# Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K

Sajid Naeem
Xinjiang Laboratory of Multi-Language Information Technology, College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, (830046), China

Aishan Wumaier
Xinjiang Laboratory of Multi-Language Information Technology, College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, (830046), China

## ABSTRACT
In the field of data mining, the approach of assigning a set of items to one similar class called cluster and the process termed as Clustering. Document clustering is one of the rapidly developing, research area for decades and considered a vital task for text mining due to exceptional expansion of document on cyberspace. It provides the opportunity to organize a large amount of scattered text, in meaningful clusters and laydown the foundation for smooth descriptive browsing and navigation systems. One of the more often useable partitioning algorithm is k-means, which is frequently use for text clustering due to its ability of converging to local optimum even though it is for enormous sparse matrix. Its objective is to make the distance of items or data-points belonging to same class as short as possible. This paper, exploring method of how a partitioned (K-mean) clustering works for text document clustering and particularly to explore one of the basic disadvantage of K-mean, which explain the true value of K. The true K value is understandable mostly while automatically selecting the suited value for k is a tough algorithmic problem. The true K exhibits to us how many cluster should make in our dataset but this K is often ambiguous there is no particular answer for this question while many variants for k-means are presented to estimate its value. Beside these variants, range of different probing techniques proposed by multiple researchers to conclude it. The study of this paper will explain how to apply some of these techniques for finding true value of K in a text dataset.

## Keywords
K-Means, Clustering, Unsupervised Learning, Pre-processing

## 1. INTRODUCTION
Due to fast development of the computer software, hardware and rapid advancement of internet technology, the massive amount of data has been collected and preserved to the databases. Average expansion of the data predicted by the researchers, doubles for each 20 months period. Despite the fact that human cannot use the raw data without obtaining knowledge appropriate for decision making, which presents its legitimate value. When the traditional human skill unable to analyze and manipulate data due to massiveness of its size, people use the computing technology to make the process automate and easy [1]. One of the emerging research activity is data mining in the field of computing Technique which define as to extract useful (non-trivial, implicit and formerly unknown) information or knowledge from extensive amount of data. Useful information accumulates by applying the data mining. Researchers afforded number of tools and techniques in the field of data mining to dig out the hidden pattern of data. These techniques are classifying in the field of classification, regression, outlier-analysis-association-rules and clustering [2]. Clustering is one of the renowned unsupervised approach, which works to divide the data into multiple related classes regardless of any prior knowledge about class definitions and used to discover groups or clusters of objects in a giant amount of data. It is one of the fundamental manner in data analysis, which applied in many fields like biology, psychology and economics. The objective is to group these objects such that objects in the same cluster should identical as much as possible and different from the object of the opposite groups or clusters as possible. Document clustering used to organize text documents, which are useful for information retrieval, data mining. Hierarchical and portioning are two types of clustering techniques. The highly used approach for achieving better quality clusters are Hierarchical techniques but they have quadratic time complexity and relatively slow. K-mean and its variants are the most extensively used partitioning techniques. Partitioning techniques have almost linear time complexity. K-mean clustering algorithm is a partitioning algorithm that grouped data into pre-defined no of clusters. It starts with random initialization of cluster centroids then assign data points to the closest (highly similar) centroids. The same operation is continuously repeating until the occurrence of termination criterion (either given number of iteration completed or clusters show no change after certain no of iteration) is met. It is centroid-based algorithm and data points assumed spherically scattered around the center of a cluster while cluster represented by a single center point and the points around. Although it was published 30 years before, but still widely used because of its simplicity, comprehensiveness and effectiveness. Arguably K-mean is the most conspicuous method of clustering that's why investigating its properties is not only area of interest to machine learning ,classification and data mining communities but also to the increase number of practitioners in bioinformatics, marketing research, customer management and other engineering application areas. Even though it is a simple and most usable algorithm but like other clustering algorithms, K-means required number of clusters to be specified in advance which is the most crucial and difficult problem to solve. Researchers have been proposed Number of methods to determine the number of clusters K for k-means algorithm. We will discuss few of them based on textual data to determine the number of clusters along with the evaluation of best and simple method to use for text documents:

A) Elbow method. B) Gap Statistic method. C) Cross-validation. D) The Silhouette method. E) K in text dataset.

F) Bayesian Information Criterion. G) Rule of thumb. H) Intra and Inter cluster distances [3-6].

## 2. CLUSTERING

Research workers have devised number of tools and techniques in field of data mining to get the inside pattern of the data. Then based on those inside patterns categorize the objects according to their similarity. This is an unsupervised approach used to find out groups or clusters of objects in enormous amount of data. Moreover, if the aim is to classify the set of documents then this process of categorization named as document clustering. It is a basic process of data analysis, which applied in many fields like biology, psychology and economics. The main objective is to divide these objects in different classes such that objects in the same class should alike as much as possible and dissimilar from the object in opposite class or clusters as much as possible. Document clustering used to organize text documents, which are beneficial for information retrieval, data mining [1-4]. It can divide the document clustering in two sub types a) hard clustering b) soft clustering these may also entitle disjoint and overlapping clustering respectively. If the document lies only in one cluster then it is a hard clustering while document lies in more than one clusters is reckoning as a soft clustering [34]. Some of the key clustering techniques are: a) Partitioning Clustering b) Hierarchical Clustering c) Grid Based Clustering d) Density-based Partitioning.

