Project 7 Executive Summary

Ari Burian

Problem:

How can 26,000 posts of Reddit be clustered into subgroups?

Solution:

Specific columns were identified which could have a correlation with the clustering of the Reddit posts into subgroups. The numeric columns of time created, self-post, number of comments, date retrieved on, and score were joined together with the word count for each unique word in the author, domain, and title of the article.

A dimensional reduction was done using PCA to limit the total number of features to 10, and a cluster model (KMeans) was used to group the articles into clusters. After looking at various amounts of clusters, the data visualizations implied that 2 clusters would be ideal, as seen in the image below.



Looking back at the original articles, it appears these 2 cluster could divide the articles into two subject areas: troubleshooting areas where problems have occurred for the author, and general curiosity that the author has about a subject area.