Document Clustering

Daisy Xu
Department of Computer Science
University of Washington
Seattle, WA 98105
daisyx@cs.washington.edu

Brian Lunder
Department of Computer Science
University of Washington
Seattle, WA 98105
lunderb@cs.washington.edu

Abstract

Document clustering can be done with a number of algorithm. One of the most common algorithm used is K-Mean Clustering. Another algorithm used is Spectral Clustering. Both are effective algorithm used for document clustering and trying to predict an article that the user may want to read next after reading their current article. This paper will predict a similar article using K-Mean Clustering and Spectral Clustering and compare their results to see which algorithm is more effective for document clustering.

1 What is done

We've obtained the data set for our experiment, and written the software for both of the algorithm looked at in this paper.

1.1 split_data.py

A program, split_data.py, split the dataset into a training and testing data set. The program split the data set using a 60-40 ratio and save the two separate data set as separate text files.

1.2 Parser.py

Parser.py takes in the data set text file, and parse the data into a pandas data frame.

1.3 k_means.py

k_means.py clusters the training data set into cluster and find the center for each cluster.

We initialize the cluster centers as a random data point in the training data set, making sure that the data point selected was not used for another cluster center. Iterating through the data set, we classify each data point to the cluster they are most close to. Then we calculate the new mean for each cluster by adding up the values of each data point in the cluster and dividing them by the number of data point in the cluster. We continue classifying and calculating the new mean until either the cluster center does not change from the last iteration or there's been 100 iterations. We check how much the cluster center change by calculating the distance between the new mean and the old mean, and check if the change in distance is smaller than 0.00001.

1.4 spectral.py

spectral.py calculate the affinity matrix, and the degree matrix to calculate the Laplacian.

Using the Laplacian, we find the k largest eigenvectors, and use K-Means to cluster the documents based on new feature space, determined by the eigenvectors.

The affinity matrix calculated by using a Gaussian Kernel between each data point. The degree matrix is calculated by the summation of the affinity values for each row. The Laplacian is the degree matrix subtracted by the affinity matrix. Using SVD, we find the k largest eigenvectors, and cluster the data using K-Means on the eigenvectors.

2 Preliminary Test Results

Using a small data set of 200 samples, we ran K-Means Clustering and Spectral Clustering to test how well each algorithm cluster the data.

2.1 K-Means

On the small data set, K-Means clusters the data, however, some clusters had zero data points in its cluster. This may have to do with how we are initializing the initial cluster center. We may not be initializing correctly, which leads to some clusters to have zero data points in them. One cluster is orders of magnitude larger than the other clusters.

2.2 Spectral

On the small data set, Spectral clusters the documents, however, because K-Means was not initialized correctly, Spectral also was not able to cluster the data correctly. Spectral depends on K-Means to compute the clusters and the cluster centers, so when Spectral finish running, some clusters have zero data points in them because K-Means was incorrect. One cluster is orders of magnitude larger than the other clusters.

3 What still needs to be done

We still need to test different number of k clusters, make predictions of the articles in the testing data set, and run K-Means Clustering and Spectral Clustering on the whole training and test data set.

For initializing the clusters in K-Means, we still need to experiment with what's the best way to initialize the clusters. We plan to set a threshold on how far away the initial clusters are from each other, and make sure that the initial cluster centers are far apart from each other that they'll converge close to the cluster centers

References

[1] Jain, Anil K. (2009). Data Clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31. Received from http://www.sciencedirect.com/science/article/pii/S0167865509002323.

[2] Khare, Purvi. (2015). Improved Clustering Technique Using K-Mean Method. *International Journal of scientific research and management*, 3. Received from http://www.ijsrm.in/v3-i12/5%20ijsrm.pdf.

[3] Singh, Aarti. *Spectral Clustering [PDF Document]*. Received from https://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture21_2.pdf.

[4] Wang, Xiang, & Qian, Buyue, & Davidson, Ian. (2012). Improving document clustering using automated machine translation. *CIKM '12*. Received from dl.acm.org/citation.cfm?doid=2396761.2396844.

[5] Zografos, Vasileios & Nordberg, Klas. *Introduction to spectral clustering* [PDF Document]. Received from https://www.cvl.isy.liu.se/education/graduate/spectral-clustering/SC_course_part1.pdf.