honor, th
first such honor bestowed upon an entertainer.
His eyes welling up with tears, Cohan expresses his gratitude by invoking his old vaudeville curtain speech: My mother thanks
ou, my father thanks you, my sister thanks you, and I thank you.
Glossing over such unsavory moments in Cohan's life as his bitter opposition of the formation of Actor's Equity -- not to men
on George M.'s intense hatred of FDR! -- Yankee Doodle Dandy offers the George M. Cohan that people in 1942 wanted to see (pr
f of the pudding was the film's five-million-dollar gross).
And besides, the plot and its fabrications were secondary to those marvelous Cohan melodies -- Give My Regards to Broadway, H
rigan, Mary, You're a Grand Old Flag, 45 Minutes from Broadway, and the title tune -- performed with brio by Cagney (who modi
es his own loose-limbed dancing style in order to imitate Cohan's inimitable stiff-legged technique) and the rest of the spir
ed cast.
Beyond its leading players, movie buffs will have a ball spotting the myriad of familiar character actors parading before the
creen: S.Z. Sakall, George Tobias, Walter Catlett, George Barbier, Eddie Foy Jr. (playing his own father), Frank Faylen, Minc
Watson, Tom Dugan, John Hamilton, and on and on and o
```

Fig 20: shows movies synopses

The third screenshot shows the tf-idf matrix generation:



```
  (0, 90)	0.11396083539270996
  (0, 154)	0.08496154572567137
  (0, 128)	0.542604143833629
  (0, 185)	0.06954135103919783
