Unsupervised Learning: K-means Clustering

Group: Soham Biswas (MDS202147)
Aniket Santra (MDS202106)

Introduction:

In this assignment we are provided with The "Bag of Words" data set from the UCI Machine Learning Repository contains five text collections in the form of bags-of-words.
We are required to consider 3 of these text collections, namely Enron emails, NIPS blog entries, KOS blog entries.
Each dataset is of different sizes.
Each row of these datasets represent a docID, a wordID and their corresponding count (i.e. the number of times the given word occurs in the given document.
We are required to cluster the documents in these datasets via K-means clustering for different values of K and determine an optimum value of K.

Jaccard Index:

As a similarity measure we have used Jaccard similarity index which measures similarity between two documents based on the overlap of words present in both documents.

Jaccard index is extensively used to measure the similarity of sets. The index is often used as a similarity measure for sparse binary data sets since it takes the common or disjoint elements in two sets. The coefficient is occasionally used to measure the similarity of text, which corresponds to quantitative, multidimensional data when handled as a bag of words.

Calculating the Jaccard index for disjoint sets results in 0 since the count of common elements is zero. In contrast, the calculation of this measure for two sets involving the same members would result in 1. The Jaccard coefficient for any two sets is within the range [0, 1]. A higher score indicates a large number of common items in comparison with the total count of elements in sets. Accordingly, high scores for the Jaccard coefficient signify a high degree of similarity. However, specifying thresholds might depend on the problem domain.

Jaccard Index $= \frac{A \cap B}{A \cup B}$ where, A and B represent 2 documents

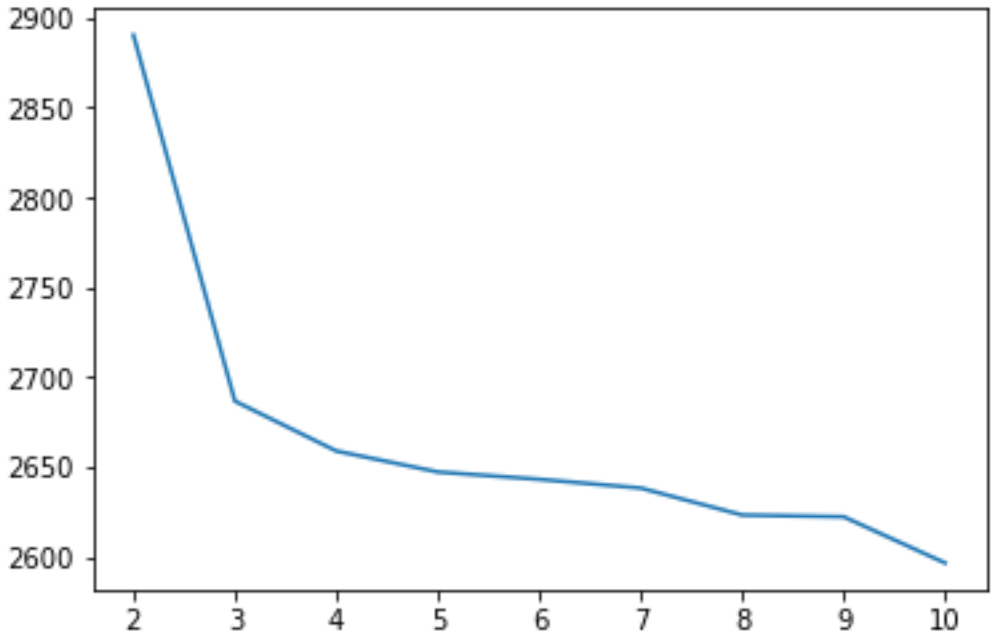
Procedure:

- First we create a binary sparse matrix for each of the datasets where the rows represent the docID (i, say) and columns represent the word ID (j, say), and the (i,j)th element represents the presence (represented by 1) and absence (represented by 0) of the jth word in the ith document.
- Thus, the dataset of the underlying model changes from "bag of words" to "set of words".
- In the case of the KOS dataset, and NIPS dataset we proceed with the further computation using the whole sparse matrix, but in the case of the Enron dataset, we use a simple random sampling scheme to sample 4000 documents song the 39861 documents and form a new sample matrix which we use in our further computations. Since the information about the various document types is unknown, other sampling methods could not be utilised and SRS therefore ensures a proper representation of the dataset.

- Using these matrices as our input, we proceed to use a modified K-means algorithm(modified based on the K-means clustering algorithm available in SKlearn package) to obtain the required clusters for each dataset.
- The modification is performed by creating the class Jaccard_K_Means().
- Jaccard_K_Means()
    - We initialize k random data points/documents as centroids using initialise_centroids function (similar to the approach followed by sklearn.KMeans)
    - The Jaccard Distance of each document from each centroid is calculated using get_j_dist function for k centroids k*(D), instead of calculating Jaccard distance of each document with all other (D* D computations).
    - Each document is then assigned to a cluster which has minimum Jaccard Distance from its centroid using assign_to_clusters function.
    - Now based on the clusters formed, we find the new centroid by finding the Euclidean mean of all data points within a cluster using centroid_update function
    - The new centroid obtained is converted into a binary array using a threshold value which optimizes the inertia for a particular value of k in k-clusters using get_new_cent function.
    - Now all data points are reassigned to various clusters based on these updated centroids. These iterations continue until either the number of iterations cross max_iter or relative distance between the previous & current centroid for the given cluster is below the tolerance (tol).
    - Using these clusters, inertia is calculated for all data points, (where inertia is the sum of squared distances of documents to their closest cluster centroid).
    - This measure of inertia is calculated for 10 values of k & plotted. The optimal k is noted at the elbow of the graph.