## 3. K-MEANS CLUSTERING

k-means clustering lies in partitioning clustering method most frequently used in datamining, the algorithm segregate N number of documents into K number of clusters while the value of K specified by users or even by making use of some the heuristic methods that discussed below to find the true K value for this division. Thus the true K will use to partitions our N documents in K different classes in which documents of same cluster must similar to each other and dissimilar from the other clusters or classes using some similarity constraints. The goal of K-means is to decrease the summation of square distance among data points and their respective cluster centers. The calculation steps required for k-means clustering method are follow as illustrated by [35].

Select the initial K cluster centers as:

$$a1(1), a2(1), a3(1) \dots ak(1)$$

Distribute the data {X} in K clusters at kth iteration using the relation below:

$$X \epsilon C_j\,(K)\ if\ \left\|x - a_j(k)\right\| < \|x - a_i(k)\|$$

For all 1,2,3,4, … , $K; i \neq j$; where Cj (k) represent the set of data points whose cluster centers is $a_j(k)$.

Calculate the new center $a_j(k+1), j = 1,2,3, \dots, K$ as the summation of the squared distances to the new cluster center from all points in $C_j(k)$ minimized. The part that work to minimize distance is simply the mean of $C_j(k)$. Thus the new cluster center is calculated as:

$$a_j(k+1) = \frac{1}{N}\sum_{x \in C_j(k)} x,\ \ j = 1,2,3, \dots, K$$

While the N j stand for the No. of samples in$C_j(k)$.

If $a_j(k+1) = a_j(k)$ for $j = 1,2,3, \dots, K$ then the algorithm become halt due to converged action, otherwise repeat step (b).

In this whole process it is clear that the final clustering results always effected by the initial seed and true value of K, but initial seeds and true value of K present in the data set required previous knowledge which is mostly impractical.

## 4. DOCUMENT PREPROCESSING AND REPRESENTATION

To divide a set of documents D in different classes or cluster is not a simple steady process because it need to follow a systematic procedure to cluster our data. The stages involved that highly effect the clustering results in the whole document clustering processes are discuses below

### 4.1 Preprocessing

Preprocessing is the primary step required to prepare the data readable for text mining process. It is useful for noise reduction in data and make the data clean. Because of preprocessing, the actual goal is to convert the original data in to machine understandable form. The process of preprocessing includes tokenization, filtering, stemming or lemmatization and stop-word removal.

#### *4.1.1Filtering:*

The words providing less value under vector models needs to remove before the actual calculation, filtering is the process to fulfill this task. Each document contains multiple words like punctuations, special-characters, stop-words and redundant words occur multiple time in each document as well as in multiple documents. It provides limited information to distinguish multiple documents, while documents also containing rare words that give no importance and needs to filter.

#### *4.1.2Stemming:*

The vital goal of stemming process is to change the words to its root (stem) words, which is highly language dependent process. The algorithm hugely take in consideration for stemming process in English language is publish in [44] and it was first introduce in [43]. It is a process that helping to raise the efficiency and decrease redundancy.

#### *4.1.3Lemmatization:*

Lemmatization is the procedure that emphasizes the lexical analysis of words and getting together number of inflected forms of words belonging to same family sorted by roots. Lemmatization can also define as a process of mapping nouns to its single form and verbs from to infinite tense. In the process to lemmatize the documents necessarily, it needs the POS definition of each word but POS is an error prone and very tedious job that is why stemming is always preferred practically instead of Lemmatization.

#### *4.1.4Stop-Word Removal:*

The perpetual occurring words like prepositions, articles, conjunctions: is, the, an, a, when, but etc. or the non-informative words and certain high frequency words are the "stop-words". While, Stop-word removal is the technique used to expel these words from the vocabulary because as a dimension of vector space they do not give any meaning and considered less significant. Stop-word removal process helps in performance and highly influences the complete clustering process [36-45].

### 4.2 Representation

Before clustering the documents in to groups, the important requirement is to convert our corpus in machine recognizable form vectors, abbreviate as VSM (Vector Space Model) written as $Z = \{x_1, x_2, x_3, x_4, \dots, x_n\}$. The term "$x_1$" in Z represents the feature vector $d$ for each document while $d = \{w_1, w_2, w_3, \dots w_n\}$. The $w_i$ is a term weight representation of term $t_i$ in a document which manifests the significance of each term in a document. For SVM vectorization of corpus it

is highly dependent and more easily understandable to make use of the methodology of TFIDF (Term Frequency Inverse Document Frequency) is excessively used one, which calculate the importance for each term of a document within corpus. The TFIDF can calculated as follow:

$$W_{ji} = tf_{ji} * idf_{ji} = tf * log_2(n/df_{ji})$$

The $tf_{ji}$ represents the frequency of term $i$ in a document $j$. While $df_{ji}$ indicate the documents number in which term $i$ has contained, and $n$ is the total number of document in a corpus. The calculation of term weight under this process deliberate the frequency of appearance for a term in a document as well as in the entire corpus. If a term occurred, more frequently in multiple documents is declare as a stop-word. TFIDF eradicate those words because the TFIDF score for stop-words turn to zero or near to zero [36] [38] [46-47].

