Authors: John Waldo and Adam Wolfson

Net Ids: jw922, amw337

CS5785 Homework 3 – Part 2

1. The analysis of the k-means algorithm was interesting. One thing that I noticed was the graph of SSE/N vs cluster size was not as clear cut as one would have hoped, so there were many options to select for K that could have been easily justified. Another point to note is that when using km.transform() for a specific cluster to turn everything into cluster space, the distance analysis returned many duplicates for titles in part 1a. This was handled by filtering and only analyzing the specific entries assigned to each cluster for distance analysis. One confusing aspect of the output was that when grouping by terms instead of titles there is some overlap between the top ten words per cluster. There was also much less overlap in titles for the first analysis in terms of max distance from the cluster center minus the mean array, likely because titles are much more distinct and unique than words.
2. I found it very interesting to extend the k-means algorithm. Especially with a dataset in which clusters may overlap or share details GMM – EM seems very powerful. In this setting, to be able to have a probability that a datapoint belongs to a cluster instead of simply declaring the most likely cluster.