Bag of Words using Jaccard Index

Shiuli Subhra Ghosh (MDS202035), Suman Roy (MDS202041)

July 3, 2021

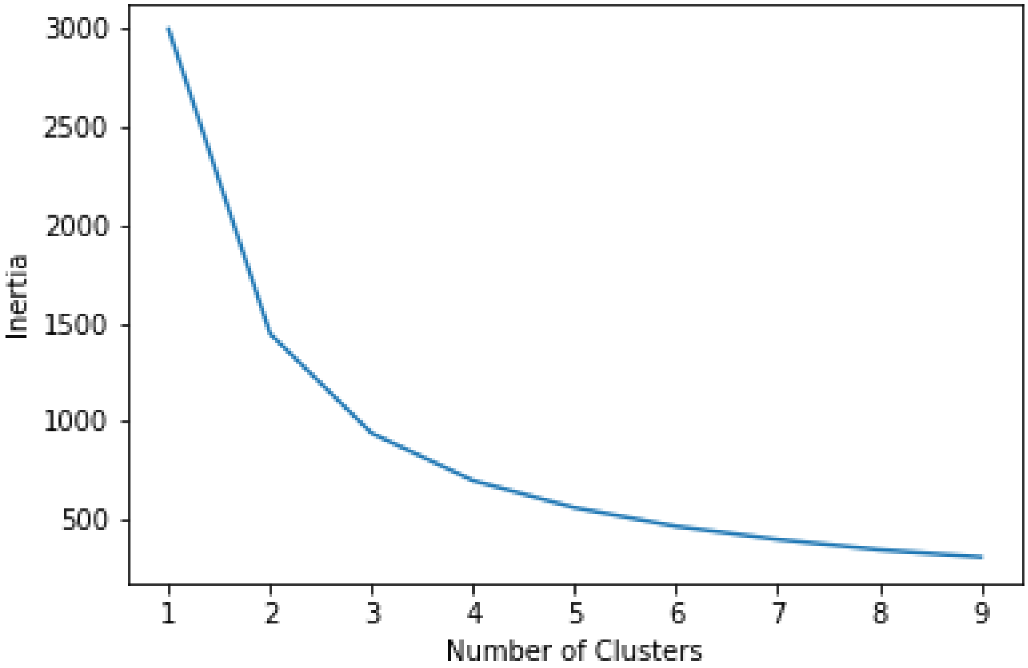
1 Procedure

- First we read the .txt files into Python and processed it in a suitable manner and put the data into a dataframe where the rows represent the word index and the columns represent document index.
- Then we calculated pairwise Jaccard distances of the documents.
- We initialised k clusters randomly.
- Then by taking only the rows of Jaccard matrix corresponding to the centroids , we compared each document's Jaccard distance with these centroids.
- Now we added the particular document to the cluster of the centroid with minimum Jaccard distance.
- For updating the centroid of the cluster we took the Jaccard matrix only of the cluster and Found the sum of the Jaccard distances corresponding to all the documents.
- We updated the centroid with the document with the lowest sum.
- This process is performed iteratively.
- Output is the final Centroid list and CLuster Labels.

2 Optimizing value of k

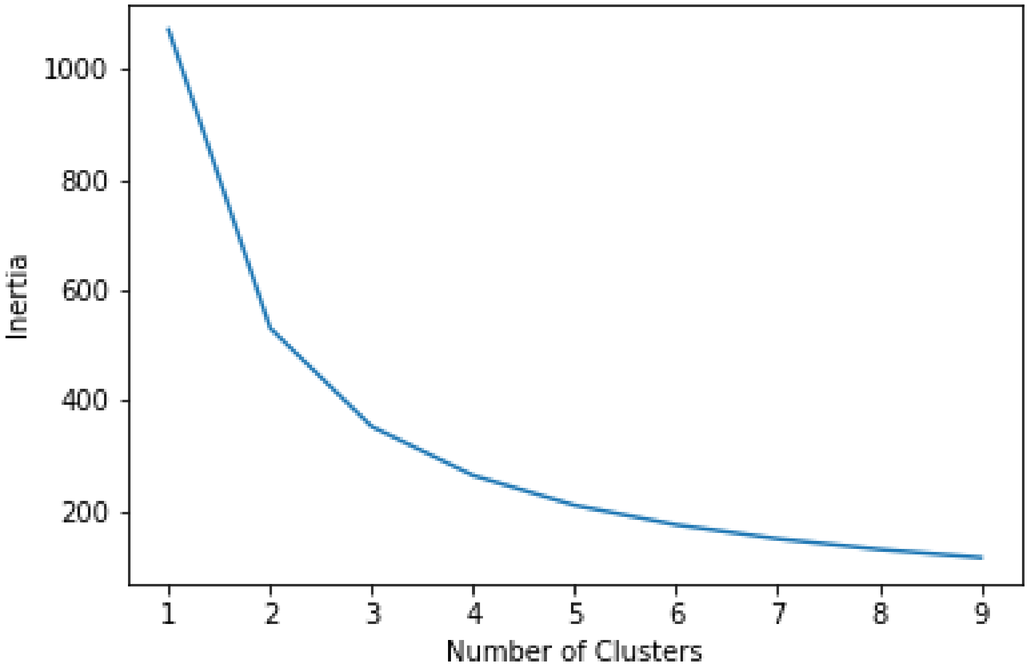
We visualised the inertia vs. no of cluster plots and found that a sharp elbow is coming at k=3 for both 'nips' and 'kos' dataset. So we published our final result taking k=3 for both these datasets.

2.1 Results for Kos Data Set:



The optimal cluster number is 3 for Kos data set.

2.2 Results for Nips Data Set:



The optimal cluster number is 3 for nips data set.

3 Results

https://drive.google.com/drive/folders/1BWV_ojivbgwwuL3focHN0GKXoH1HOH61?usp=sharing