## 5. PREVIOUS WORK

Himanshu Gupta et al., [3] prescribed a technique for document clustering to find the K value for K-means using SVD (Singular Vector Decomposition) and the clustering improved by features voting that enable the algorithm comparatively much faster.

Trupti M.Kodinariya et al., [6] elaborated six different approaches for the selection of K value for K-Mean clustering algorithm in a dataset. He concluded that clusters are in a viewing eye and analyzed the situation when clusters, though not definitely typical, are in data.

Ahmed Shafeeq B M at al., [7] represented an approach to modified algorithm of k-mean for fixing the required optimal cluster numbers with improving the cluster quality. K-means algorithm required (k) number of cluster from the user. Practically it is very difficult to fix the cluster numbers in advance. Proposed method work in both cases for known and unknown clusters. User have the choice to fix the k number or input required minimum number of clusters.

The methodology elaborated by Azhar Rauf et al., [8] found the initial centroid instead of random selection, which causes improvement in two aspects elapsed time improvement along with decreased iteration number.

The process presented by Youguo Li et al., [9] combined the traditional K-Means algorithm with largest minimum distance algorithm that overcome the deficiencies found in the traditional way which helps to determine the starting focal point.

Siddheswar Ray et al., [10] narrated a way to determine number of clusters automatically using the method of intra, inter clustering distance measures. The primary method contains all the segmented images from two clusters up to K clusters, where K max displays a highest cluster number.

Madhu Yedla et al., [11] suggested a method for reducing the time complexity and finding the suitable initial centroids. He also proposed an efficient way to determine that data points assigned to its suitable clusters.

K. A. Abdul Nazeer et al., [12] describe a mechanism in which he enhanced the traditional simple k-mean, which can efficiently assign the data points to clusters along with the systematic method for finding the initial centroids.

Madhuri A. Dalal et al., [13] designed an improved algorithm for better starting points to start K-Means. Initial starting point effected the local minimum that is why selecting the best starting point make the algorithm fast and provide best resulting clusters.

Deepika Khurana et al., [14] Presented the new dynamic way based on the silhouette validity index and give a solution for picking the initial cluster centroids. Besides the fact to run the algorithm for various k values , user only need to give the initial value of k to input and then algorithm determine the adequate k value for the given dataset. From experimental results, it has been clearly testified that the proposed system enhance the overall computation time plus initial selection of centers.

Chunfei Zhang et al., [15] elaborated the means for patronizing K-means algorithm based on k values determination and initial focal points. Experimental result from simulation make it obvious that the final clustering ends are more efficient and precise because of enhancement the clustering algorithm. The improved algorithm not only avoid the noise impact in the dataset but also more poised in a clustering process.

The new clustering algorithm advised by Nidhi Gupta et al., [16] takeout the shortcoming of K-Means algorithm. Her proposed approach does not need to define the K value, i.e. required cluster number.

Pallavi Purohit et al., [17] expressed the K-Mea in a new algorithmic form, according to the users requirement firstly it can determine the initial centroids and then give suitable, uplifted and good cluster beyond scarifying certainty to the users. The displayed description takeover the weaknesses of existing K-Mean. It also decreases the mean square errors and enhance the quality of clustering. It also provokes stable clusters to boost accuracy.

Sharddha Shukla et al., [18] wrote a review paper on k-mean on various efforts made by different researchers to overcome the shortcoming found in the traditional k-means. She also discussed the application of the k-means and concluded that the most of work done is to improving the accuracy and efficiency of clusters, while deciding the k values remains a challenging problem that's why it's still provide a scope to researchers for future enhancement.

D T Pham et al., [19] reviewed the existing methods of K value determination and elaborated the factors that consequence the selection. The suggested new method aided the selection with the analysis of outcome for the proposed measure against different data sets to resolve the issue of K value. The propose method suggested numerous values of K to users for cases when multiple clustering conclusions could be get with distinct levels of detail. However, the recommended approach is computationally expensive for huge datasets because it desires various application of K-means before it propose a suitable value for K.

Greg Hamerly et al., [20] proposed and improved clustering algorithm for learning K value "G-mean" based on statically test, which explain that the subset of data follows Gaussian distribution. The proposed algorithm runs in a hierarchical fission with increasing K until the test accept the hypothesis the result is Gaussian after assigning data to each center of K-mean.

Jian Di et al., [21] proposed an improved bisecting K-means algorithm for automatic determination of K value and optimized the cluster center. In proposed method, the initial cluster centers are selecting by using the point density and distance function. The function of K values detection done by

inter cluster difference and intra cluster similarity. The algorithms overcome the effect of noise points, outliers and enhance the efficiency of clustering outcome.

