“DOCUMENT CLUSTERING BY K MEANS ”

BY ARGHYA MAZUMDAR

08/06/1996

ABSTRACT

There is an important requirement to classify unknown documents based upon their content for similarity checking. This project aims to take a dataset of research papers and classify them into subtopics by using the K Means Algorithm so that researchers can easily follow work which is closely related to their own field.

INTRODUCTION

The dataset consisted of research papers which were accepted by the University Of California from 2012-2016. Each of the documents were tokenized and stemmed and term frequencies (tf) of the significant words were found out. The information retrieval was done by using tf-idf weight vector matrix which was then passed into K Means Algorithm to form cluster of words. The clusters were then pictorially represented by multidimensional scaling and the pandas library present in the Python language.

METHODS

The two lists are first segregated from the columns of the dataset which include the title and the abstract of the research papers. Then the abstract is stemmed and tokenized. NLTK's list of English stop words which are words like "a", "the", or "in" which don't convey significant meaning are removed from the abstract to improve classification accuracy. The Snowball Stemmer which is a part of NLTK breaks a word down into its root(stemming). Then this cleared up abstract list is passed to the tf-idf vectorizer which assigns a weight vector to every research paper respectively. The **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

The list Dict is defined as 1 - the cosine similarity of each document. Cosine similarity is measured against the tf-idf matrix and can be used to generate a measure of similarity between each document and the other documents in the corpus (each synopsis among the synopses). Subtracting it from 1 provides cosine distance which is used for plotting on a Euclidean (2-dimensional) plane.

The tf-idf sparse matrix is then passed to the K Means Clustering Algorithm with the value of K chosen as per described below. After the clusters are formed, they are pictorially represented using MDS and then displayed by using the matplotlib.

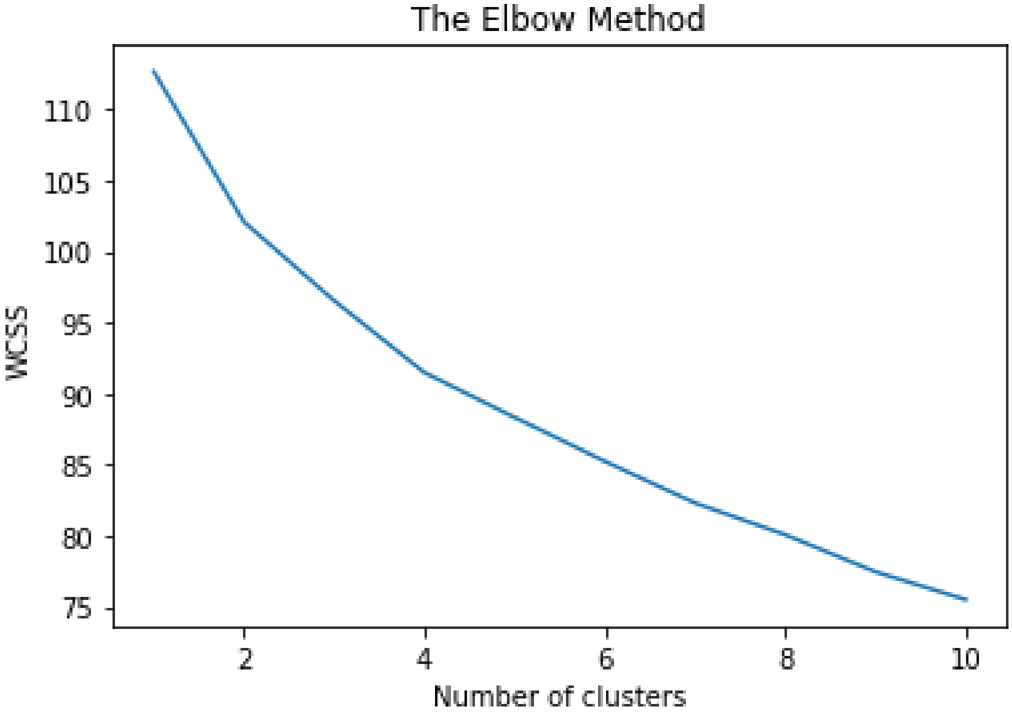
CHOOSING OPTIMUM VALUE OF K

The value of K was tried to be determined first by the “elbow” method.

