11-741 k-means Clustering

Xin Qian

September 18, 2016

Due: Sept 22, 2016 11:59 PM

1 Corpus Exploration

Dataset	Development set	Test set
Total number of documents	942	942
Total number of words	254852	249516
Total number of unique words	14063	13924
Average number of unique words per document	174.59	173.37

For the first document in the development set, the total number of unique words is 161. All twice occurrence word ids are, 2, 5, 10, 18, 23, 27, 28, 30, 32, 42, 44, 45, 46, 50, 52, 60, 62, 69, 79, 87, 91, 99, 102, 107, 114, 141.

60 clusters: 0.518418660296 0.000441304462641
65 clusters: 0.544281681506 0.00103714651896
67 clusters: 0.543796820512 0.000707279336919
70 clusters: 0.532671281996 0.000507790508475
75 clusters: 0.55651164239 0.00166091932081
80 clusters: 0.571080054113 0.000701041384392