# 6. METHODS:

## 6.1 Elbow Method

Determining the proper number of cluster is one of the basic drawback in k-means algorithm. The correct choice of k is often ambiguous; to solve this problem different practitioner used different approaches Elbow method is also one of them to find the right number of K for K-mean algorithm. It is a visual and oldest approach, which is more often used approach toward this problem for finding true value of K. It is a process of assessing the ratio of variance outcome as a function of the number of cluster. This method based on logic if make update in k value for clustering the same dataset one by one will not provide much better modeling for dataset. Then plot the examined variance after increasing number of clusters plotted against the number of clusters. The basic notion is to initiate K=2 and keep incrementing it in each step by 1 and for each value of k evaluate the sum of squared errors (SSE) or distortion and clusters that anticipated with training. Starting Clusters will figure enough information while at a particular spot the marginal gain dramatically descends and allocate an angle graph. At that value, when the cost drop down dramatically after that the graph increase horizontally and further increase in number of clusters it gets plateau. Henceforth when plot a line chart of the SSE (measures the compactness of the clustering and our desire it to be as nominal as possible) for respective value of k. If the line chart visible like an arm, then the "elbow" on the arm is the value of k that is the pertinent k. The concept is that search for a meagre SSE and SSE tends to decrease toward 0 as escalate k (the SSE is 0 when k is synchronize to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). The justification is after this, if make increase in the clusters in numbers, the new cluster will very close to some of the existing. Therefore, our aim is to acquire a small value of k that still has a minor SSE, and the elbow usually represents where have to start. Despite this, there is still a problem with elbow method because it cannot be clear always and sometime gives multiple elbow for some datasets or even show no elbow. So in the situation if have not clear elbow then as alternate go for use of some other method to know the true value of k or to check whether our dataset is suitable for clustering or not. If the resultant elbow is vague elbow then approach of 'Kneedle' which is a method explain by Ville Satopa [24], can helps to represent the best curve point knee for elbow method because of dealing with a mathematical concept of curvature. The python SK-learn provide a library for finding the knee point in elbow. Even though it give mostly good results and also simple to understand and implement, but its need of biased judgment for pointing out where the actual elbow is located, and as [25] prove that the approach can easily fail. Also for high dimensions in our experiment, the elbow is not clear to select the true k.

*Steps for Elbow Method:*

- K =2 K start from 2
- Increase the k value by one 1
- Measure the SSE or distortion
- The point at which the cost drops dramatically
- Select that point as a true K

## 6.2 Gap Statistic Method

The method presented by Tibshirani et al for estimating number of cluster in a dataset which is compatible with any type of clustering such as hierarchical and bisecting etc. This method works to compare within cluster dispersion change$Wk$. The error rate of $Wk$ decrease with the increase of $K$and show an 'elbow' shape to recognize the true $K$ value as mentioned before that give graphical view. While gap statistic works to standardize $log(Wk)$ graph using null reference distribution of data for comparison and select smallest $K$ as a true $K$ value, which maximize the $Gap(k)$[29].

*Steps for Gap Statistics:*

1) Cluster the dataset using range of k values from $K = 1 \dots Kmax$ and calculate the $Wk$ for each $K$ value.
2) Make a reference datasets B using the two methods mentioned in original paper and also cluster it them with range of $K = 1 \dots Kmax$ and calculate the predicted gap statistics.

$Gap(k) = (1/B) \sum_{b=1}^{B} logW_{kb}^{*} - logW_k$.

3) Now with $\bar{w} = (1/B) \sum_b logW_{kb}^{*}$ ,Compute the standard deviation

$sd(k) = [(1/B) \sum_b (logW_{kb}^{*} - \bar{w})^2]^{1/2}$ and define $s_k = \sqrt{1 + 1/B}\, sd(k)$.

4) After all choose the true clusters value K through this process such that smallest K having $Gap(k) \geq Gap(k+1) - s_{k+1}$ [29].

## 6.3 Cross-Validation

Cross-validation is another mechanism for scrutinizing the number of clusters devised by Smyth [22], based on cluster stability to figure out the true k. This method breach the data into two or more parts it means, the dataset classifieds in to v parts. One portion of dataset used for clustering and the other parts used for validation. The concept that explain clustering stability is that a "good" algorithm contribute for repeatedly producing similar clustering on data derive from same origin. In other word, the algorithm is durable according to input randomization this metric is called v-fold cross validation. The true k value that explaining the number of classes in a dataset is really nuisance criterion of clustering model could estimate through this possible method. In broad context, simply applying this method to different cluster numbers in k-mean and examined observations average distance (in cross-validation samples) from centers of their cluster. The [25] explain the cross validation for estimating the values of k in a large dataset and of high dimensionality they provide a novel metric which works better for their dataset than ordinary method. They use cross validation error of each k, and choose the K having least cross validation error.

*Steps for Cross-Validation:*

- Select a folds number.
- Select the K (clusters number) value in range that should try for clustering.
- Then analyzed the variance details in percentage among the successive values of K that have used.
- Specify a precedent to that helps in diction of elbow automatically.

## 6.4 The Silhouette Method

The average silhouette of the data is another lucrative and precise way for determining the natural number of clusters. According to [26] the silhouette method consumed lot of time due to calculation of distance and took more CPU time. Along with this, describe a new method for silhouette to minimize the computation time with reducing addition operations amount during distance calculation, which has experimentally proven that about 50% CPU time gained. It is also a measure that help in concluding clustering legitimacy and selecting the optimal K value to divide a ratio scale data in distinct classes [27]. For true K the preferable number of clusters whose silhouette value predicted large enough. Still it is too much complex because the value variation of silhouette of two clusters A and B are very imperceptible and will use to plot it for definite understanding. The silhouette of a data instance is a measure of how closely it is matches to data point within its own cluster and how loosely it is match to data of the neighboring cluster. The silhouette values occurs in the scope from -1 to 1. If the silhouette width value for an entity is about to zero, likewise the entity could be appoint to another cluster as well. If the silhouette width value is near to 1, it means that the entity is misclassified. If all the silhouette width values are intimate to 1, it means that the set I is well clustered. The average silhouette of individual entities can characterize a clustering. The largest average silhouette width, over different K, indicates the best number of clusters.

