Text Clustering

By James Edwards

Formatting The Data

- Read in the documents
- Remove punctuation
- Convert into a list in “bag of words” format

Create The Corpus

- Read through each document
- Add each unique word to the corpus

Term Frequencies

- Represent each document with a list the size of the corpus
- Fill the list with the count of the term at the respective index of the corpus
- If count is not zero, replace with 1 + log(count) to weight the term

Cosine Similarity

- Finding the similarity score of two documents
- Find the product of the values at each matching index
- Add them all up to get the similarity score
- The higher the score, the closer the documents are to one another

K-Means Clustering

- Choose K random points to be the starting centroids
- Assign the rest of the points to the nearest centroid
  - The one with the greatest cosine similarity
- Calculate the new centroid of each cluster by averaging all the points in the cluster
- Reassign all the points based on the new centroids and recalculate the new average
- Repeat until the average doesn't change or barely changes

DBSCAN

- Use same methods as K-Means to read in data, convert to weighted term frequency, calculate cosine similarity and cluster average
- Difference comes with how points are assigned
- No set number of clusters
- Eliminate noise points – points with fewer than a predetermined number of neighbors within a set range

DBSCAN Steps

- Go through all points
- If point isn't assigned a cluster, increment cluster number by one and assign it to point
- Assign any points within the predetermined epsilon range of the core point to that same cluster

SSE

- Score assigned to determine how good a cluster is
- Used to determine which implementation is best
- Calculated by adding the similarity score of the mean of each cluster and every other point in the cluster
- The higher the total score, the more effective the implementation

Results

- K-Means clustering was the most effective implementation
- K = 10 was the best K value
- All but K = 5 were better than DBSCAN

Results