Elbow Method:

The elbow method looks at the percentage of variance explained as a function of the number of clusters. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information, but at some point; the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test.

The method indeed proved to be ambiguous as no clear cut “elbow” was seen in the graph.



Here WCSS(Within-Cluster-Sum-of-Square) is the Implicit **objective function in k-Means** measures sum of distances of observations from their cluster centroids, called (WCSS). This is computed as

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

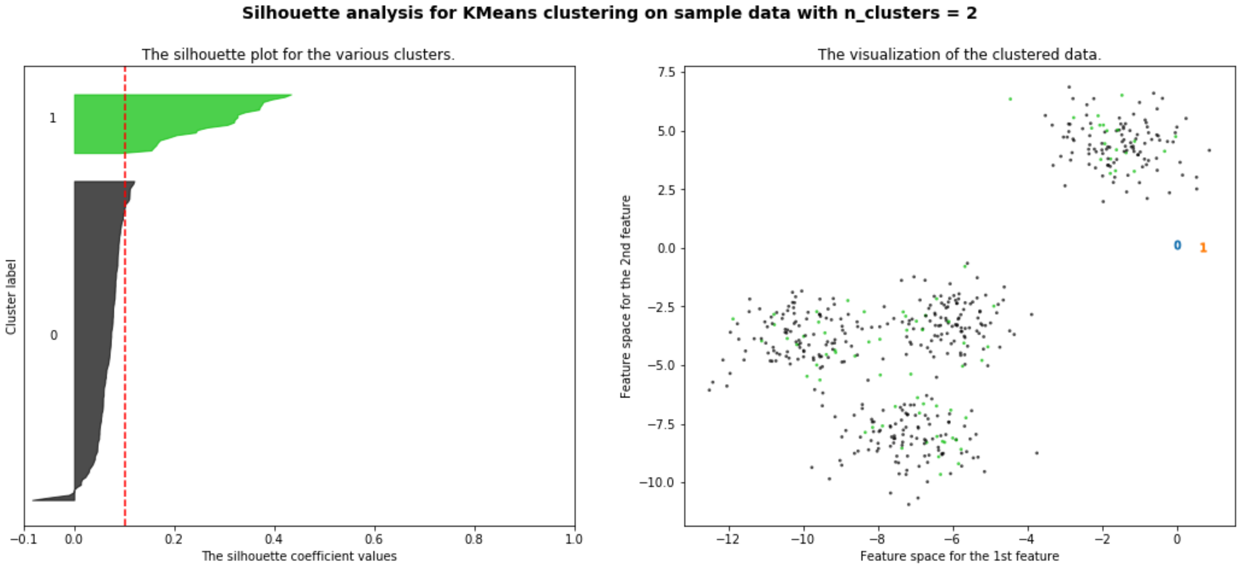
where Y_i is centroid for observation X_i. By definition, this is geared towards maximizing number of clusters, and in limiting case each data point becomes its own cluster centroid.

Cluster Quality using Silhouette Coefficient

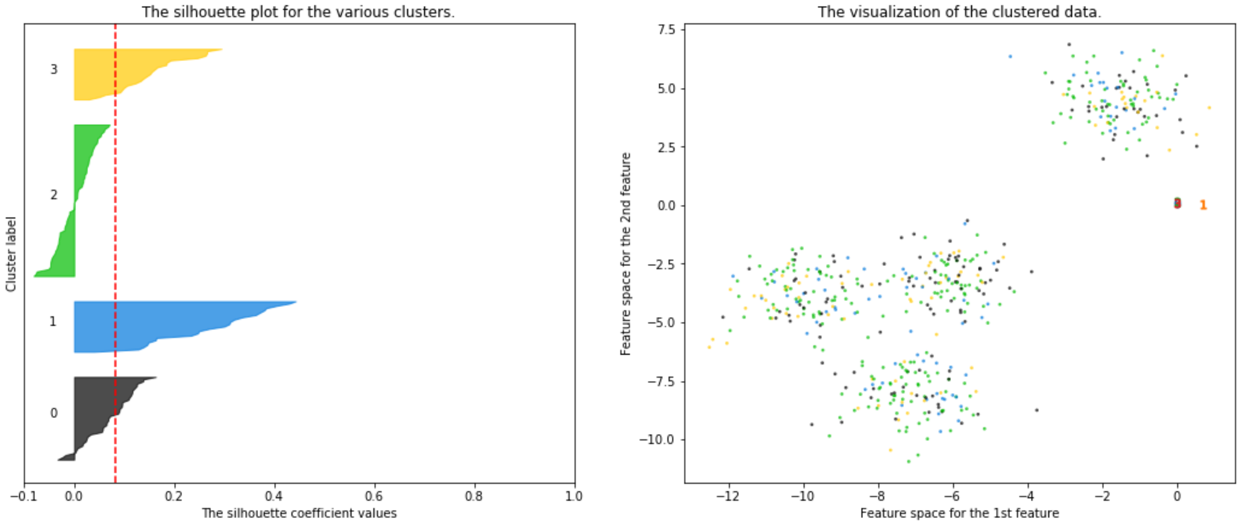
Silhouette-Coefficient of observation *i* is calculated as