## 6.5 For Text Dataset

In a text database, a document archive defined by a term $D$ matrix (of magnitude m by n, m: number of documents, n: number of terms). Number of clusters can grossly anticipated by the following formula $(MN/T)$, $T$ is the number of non-zero entries in $D$. Note that in $D$ each row and each column essentially possessed at least one non-zero element. Beside this, the actual value of k in text dataset could be find using method presented in [28] "cover coefficient method" which have proven more suitable and reliable.

## 6.6 Bayesian Information Criterion

The method of Bayesian information criterion introduce by Gideon E. Schwarz [30]. Even though it can also say "Shwarz" criterion, which is not particularly use for model selection only, but also gives help in determining the number of clusters [32]. Bayesian Information Criterion method choose a model that resulting with the maximization Of BIC and focus on calculating the likelihood with the range of clusters K. The K which maximize likelihood should be selected but meanwhile carefully keep focused on problem if in case continuously increase the number of clusters can maximize the likelihood up to the extent in which our data may clustered each data point per group then every data point belong to its own cluster. As it has known from SSE calculation that increase in K cause decrease in SSE and when the number of data points and K become equal the SSE reduce to zero. That is the reason primary need is to find and choose the K, which is not entirely depends upon likelihood [6]. The selected true K for our dataset is, which return the maximum BIC score with a smallest K assumed for k-means clustering by Pelleg et al. for applying the x-mean with some extra acceleration feature and extend k-means algorithm to its new variant for x-means clustering algorithm. The idea to choose the best K for clustering is to investigate the model for the type spherical Gaussians as assumed by k-means [31-33].

This BIC could be mathematically compute using the following formula [33]:

$$BIC(M_k) = \hat{l}k(D) - \frac{p_k}{2} log(n)$$

In the above equation $\hat{l}k(D)$ represent the maximum log-likelihood for data regarding model $M_k$, while $p_k = k(d + 1)$are number of parameters for $MK$. The log-likelihood for overall data points belonging to every centroid is the measure of summation to log-likelihood of individual centroids while the $N$ used to represent total number of points associated to all centroids in consideration. [31]

## 6.7 Rule of Thumb

This is a heuristic method presented in [6] which is used for K calculation in a dataset but it have no mathematical proof but still preferred by some researchers on the internet due to simplicity in calculation. It may work some time in some condition but easy to estimate the K for any type of dataset.

$$\boldsymbol{K \cong \sqrt{n/2}}$$

In the above equation, $n$ represent the number of objects known as data point.

## 6.8 Intra and Inter Cluster Distances

In clustering techniques, the main aim to get a condensed and well-separated resultant clustering after applying the clustering algorithm. As a result, it is urge to minimize the intra-cluster (with-in cluster scatter distance) distance and maximize inter-cluster (between cluster separations) distance [48]. While the aim of K-means algorithm is to decrease the summation of square distance among data points and their respective cluster centers [35]. Therefore, compactness of clusters in K-means can be measure by using the average of intra-cluster distance measurement. Inter-cluster distance is the distance between centers shows clusters separation, which must need to maximize. The combination of these two measures may help to calculate clustering excellence as defined by [10].

Validity = intra/inter

In this equation, basic aim is to decrease the validity. The K-means should run from $K = 2$ up to $Kn$. While $Kn$ represents the maximum value of $K$. The $K$ which calculate minimum validity, will be treated as a true value of $K$ [10]. In the experiments K-means is initialized with "K-mean++" initialization provided by "Sklearn" which select the best initial centroid selection for our experiment.

# 7. EXPERIMENTAL ENVIRONMENT AND RESULTS

The data used for our experiment archived from UCI (University of California, Irvine) a machine-learning repository developed in 1987 by David Aha with his graduate students [49]. The UCI is extensively used repository for research and analysis purpose providing freely available corpus collections. These databases are freely available along with their descriptions. Beside to use freely available corpus, even it could also possible to develop our own crawler to collect text from the internet. The system description used in our experiments are Computer CPU Intel ® Core I7; 8 GB RAM with Windows 10, 64-bit OS. Programming language Python.3.6 with anaconda environment and IDE Pycharm.

