Data Mining Assignment-4

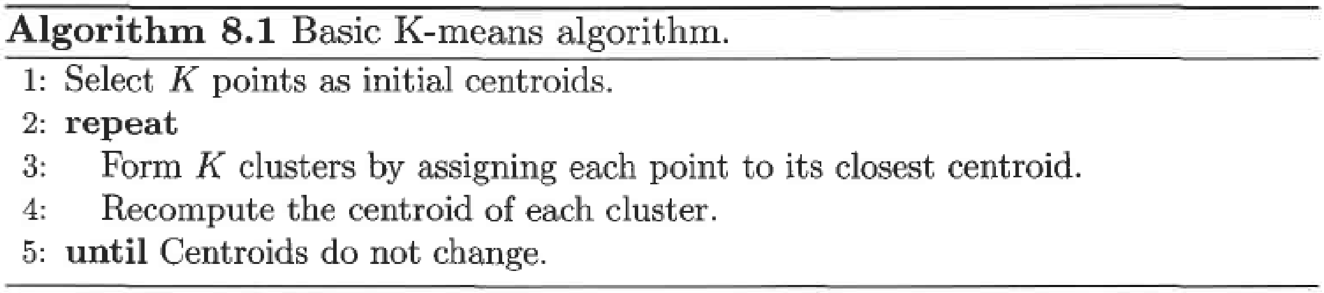
CLUSTER ANALYSIS

Prashant Pathak
MT19051
DMG

13-Sep,2019

K-means

K-means is a prototype-based clustering technique which clusters the data point based on the distance between them. It works on the idea that the points that are closer to one another have similar feature. So, one representative form the cluster is chosen which represent the behavior of the entire cluster. In k-means the Centroid is the element which is the representative of the cluster.



```
Algorithm 8.1 Basic K-means algorithm.
1: Select $K$ points as initial centroids.
2: repeat
3:    Form $K$ clusters by assigning each point to its closest centroid.
4:    Recompute the centroid of each cluster.
5: until Centroids do not change.
```

Figure 1 Algorithm of the k-means clustering

Usually the starting points are chosen at random. And then the clustering is applied.

In Step 3 the notion of closest is used. Here some distance or similarity notion can be used to find the closest point. Example of distance measure are Euclidean distance, Manhattan distance. Example of similarity are Jaccard similarity and cosine similarity.

For recomputing the centroid position, we take the mean of all the point falling into the same cluster.

Time Complexity

The time complexity of the k-means: **O (I x K x m x n)**, where, I is number of iterations, K is the number of clusters is the data point and n are number of attributes.

Advantages

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.

Disadvantages

- We have to choose value of k manually.

  One example to choose the value of k would be to plot the SSE vs k and see where an elbow shape is made. That is our value of k.
- Being dependent on the initial value.
- Can cluster data of various size and density.
- Clustering outlier

  The presence of noise in the cluster can make the k means algorithm to perform worse.
- Curse of dimensionality.

