Kyle Benson
CS 273A - Machine Learning: Fall 2013
Homework 5

Problem 1: Basics of Clustering

(a) TODO

Problem 2: k-means Clustering on Text

(a) Done

(b) Cost function results: 2.7047, 2.5655, 2.1498, 2.5837, 2.6072
Cluster 1 has 26 articles
Most common words: mississippi, play, scored, game, half, points, ranked, thomas, night, team,
Cluster 2 has 73 articles
Most common words: issue, number, days, times, 500, front, paper, 000, april, desk,
Cluster 3 has 1 articles
Most common words: washington, clinton, mall, lincoln, memorial, america, celebration, crowd, fireworks, friends,
Cluster 4 has 2 articles
Most common words: air, airlines, friday, traffic, aviation, commercial, midnight, 2000, canceled, control,
Cluster 5 has 40 articles
Most common words: children, including, post, produced, programs, television, american, art, black, calif,
Cluster 6 has 2 articles
Most common words: city, mexico, y2k, millennium, computer, government, red, armed, business, called,
Cluster 7 has 4 articles
Most common words: millennium, celebrate, millions, square, times, vast, cities, computer, crowds, fireworks,
Cluster 8 has 7 articles
Most common words: millennium, thousand, calendar, western, days, end, religious, africa, agree, ancient,
Cluster 9 has 1 articles
Most common words: times, york, square, 2000, saturday, midnight, early, 1999, dec, newyear,
Cluster 10 has 2 articles
Most common words: night, parade, rain, family, millennium, eve, friends, want, calif, colorado,
Cluster 11 has 2 articles
Most common words: mall, celebration, friday, police, close, crowd, evening, millennium, night, president,
Cluster 12 has 1 articles
Most common words: feet, square, broadway, seventh, street, side, times, avenue, block, million,
Cluster 13 has 12 articles
Most common words: completed, dancers, cast, director, ends, stage, thomas, 1995, advantage, afternoon,
Cluster 14 has 1 articles
Most common words: hong, kong, 000, evening, horse, millennium, chinese, eve, midnight,

party,
Cluster 15 has 6 articles
Most common words: 2000, problem, computer, problems, computers, analysts, fix, glitches, government, y2k,
Cluster 16 has 1 articles
Most common words: susan, y2k, smiths, farm, call, food, going, jim, real, thing,
Cluster 17 has 7 articles
Most common words: susan, y2k, smiths, farm, call, food, going, jim, real, thing,
Cluster 18 has 1 articles
Most common words: jerusalem, friday, christians, city, mount, 000, christian, eve, gate, jesus,
Cluster 19 has 37 articles
Most common words: square, times, home, midnight, 2000, ball, watching, celebration, hour, party,
Cluster 20 has 2 articles
Most common words: times, square, abc, brown, perfect, america, beach, broadcast, broadway, business,

We can see that most of the clusters seem to have some sort of common theme between the words, though in some cases we may need to look at more words to discern what that is. Interestingly, we see that clusters 16 and 17 should probably have been merged as they have the exact same most common words.

(c) The documents do not seem to form coherent groups very well. It seems as if some of the clusters should be split into several others as there appear to be multiple categories of documents within each cluster. It also doesn't seem like the documents were assigned to their ideal clusters. For example, one cluster had a bunch of documents about Putin and other Russian politics, but the common words associated with that cluster indicate that it should be something about dancing or performances.

(d) The sum of squared distances is **1.7278**, which is significantly lower than the previous clustering.1

Problem 3: EigenFaces

(a) TODO

Problem 4: Project

Yessir!