$$s_i = \frac{b - a}{\max{(a, b)}}$$

where *a* is average distance to all other observations within same cluster as that of observation *i* while *b* is minimum of average distance to all other observations from all other clusters. Silhouette coefficient of clustering result is average of s_i for all observations *i*. This metric is between +1 representing best clustering and -1 representing worst clustering. This method gave appreciable results as shown below.



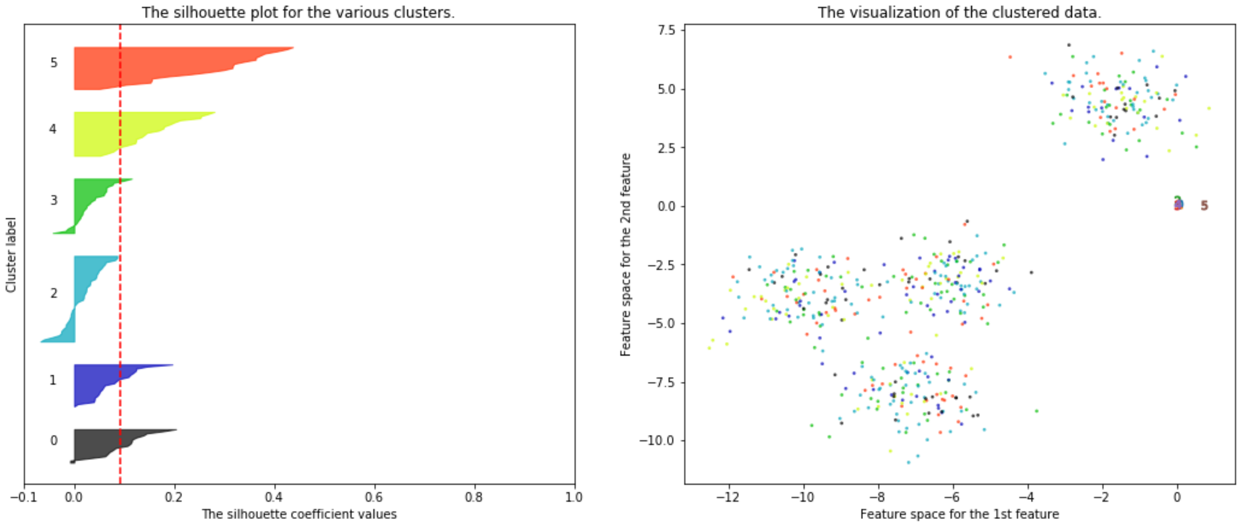
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



The silhouette plot for the various clusters.

The visualization of the clustered data.