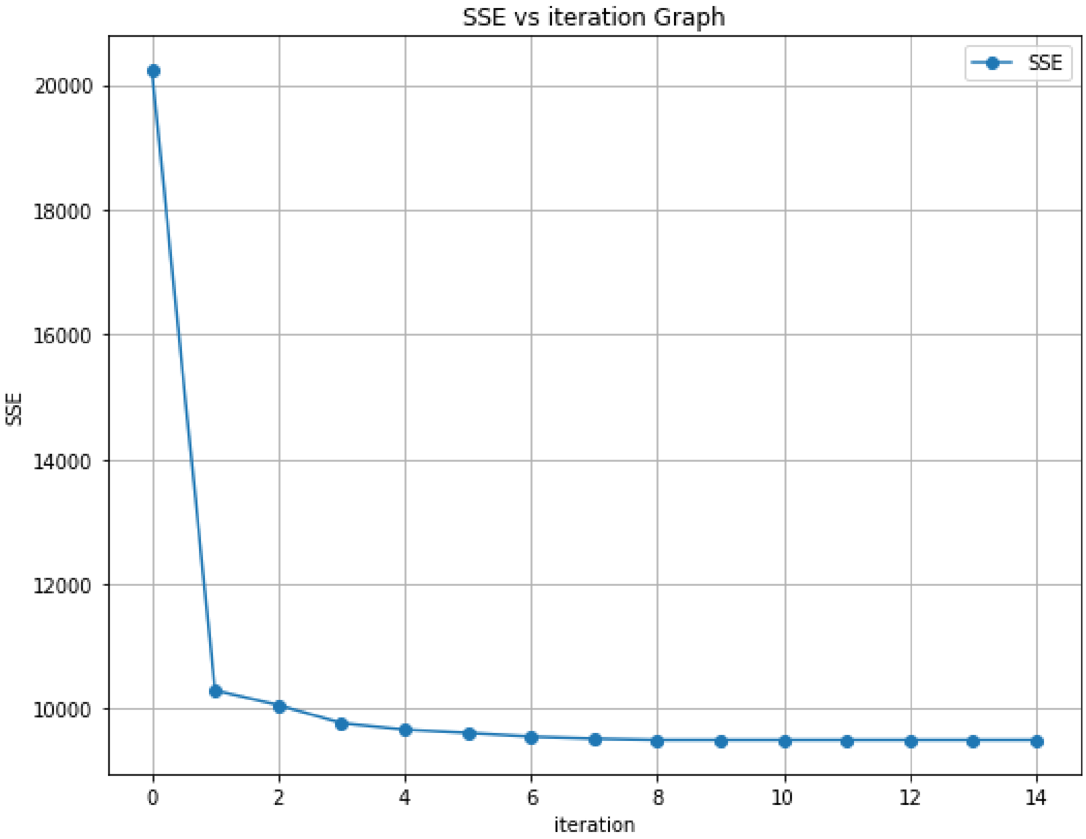
About the data

We have been given 4 files with have different category of the word. Example. there is a file which have word related to animals. Like wise we have 4 files with different category like, animal, veggie, fruit and country.

Implementation

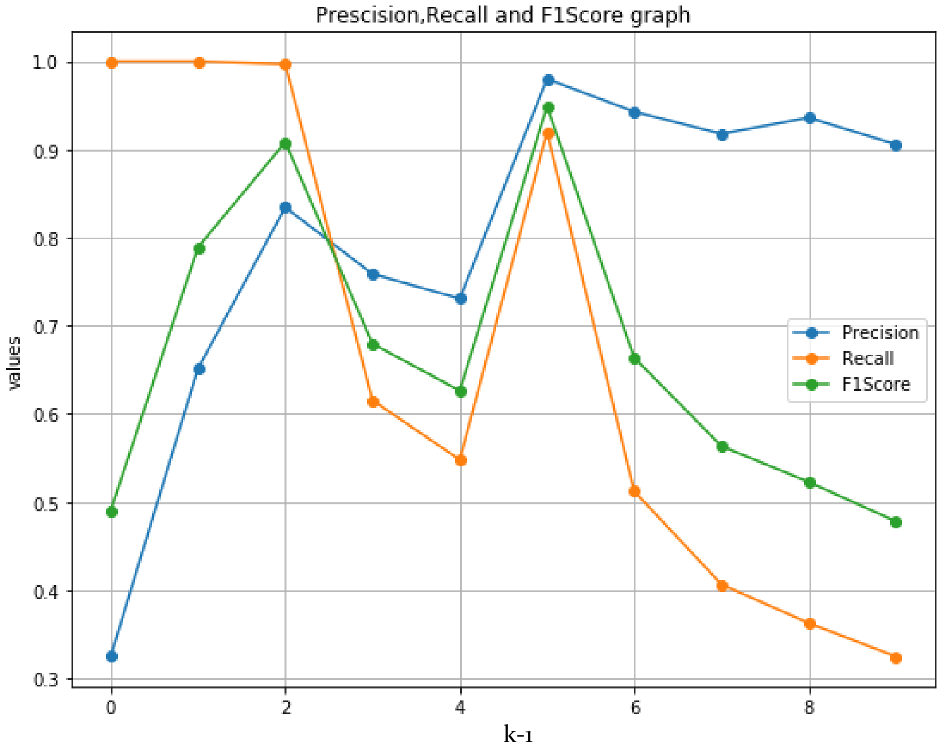
We have implemented the k-means form the scratch. Initially we have taken the distance to compute the cluster