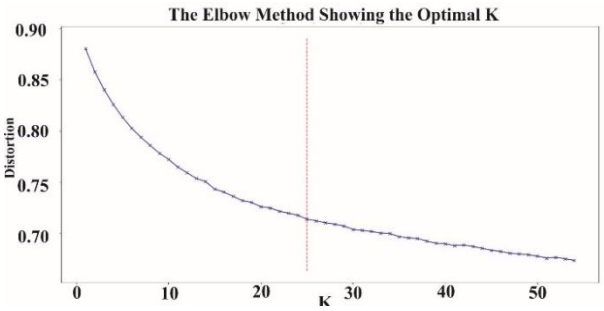


**Fig 1: Elbow Graph**

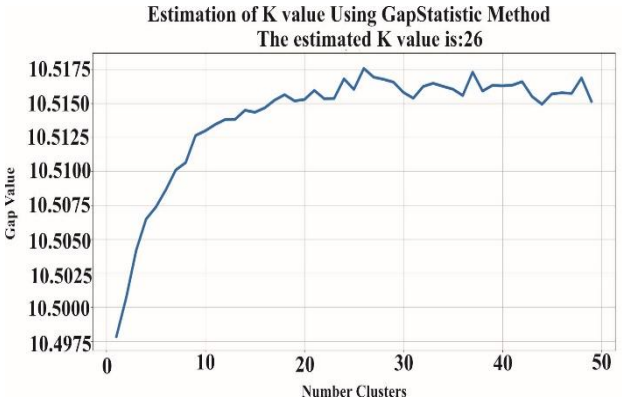


**Fig 2: GapStatistic**

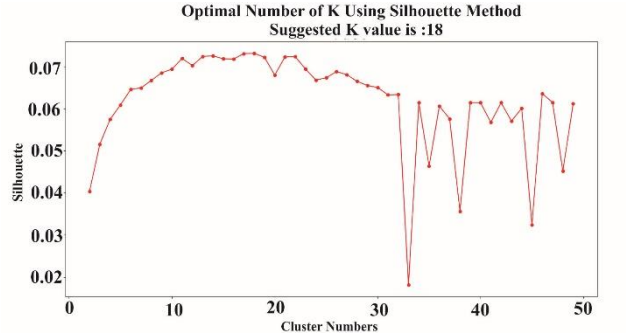


**Fig 3: Silhouette**

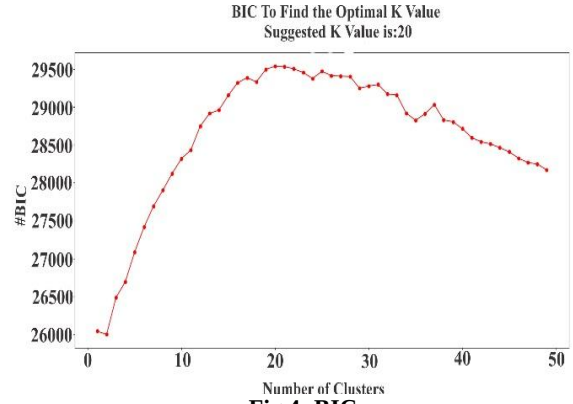


**Fig 4: BIC**

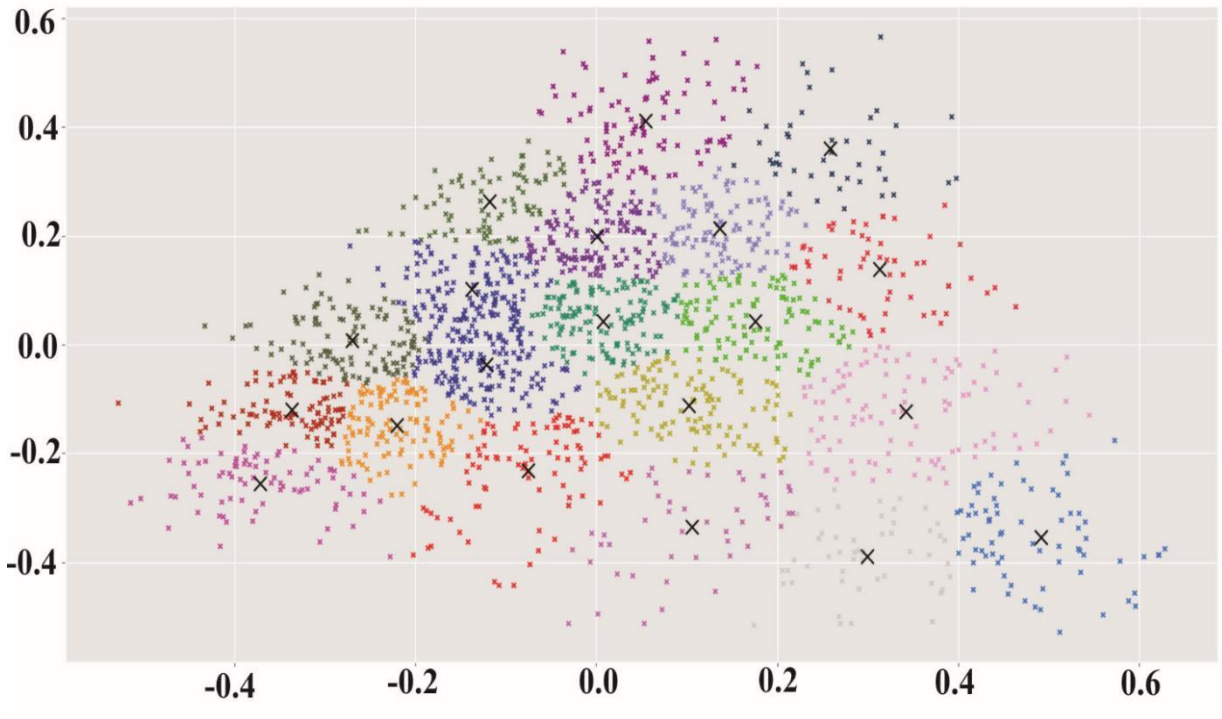

**Fig 5: Final Data Representation in Two**

**Dimension Graph with K=20**

**Table 1: Dataset Description:**

| **Dataset** | Number of Documents | Number of Tokens Before Preprocessing | Number of Tokens After Preprocessing |
|---|---|---|---|
| **Division** | 2000 | 1348129 | 467153 |

