## **CSCI 6370 IR and Web Search**

## Homework Assignment Five

Total points: 100. Due: Monday, July 6.

1. [35 points] Assume that we use cosine similarity as the similarity measure. In the hierarchical agglomerative clustering (HAC), we need to define a good way to measure the similarity of two clusters. One usual way is to use the group average similarity between documents in two clusters. Formally, for two cluster and , let , we define

Where is the cosine similarity between and .

Given a list of clusters , assume that their pairwise similarities are saved in a two dimensional array of size . Show that there is a way to find two mostly similar clusters in time.

1. [30 points] For a list of m documents in d-dimensional vector space, each iteration of the k-means clustering has a time complexity of , which the hierarchical agglomerative clustering (HAC) has a time complexity of . We know that the overall performance of the k-means clustering depends on the choices of initial centroids. However, there is no such an issue for HAC. Describe a method to use HAC to help the k-means clustering, but the method shall maintain the same time complexity for the k-means clustering.
2. [35 points] Assume you are working for amazon.com. You need to find some way to recommend products for an online shopper without asking for relevance feedback from any user. Provide one solution.