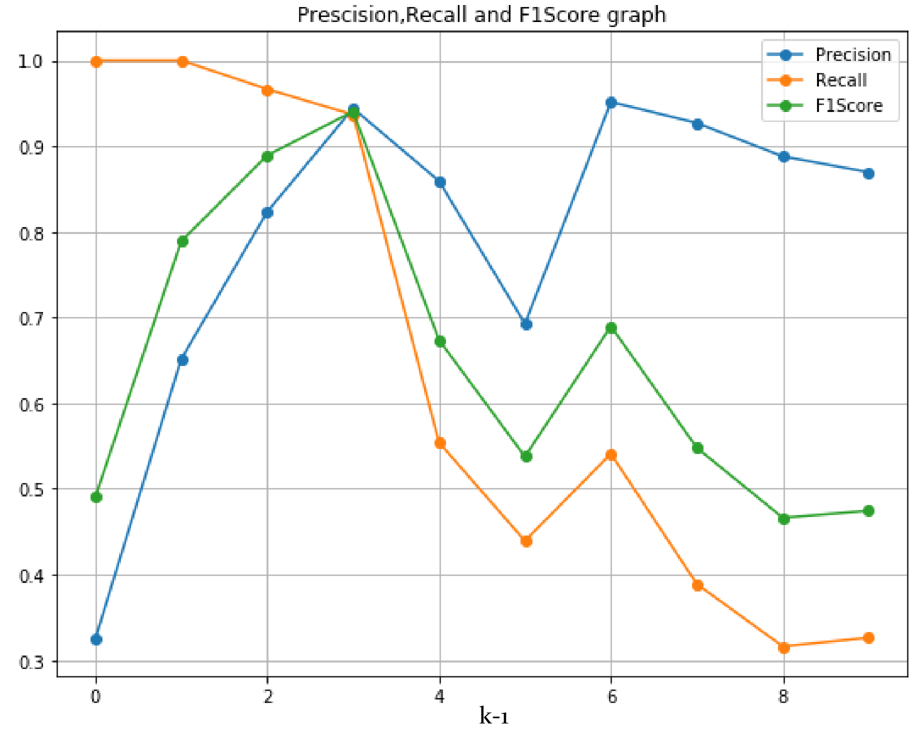
PART 1: SSE VS K GRAPH



The above figure shows that with each iteration one SSE (sum of square error) is decreasing with every iteration and thus converging on a local optima. This shows that our Euclidian distance is correct.