**Table 2: Experiment results based on each method:**

| **Methods** | Elbow | Gap Statistic method | The Silhouette method | Bayesian Information Criterion | Rule of thumb | Intra and Inter cluster distances |
|---|---|---|---|---|---|---|
| **True KValue** | 25 | 26 | 18 | 20 | 30 | 21 |

# 8. CONCLUSION

Clustering is one of the popular unsupervised approach as for as it have notorious drawbacks and substitutes.
Still used in many fields (data mining, classification and machine learning etc.) due to its simplicity and speed. It provide the opportunity to organize a large amount of scatter text in meaningful clusters while document clustering is a key unsupervised process for grouping massive freely available archives on the internet and it remains the field of interest for many researchers since decades. K-means have multiple variants and many researches already been done for its accuracy and efficiency while the automatic K selection is always remains ambiguous and challenging. There is no proper solution to the problem true K value estimation, which is trustworthy in each dimensions but there is some heuristic rules used to determine the value of K. Due to limits of the paper all the dimension are not explained for these techniques regarding k-means document clustering. In the final conclusion of experiments it is experienced clearly that beside the ambiguity in the True K value k-means is also highly effected by initial centroid selection, outliers and noise, preprocessing and high dimensionality (large spare data) because in document clustering the final clustering results is highly impacted by preprocessing step that's . Many researchers such as in [50] it has been detailed that better to use dimensionality reduction despite directly apply k-means in high dimension data and can use PCA (principle component analysis) for dimensionality reduction. In our experiment the combination of last method (Intra and Inter cluster distances) with k-mean++ to select the initial centroid carefully and it is exponentially faster and this technique may use in many dimensions to get to the actual result. In this paper, it has described that the most critical issue which remains an open question for many researcher in recent years we will focus on some methods to find initial cluster centroids (initial seed selection) selection.

# 9. REFERENCES

[1] Shraddha, S.et al. 2014, "A Review ON K-means DATA Clustering APPROACH" International Journal of Information and Computation Technology.

[2] Y,S,Patail , M.B. Vaidya 2012, "A Technical survey on Clustering Analysis in Data mining" International Journal of Emerging Technology and Advanced Engineering.

[3] Himanshu Gupta, Dr.Rajeev Srivastav 2014, "K-means Based Document Clustering with Automatic 'K' Selection and Cluster Refinement" International Journal of Computer Science and Mobile Applications.

[4] Greg Hamerly and Charles Elkan 2003, "Learning the k in k- means" In Neural Information Processing System, MIT Press.

[5] Anil K Jian 2009, "Data Clustering: 50 Years beyond K-Means, Pattern Recognition Letters".

[6] Trupti M.Kodinariya and Dr.Prashant R. Makwana 2013, "Review on determining number of cluster in K-Mean Clustering" International Journal of Advance Research in Computer Science and Management Studies.

[7] Ahmad Shafeeq B M, Hareesha K S 2012, "Dynamic Clustering of Data with Modified K-Means Algorithm" International Conference on information and Computer Networks, Vol. 27.

[8] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro, Huma Javed 2012, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity" Middle-East Journal of Scientific Research, pp. 959-963.

[9] Youguo Li, Haiyan Wu 2012, "A Clustering Method Based on K-Means Clustering Algorithm" International Conference on Solid State Devices and Materials Science, pp. 1104-1109.

[10] Siddheswar Ray, Rose H.Turi, 1998, "Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation".

[11] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa 2010, "Enhanced K-Means Clustering Algorithm with Improved Initial Center" International Journal of Computer Science and Information Technologies, Vol. 1, pp. 121-125.

[12] K. A. Abdul Nazeer, M.P. Sebastian July 1-3, 2009, "Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm" Proceedings of the World Congress on Engineering, London, UK.

[13] Madhuri A. Dalal, Nareshkumar D. Harale, Umesh L.Kulkarni July 2011, "An Iterative Improved K-Means Clustering" ACEEE International Journal on Network Security, Vol. 02.

[14] Deepika Khurana, Dr. M.P.S Bhatia May-June 2013, "Dynamic Approach to K-Means Clustering Algorithm" International Journal of Computer Engineering & Technology and Research, Issue 3, Vol. 4, pp. 204-219.

[15] Chunfei Zhang, ZhiyiFang 2013, "An Improved K-Means Clustering Algorithm" Journal of Information & Computational Science.

[16] Nidhi Gupta, R.L. Ujjwal 2013, "An Efficient Incremental Clustering Algorithm" World of Computer Science and Information Technology Journal, Vol. 3.

[17] Pallavi Purohit, Ritesh Joshi March 2013, "A New Efficient Approach towards K-Means Clustering Algorithm" International Journal of the Computer Applications.

[18] Shardda Shukla, Naganna S 2014, "A Review ON K-mean DATA Clustering APPORACH" International Journal of Information and Computation Technology.