  (0, 138)	0.028306314993671748
  (0, 291)	0.024008069432515523
  (0, 155)	0.03357550062022124
  (0, 40)	0.07476462045471649
  (0, 282)	0.11751425217077432
  (0, 91)	0.02711697533491149
  (0, 218)	0.14243546287228073
  (0, 30)	0.10125529264174635
  (0, 329)	0.021714227776699704
  (0, 375)	0.07143039223837555
  (0, 280)	0.041492219089579674
  (0, 0)	0.023180450346399276
  (0, 34)	0.023180450346399276
  (0, 57)	0.20444128015551136
  (0, 365)	0.032398430061176056
  (0, 352)	0.02957500091323846
  (0, 233)	0.044362501369857685
  (0, 8)	0.03712496974442534
  (0, 318)	0.13236085776084963
  (0, 287)	0.0891160800693765
  :	:
  (99, 163)	0.08037277308539889
  (99, 93)	0.07658812942558954
  (99, 245)	0.05948453875691722
  (99, 243)	0.17577396630034062
  (99, 373)	0.07658812942558954
  (99, 290)	0.08560737021533749
```

Fig 21: shows tf-idf matrix

4.2 Results:

4..2.1 Kmeans result:

Here you can see that the clusters are formed but they are not well defined

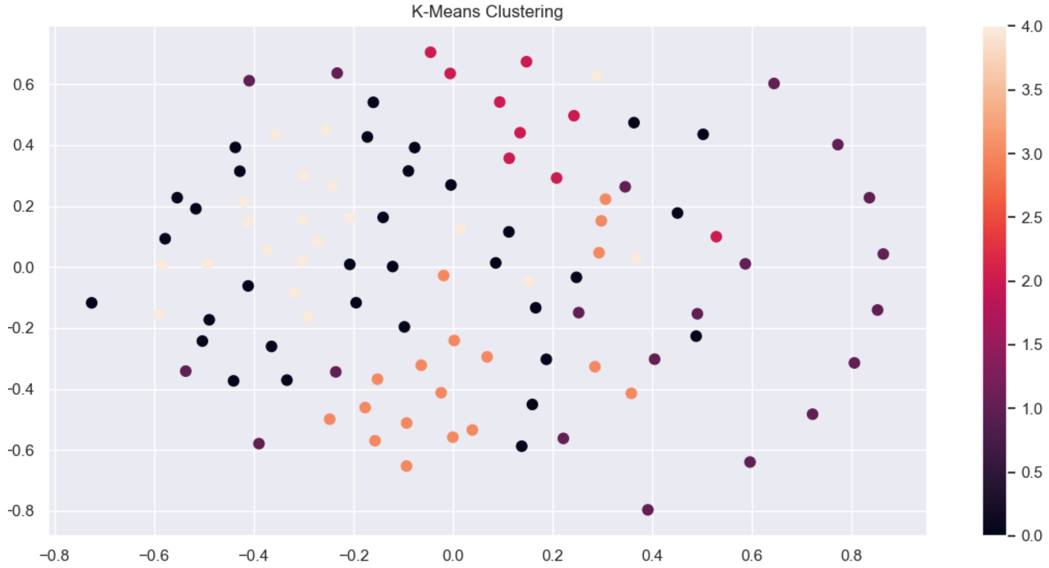


Fig 22: shows clusters formed by applying kmeans

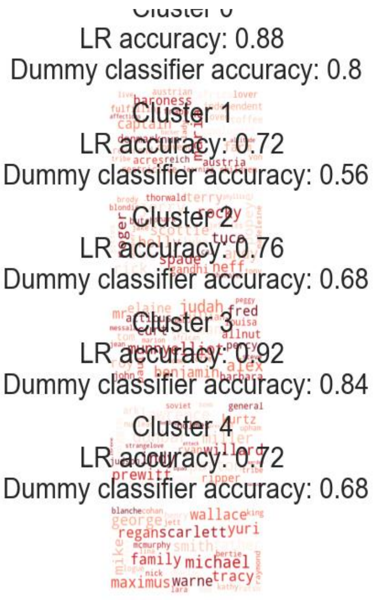


Fig 23: wordcloud representation of clusters

4.2.2 Agglomerative hierarchical clustering result:

You can see the clusters are formed and they are better than k means cluster.

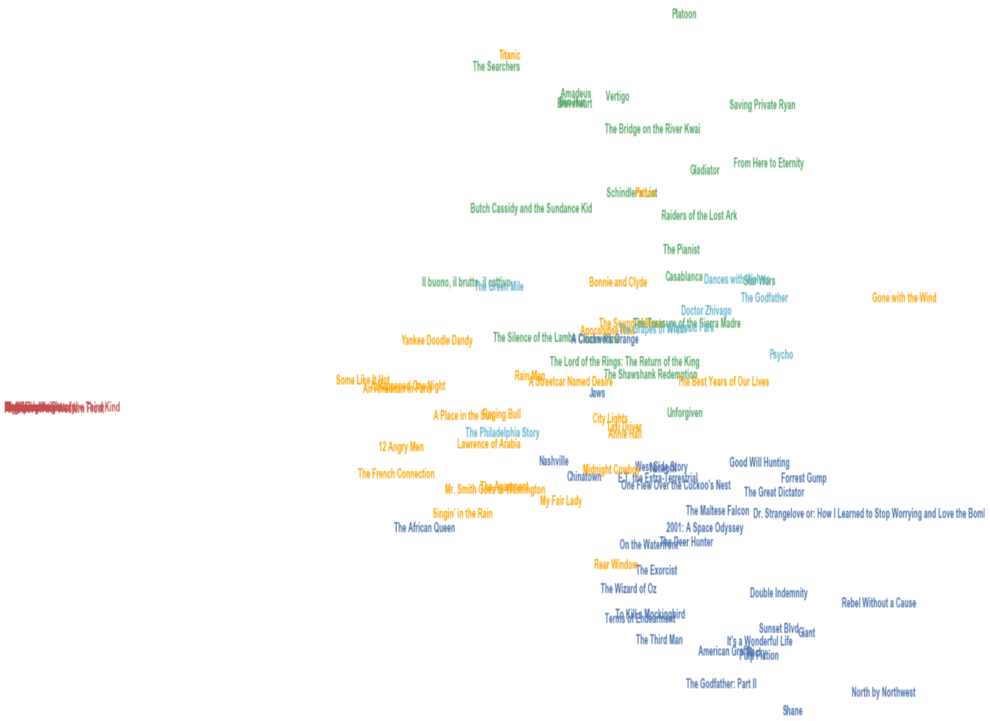


Fig 24: shows clusters formed by agglomerative hierarchical clustering

4.2.3 Fuzzy c-means result:

Here the clusters formed are better than the two. They are well defined.