PART2 QUALITY OF MEASURE- PRECISION, RECALL AND F1 SCORE OF EUCLIDEAN DISTACNE

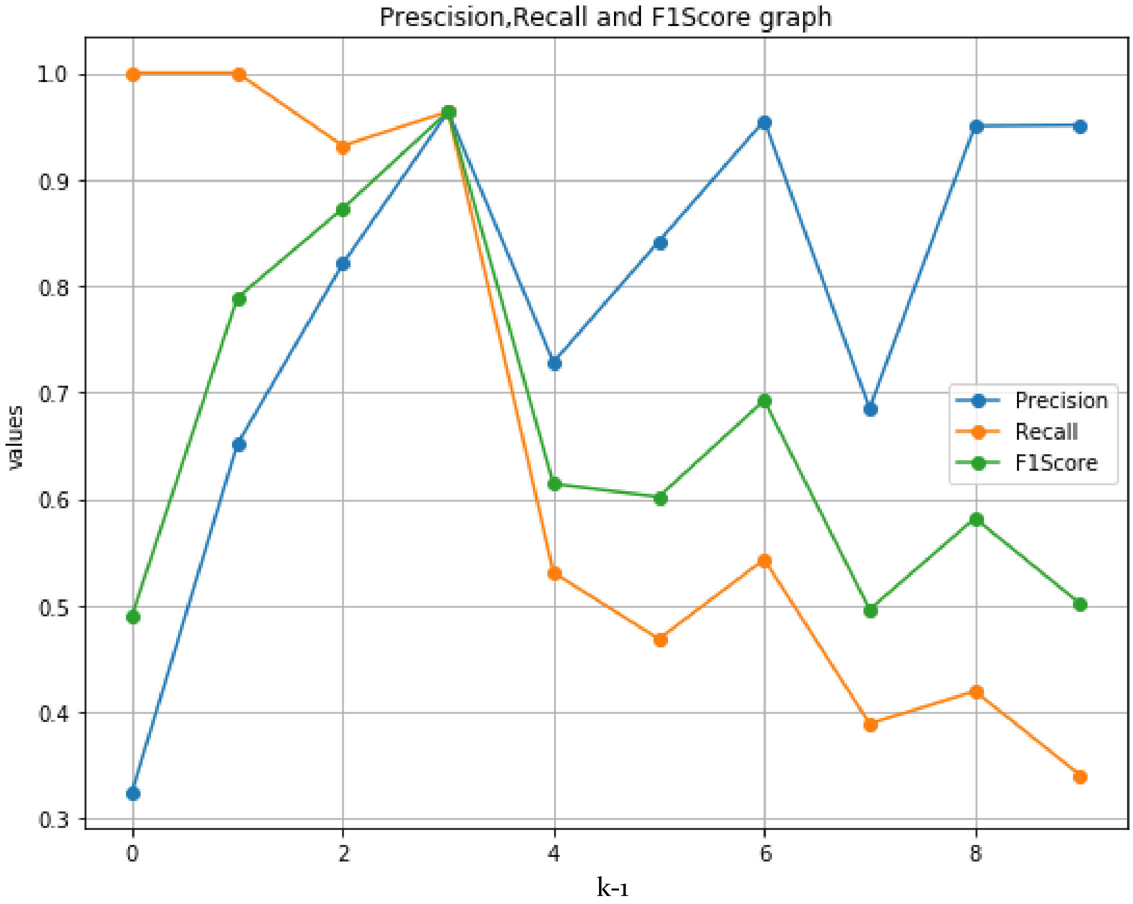




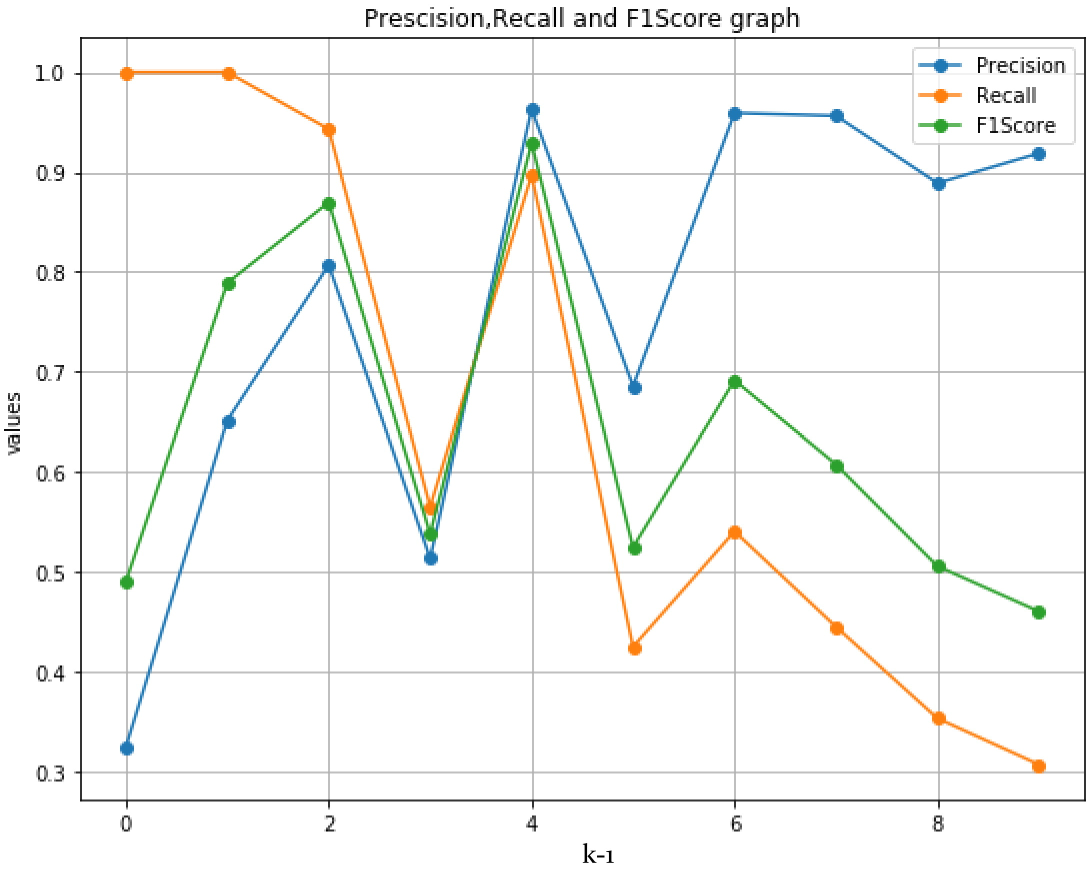
We can see that running multiple iteration of k-means can lead to different graph. The cluster which has highest precision and recall is conceded the best graph

In the first graph we see that 6 cluster gives the best precision and recall but in the second graph we can that 3 cluster gives us highest precision and recall.

