CSCE578 Assignment 3 Proposal

Sebastian Martin

March 18, 2019

For the third assignment I will be choosing Option one on the proposal assignment sheet. The reason that I have chosen this is because it will allow me to explore the code necissary for the cluster graphing and fully figuring out how TF-IDF functions in order to properly implement this into the final project.

For this assignment all of the student essays will be taken and catagorized by their general themes using the clustering technique. The first step will be to take all of the Essays and compute a TF-IDF on each on in order to find all of the most important terms within each document. Once this has been completed for all of the documents the matrix will be created in order to store all of these values. This matrix will be normalized along each column (document space), forcing each vector to be a unit vector. These will then be graphed and analyzed wo determine which documents fall into which cluster. Based off this we will be able to determine which documents are related based on the distance of the points to each other and the proximinty of the cluster.

If time permits a convex hull may also be used to determine the spread of each cluster. This would allow us to determine the spread that all of the points take, and how densly packed related documents are to each other.