Results:

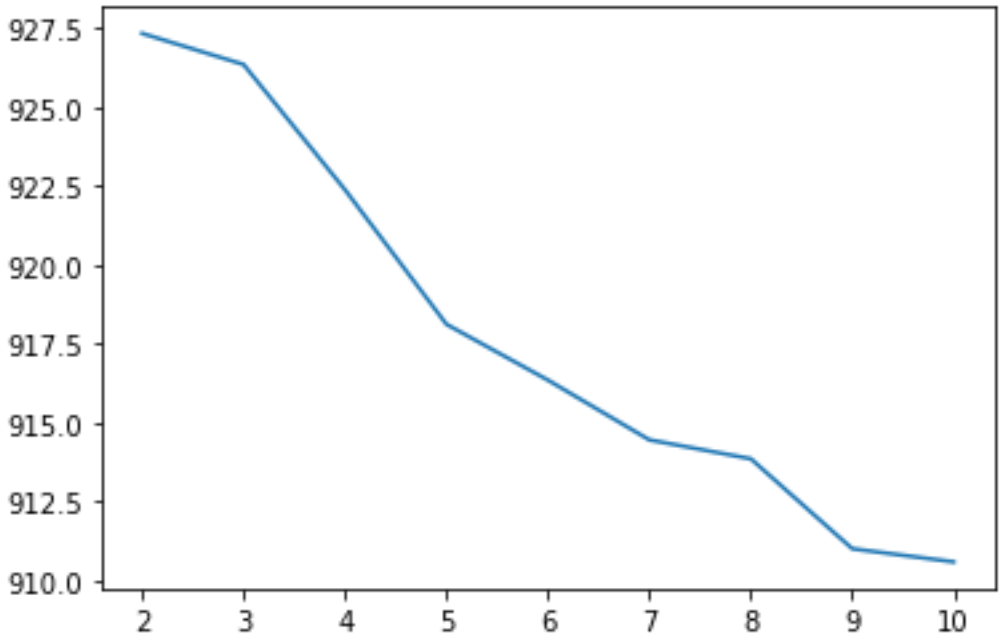
For KOS Blog Entries we get the inertia graph as:



From the graphs we conclude that the elbow is at k = 3.

That is the optimal value of k for this dataset is 3.

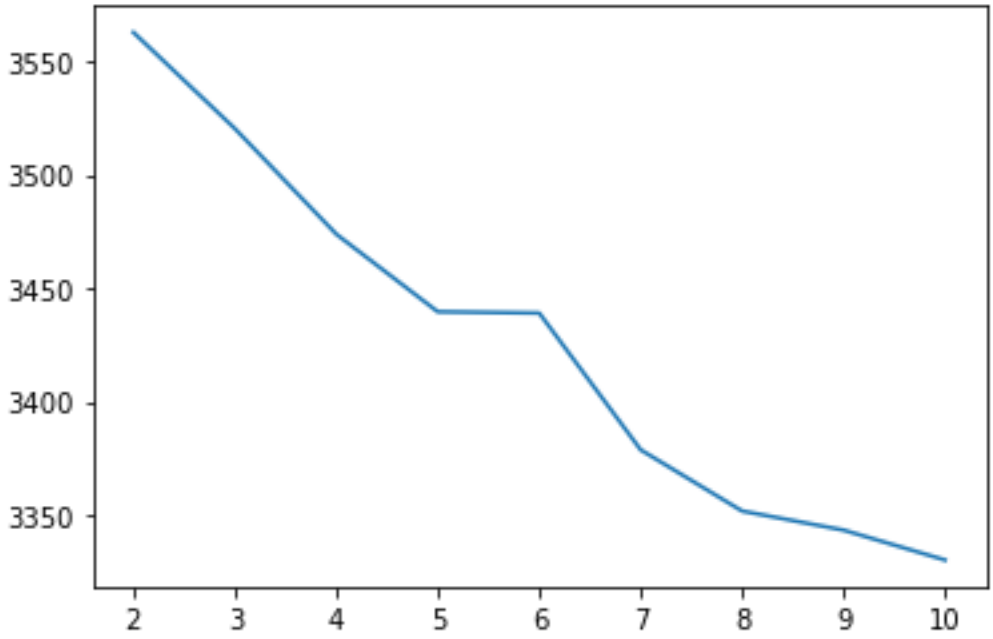
For NIPS full papers we get the inertia graph as:



From the graphs we conclude that the elbow is at k = 9.

That is the optimal value of k for this dataset is 9.

For Enron emails we get the inertia graph as:



From the graphs we conclude that the elbow is at k = 8.

That is the optimal value of k for this dataset is 8.

----------X----------