

Written Assignment 5

CS 510 TOP: Information Retrieval on the Internet

Assigned: May 18, 2010

Due: May 25, 2010

Possible points: 8

1. Consider the six data points given in Figure 16.7 in the textbook. Demonstrate, showing your calculations, the claim made in the caption, that 2-means clustering using $\{d_2, d_5\}$ as initial centroids converges to a different clustering than if the initial centroids are $\{d_2, d_3\}$. (b) Use the final RSS values to show that the latter yields a better clustering than the former. (2 points)
2. Consider the eight data points shown in Figure 17.4 in the textbook. (a) Demonstrate, showing your calculations, how the two different hierarchical clusterings (left and right figures) are created using single-link and complete-link similarity measures. (b) Show what clustering of these points would be created using centroid similarity (illustrated in Figure 17.3(c)). (2 points)
3. Read the short (and clear!) tutorial on LSI at <http://www.miislita.com/information-retrieval-tutorial/latent-semantic-indexing-fast-track-tutorial.pdf>

Repeat the steps given in this tutorial to determine the ranking of the three documents given the query “fire gold truck”. Show your work. (3 points)
4. Describe, in a paragraph or so, an experiment you could do to determine if Google is using Latent Semantic Indexing as part of its ranking strategy. (1 point)