Clustering

Cluster the words belonging to four categories: animals, countries, fruits, and veggies. The words are arranged into four different files. The first entry in each line is a word followed by 300 features (word embedding) describing the meaning of that word.

1) Train a K-means model with Euclidean distance to cluster the instances into k clusters.

2) Vary the value of k from 1 to 10 and compute the precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and precision, recall, and F-score in the vertical axis in the same plot.

3) Now re-run the k-means clustering algorithm you implemented in part (1) but normalize each feature vector to unit l2 length before computing Euclidean distances. Vary the value of k from1 to 10 and compute the precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and precision, recall, and F-score in the vertical axis in the same plot.

4) Now re-run the k-means clustering algorithm you implemented in part (1) but this time use Manhattan distance over the unnormalized feature vectors. Vary the value of from 1 to 10 and compute the precision, recall, and F-score for each set of clusters. Plot k in the horizontal axis and precision, recall, and F-score in the vertical axis in the same plot.

5) Now re-run the k-means clustering algorithm you implemented in part but this time use cosine similarity as the distance (similarity) measure. Vary the value of k from 1 to 10 and compute the precision, recall, and F-score for each set of clusters. Plotkin the horizontal axis and precision, recall, and F-score in the vertical axis in the same plot. 6) Comparing the different clusterings you obtained in (2)-(5), discuss what is the best setting for k-means clustering for this dataset.