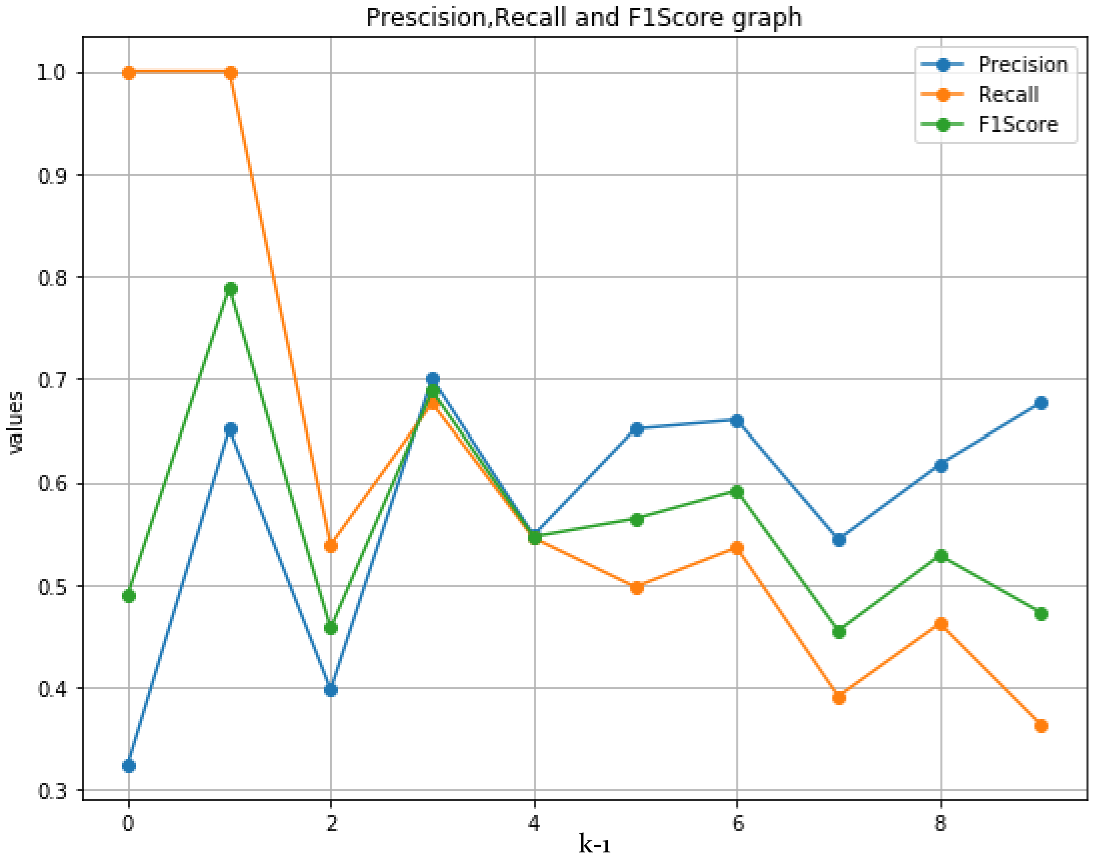
PART3: EFFECT OF NORMALIZATION ON CLUSTER



PART4: QUALITY OF MEASURE WITH MANHATTAN DISTANCE



PART5: QUALITY OF MEASURE WITH COSINE SIMILARITY



PART6: COMPARISION BETWEEN ALL THE ABOVE CLUSTER

Metric Used	Precision (approx.)	Recall(approx.)	F1 -Score (approx.)	Value of k
Euclidean Distance	.94	.94	.94	4
Euclidean Distance with Normalized data	.95	.95	.95	4
Manhattan distance	.92	.90	.91	5
Cosine Similarity	0.68	0.68	0.68	4

- We can see from the above table that Euclidean distance works better in for our dataset because we get precision and recall of around 95%.
- After Euclidean distance Manhattan distance works better and then cosine similarity
- The best value of K that we get is 4.
- After 3 the best value of k that we get is 7.

So, to get best result we should choose 4 as value of k and use Euclidean distance as our distance measure.

References

- https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html