ELEG 6913: Machine Learning for Big Data
Fall 2016
Department of Electrical and Computer Engineering
Prairie View A&M University

Project 2

(a) Text clustering is to automatically group textual documents (for example, documents in plain text, web pages, emails and etc) into clusters based on their content similarity. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering. The target of this project is to construct a text cluster to group texts into different clusters. Text data containing 5 classes for the project is available at https://github.com/BruceDong/Resources-for-Projects-of-Machine-Learning/tree/master/Datasets/cluster.

(b) Choose 2 classes of data from the text data set to build the clustering data.

(c) Use the feature extraction method: binary feature to represent the clustering data as feature data.

(d) Choose k-means algorithm as the text clustering algorithm to group the feature data into two classes.

(e) Evaluate the clustering results by calculating Accuracy.

(f) Repeat everything in (b) with another feature extraction method: term frequency–inverse document frequency (TF-IDF).

(g) Repeat everything in (d-e). Compare your results with (e).

(h) Repeat everything in (c-g) on the whole feature data set.

(i) Summarize your observations and conclusions.

Bonus Section:

If you can implement "Fast k-means algorithm clustering"[1] *with C++ as the clustering algorithm for this project, you will obtain extra 20% scores. Requirements:*

(1) Comparing Accuracy between k-means and Fast k-means

(2) Comparing Speed between k-means and Fast k-means

Submit your programs and supporting documents as one zip file in ecourses by November 28, 2016.

[1] https://arxiv.org/abs/1108.1351