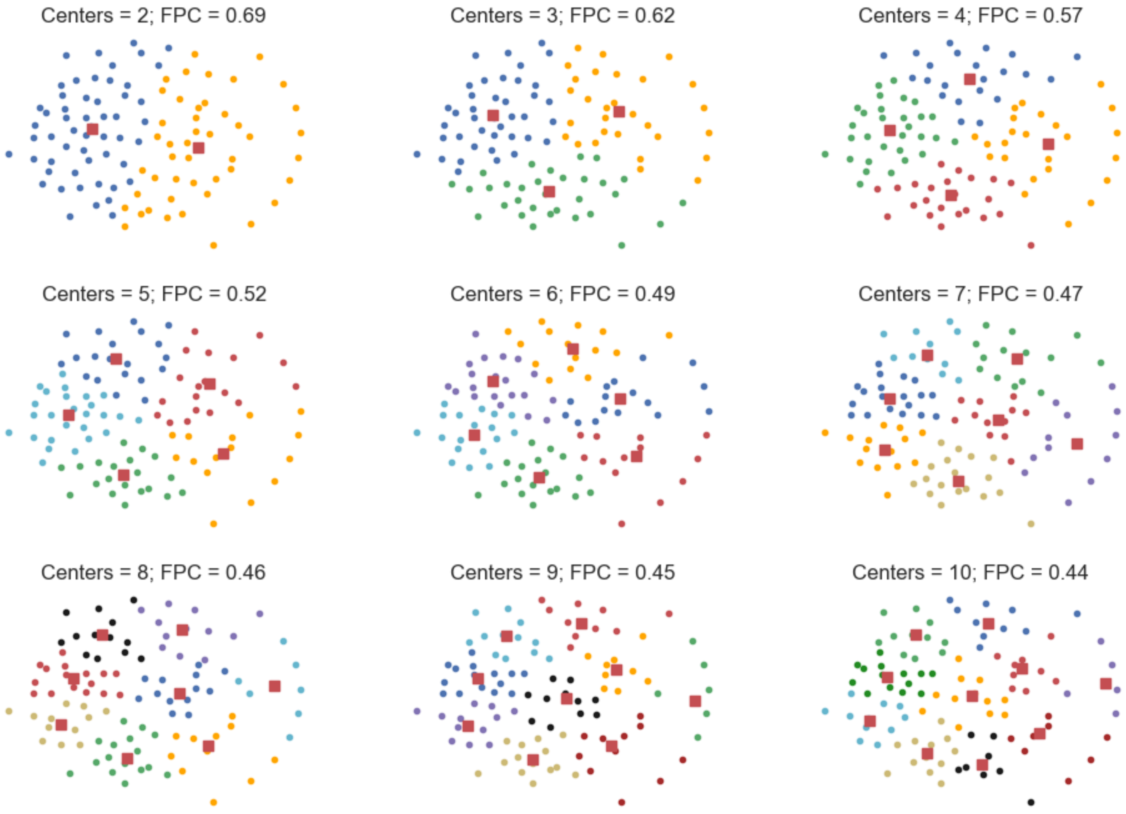


Fig 23: shows clusters formed by fuzzy c-means

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENTS

5.1 Application of algorithms:

Clustering is the extensively recognized type of unsupervised learning and is a remarkable device in numerous business and science fields in various applications. Hence, we outline the basic headings used in clustering.

•Finding documents that are identical: This component is frequently used when the user in a query result has spotted a "great" document and needs more – like this. Clustering can find documents that share a lot of similar words.

•Organizing Large Document Collections: This component is frequently used when the user in a query result has spotted a "great" document and needs more – like this. Clustering can find documents that share a lot of similar words. The test here is to compose these documents in a scientific categorization that is indistinguishable from the one that people would give enough opportunity to use as a perusing interface to the first document collections.