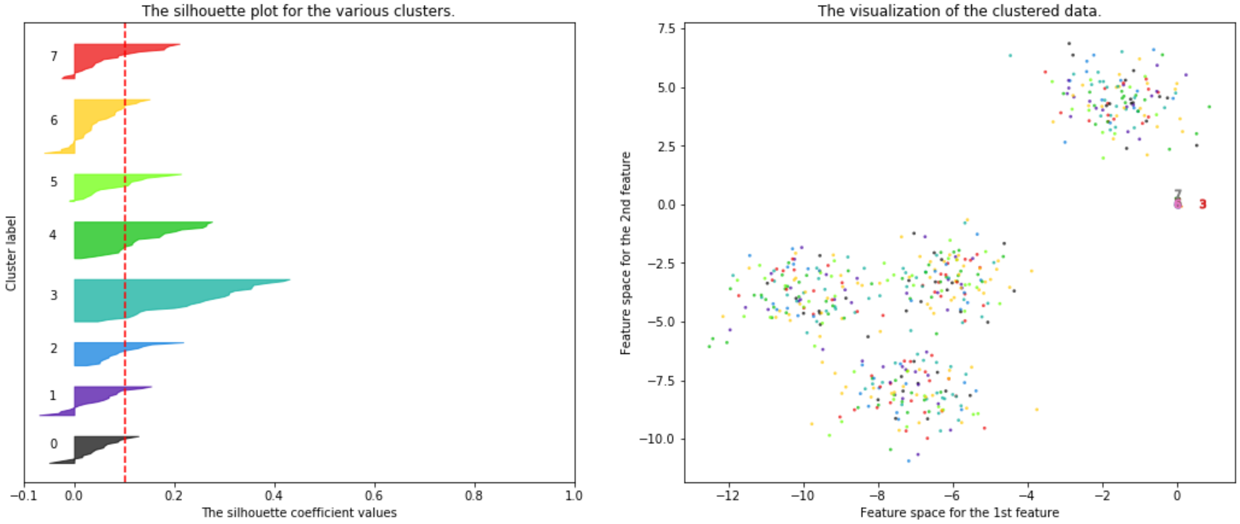
Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



The silhouette plot for the various clusters.

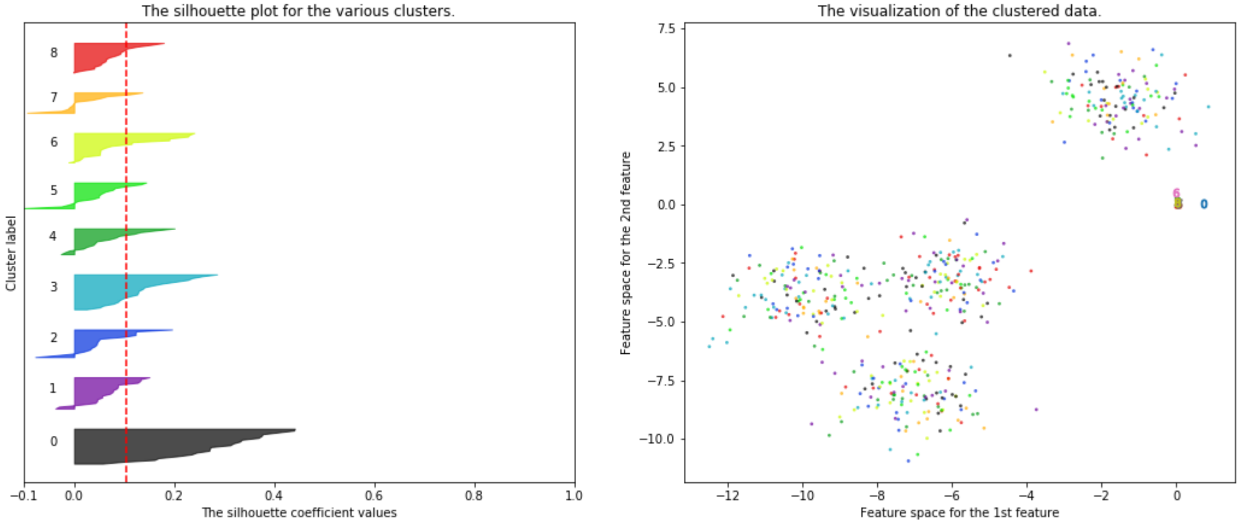
The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 8


The silhouette plot for the various clusters.

The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 9


The silhouette plot for the various clusters.

The visualization of the clustered data.