[19] D T Pham,S S Dimov, C D Nguyen 2004, "Selection of K in K-means clustering" Manufacturing Engineering Centre, Cardiff University, Cardiff, UK.

[20] Greg Hamerly Charles Elkan 2004, "Learning the K in K-means" Advances in neural information processing systems, Vol. 16.

[21] Jian Di Xinyue Gou 2017, "Bisecting K-means Algorithm Based on K-valued Self-determining and Clustering Center Optimization" Journal of Computers.

[22] Smyth, P. 1996, "Clustering using Monte Carlo Cross Validation" In Proc.2nd Intl. Conf. Knowl. Discovery and Data Mining (KDD-96), Portland.

[23] J. B. MacQueen 1967, "Some Methods for classification and Analysis of Multivariate Observations" Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press.

[24] Ville Satopa et al. "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior" International Computer Science Institute, Berkeley, CA.

[25] Wei Fu and Patrick O. Perry February 10, 2017, "Estimating the number of clusters using cross-validation" Stern School of Business, New York University.

[26] Moh'd Belal Al- Zoubi and Mohammad al Rawi, "An Efficient Approach for Computing Silhouette Coefficients" Department of Computer Information Systems, University of Jordan, Amman 11942, Jordan.

[27] Tippaya Thinsungnoena et al 2015, "The Clustering Validity with Silhouette and Sum of Squared Errors" The 3rd International Conference on Industrial Application Engineering (ICIAE2015).

[28] Fazli Can and E. A.Ozkarahan, December 1990, "Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases" ACM Transactions on Database Systems, Vol. 15, No. 4, pp. 483-517.

[29] Robert Tibshirani et al. 2001, "Estimating the number of cluster in a dataset via the gap" Royal Statistical Society, Standford University, USA. Part 2, pp. 411-423.

[30] Schwarz, Gideon E. March 1978, "Estimating the dimension of a model". Annals of Statistics Vol. 6, No. 2, pp. 461–464.

[31] D. Pelleg and A. Moore July 2000, "X-means: Extending k-means with efficient estimation of the number of clusters" Proceedings of the Seventeenth International Conference on Machine Learning, pp 727-734

[32] C. Fraley and A. E. Raftery 1998, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis" The Computer Journal, Department of Statistics University of Washington USA, Vol. 41, No. 8, pp. 578-588.

[33] R. E. Kass and L. Wasserman 1995, "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," Journal of the American Statistical Association, pp. 928– 934.

[34] Chun-ling Chen,S.C. Tseng and Tyne Liang Nov. 2010, "An integration of Word Net and Fuzzy association rule mining for multi-label document clustering" Data and Knowledge Engineering, pp. 1208-1226.

[35] J.T. Tou and R.C. Gonzalez 1974, "Pattern Recognition Principles" Massachusetts: Addison-Wesley.

[36] Mehdi Allahyari.et al. August 2017, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques" In Proceedings of KDD Bigdas, Halifax, Canada, 13 pages.

[37] Twinkle Svadas, Jasmin Jha June 2015, "Document Cluster Mining on Text Documents" International Journal of Computer Science and Mobile Computing Vol.4, pg.778-782.

[38] Neepa Shah, Sunita Mahajan October 2012, "Document Clustering: A Detailed Review" International Journal of Applied Information Systems (IJAIS) Vol. 4.

[39] Abdennour Mohamed Jalil, Imad Hafidi et al. 2016, "Comparitive Study of Clustering Algorithms in Text Mining Context" International Journal of Interactive Multimedia and Artificial Intelligence Vol. 3, No. 7.

[40] Jonathan J Webster and Chunyu Kit 1992, "Tokenization as the initial phase in NLP" In Proceedings of the 14th conference on Computational linguistics Vol. 4, pp. 1106–1110.

[41] Hassan Saif et al 2014 "On stopwords filtering and data sparsity for sentiment analysis of twitter" School of Engineering and Applied Science, Aston University, UK.

[42] Catarina Silva and Bernardete Ribeiro 2003, "The importance of stop word removal on recall values in text categorization" Proceedings of the International Joint Conference on Neural Networks IEEE, Vol. 3, pp. 1661–1666.

[43] Julie B Lovins 1968, "Development of a stemming algorithm. MIT Information Processing Group" Electronic Systems Laboratory.

[44] Martin F Porter 1980, "An algorithm for suffix stripping" Program: Electronic Library and information system, pp. 130–137.

[45] David A Hull et al. 1996, "Stemming algorithms: A case study for detailed evaluation" JASIS, pp. 70–84.

[46] Everitt, B., 1980. "Cluster Analysis" 2nd Edition. Halsted Press, New York

[47] M. Meila, and D.Hackerman 1998, "An Experimental Comparison of Several Clustering and Initialization Method" Microsoft Research Redmond, WA.

[48] D.L Davies and D.W. Bouldin 1979. "A cluster separation measure" IEEE Trans. Pattern Anal. Machine Intell. Vol.1, pp. 224-227.

[49] The corpus taken from UCI repository "https://archive.ics.uci.edu/ml/datasets.html".

[50] Sanjoy Dasgupta 2000, "Experiments with random projection" In Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference San Francisco, CA. Morgan Kaufmann Publishers, pp. 143-151.