•Identical Content Detection: In most of the applications, a large number of documents need to find near-duplicates or the duplicates.

•Recommendation System: In this application, articles depending on the articles that the client has just perused are suggested to a user. Clustering the articles makes it continuously conceivable and greatly improves the quality.

•Search Optimization: Clustering makes it possible to significantly improve the quality and skill of web crawlers as the customer can first be contrasted with the bunches as opposed to legitimately contrasting them with the documents and the list items can also be effectively organized.

5.2 problems with clustering

Different issues among which few are considered in clustering are:

•Recent clustering systems do not satisfactorily call each of the preconditions.

•Things can be hazardous due to time complexity when dealing with an expansive dimensions and a vast data.

•The technique's viability is dependent on the meaning of "distance" (clustering based on separation).

•There is no undeniable measure, we should ' characterize ' it, which isn't always easy, particularly in the case of multi-dimensional spaces.

The main prerequisites that should be met by a clustering algorithm are:

•The capability to change in size and shape.

•Capability to form clusters

•Input parameters can be found by minimal knowledge of data

•noise and outliers dealing

•High dimensionality

•Interpretability

•Usability

5.3 Future Enhancements and conclusion

Document (or text) clustering is a subdivision of the broader data clustering field that takes the concepts from others, the areas of natural language processing (NLP), information retrieval (IR), and machine learning (ML). Clustering documents allows for expanding responses by including documents that are similar to those that a query has retrieved. Due to the increasing availability of electronic documents from the various variety of sources, Recently, document clustering studies have become more important. semi-structured and

Unstructured information resources include the governmental electronic, global web repositories, biological databases, news articles, digital libraries, chat rooms, electronic mail, online forums, and blogs repositories. Taking out information from these resources and proper categorization and discovering knowledge is therefore an important research area.

Machine Learning techniques, Natural Language Processing, and Data Mining work with one another to find patterns automatically in the documents. Text mining is the technique to deal with summary, trend and association analysis of retrieval, operations, classification (unsupervised, supervised and semi-supervised). Text mining's main goal is to allow the users to extract the relevant information from text resources. How the documented can be properly presented, annotated and classified, so the documents categorization consists of several challenges, proper annotation of documents, adequate representation of documents, a classification method that is appropriate to achieve generalization that is good.

As future work what we can do is we can use noise removal algorithms can be used to remove inconsistency in data.

References

[1] http://glaros.dtc.umn.edu/gkhome/fetch/papers/docclusterKDDTMW00.pdf

[2] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., “Chapter 4: A Survey of Text Clustering Algorithms,” in Mining Text Data. NewYork: Springer, 2012. [15] Schafer J. B., Frankowski D., Herlocker J., and Sen S., “Collaborative filtering recommender systems,” Lecture Notes In Computer Science, vol. 4321, p. 291, 2007.

[3]https://www.ijraset.com/fileserve.php?FID=10177

[4]https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1032&context=cs_faculty_pubs

[5] A. Huang, “Similarity measures for text document clustering,” In Proc. of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC, pp. 49—56, 2008.

[6] Pankaj Jajoo, “Document Clustering,” Masters’ Thesis, IIT Kharagpur, 2008

[7] MS. K.Mugunthadevi, MRS. S.C. Punitha, and Dr..M. Punithavalli, “Survey on Feature Selection in Document Clustering,” Int’l Journal on Computer Science and Engineering (IJCSE), vol. 3, No. 3, pp. 1240-1244, Mar 2011

[8] Minqiang Li and Liang Zhang, “Multinomial mixture model with feature selection for text clustering,” Journal of Knowledge-Based Systems, vol. 21, issue 7, pp. 704- 708, Oct. 2008

[9] Junjie Wu, Hui Xiong, and Jian Chen, “Towards understanding hierarchical clustering: A data distribution perspective,” Neurocomputing 72, pp. 2319–2330, 2009