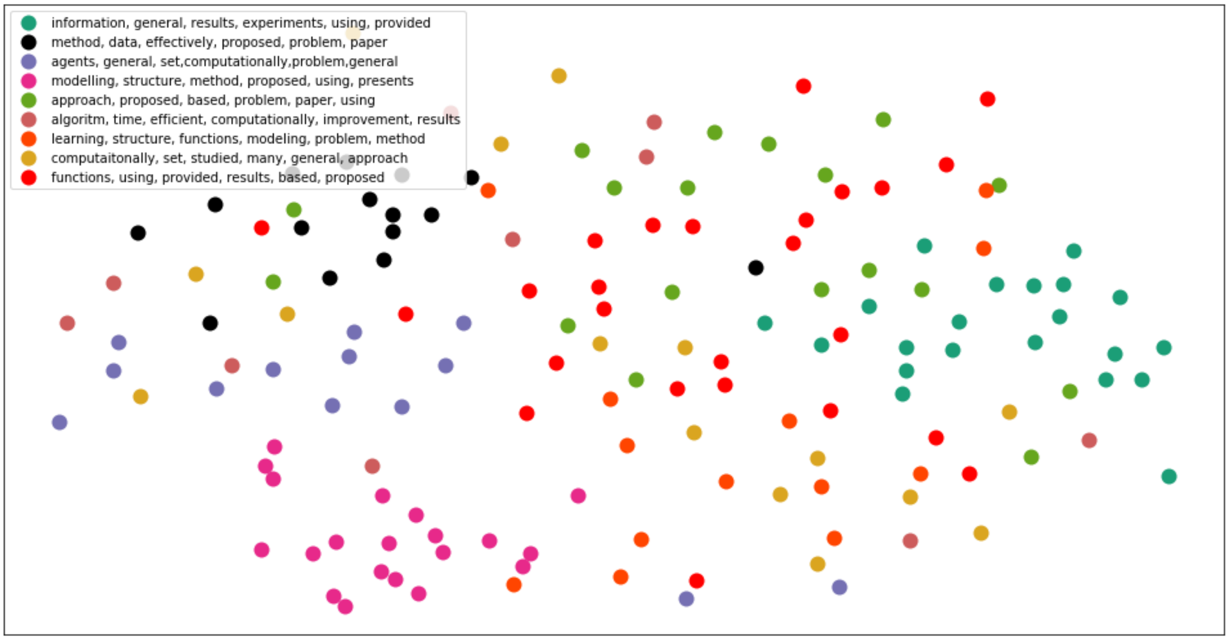
Silhouette Scores:

K Value	Silhouette Score
2	0.101360679412
3	0.0839263423597
4	0.0830329997309
5	0.0830329997309
6	0.0910970720602
7	0.0910970720602
8	0.101361882771
9	0.103257371795
10	0.0918778421579

From the table, we see that Silhouette Score is max at K=9 (closest to 1) thus K=9 is the optimal choice here with only the cluster 0 seeming irregular (not having comparable width) in the picture shown above.

RESULTS

The following 9 clusters of words were obtained.
0: 'information, general, results, experiments, using, provided',
1: 'method, data, effectively, proposed, problem, paper',
2: 'agents, general, set, computationally, problem, general',
3: 'modelling, structure, method, proposed, using, presents',
4: 'approach, proposed, based, problem, paper, using',
5: 'algorithm, time, efficient, computationally, improvement, results',
6: 'learning, structure, functions, modeling, problem, method',
7: 'computationally, set, studied, many, general, approach',
8: 'functions, using, provided, results, based, proposed'



Visual representation of the clusters is shown above.

DISCUSSIONS

K-means initializes with a pre-determined number of clusters (chosen 5). Each observation is assigned to a cluster (cluster assignment) so as to minimize the within cluster sum of squares. Next, the mean of the clustered observations is calculated and used as the new cluster centroid. Then, observations are reassigned to clusters and centroids recalculated in an iterative process until the algorithm reaches convergence.

The clustering algorithm though provides necessary accuracy but it is not of a particularly precise one. The LDA and Hierarchical Clustering algorithms may be better suited to cluster the data and not result in so many overlaps.

CITATIONS:

Aljaber, B., Stokes, N., Bailey, J. et al. Inf Retrieval (2010) 13: 101. doi:10.1007/s10791-009-9108-x

Chik, F., Luk, R., & Chung, K. (2005). Text categorization based on subtopic clusters. *Natural Language Processing and Information Systems, 3513*, 203–214.