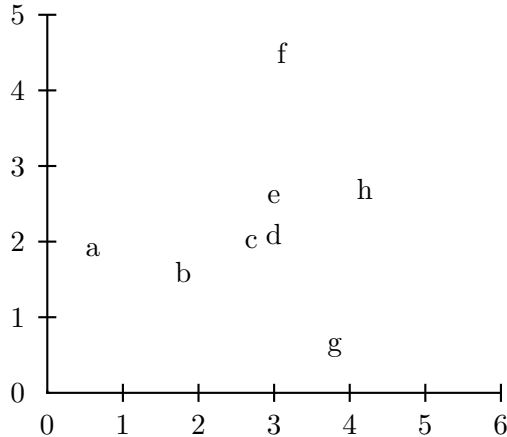


Information Retrieval and Text Mining: Assignment 5

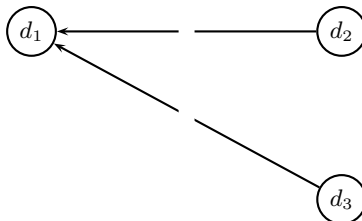
Problem 1. (20 points)



The points have the following coordinates: a: (0.6,1.9), b: (1.8,1.6), c: (2.7,2.0), d: (3.0,2.1), e: (3.0,2.6), f: (3.1,4.5), g: (3.8,0.6), h: (4.2,2.7). Define the similarity of two points as $-(x_1 - x_2)^2 - (y_1 - y_2)^2$.

Compute (i) single-link and (ii) complete-link clusterings of this set of points and depict them as dendrograms. Make sure to indicate the y-value of each horizontal “merge” line. (iii) Suppose you want to compute a GAAC of the set. Discuss what difficulty arises. Suggest a solution. (You don’t have to compute the GAAC clustering.)

Problem 2. (20 points)



For the web graph in the figure, compute PageRank, hub and authority scores for each of the three pages. Also give the relative ordering of the 3 nodes for each of these scores, indicating any ties. Assume that the PageRank teleport probability is 0.1. Normalize the hub and authority scores so that the maximum hub/authority score is 1.

Hint: Using symmetries to simplify and solving with linear equations might be easier than using iterative methods.

Problem 3. (10 points)

Rank the documents in collection $\{d_1, d_2\}$ for query q using the language model approach to IR introduced in class. Use the mixture coefficient $\lambda = 0.4$.

- d_1 : Scottish sheep getting smaller due to climate change study says
- d_2 : The analysis has shown a dramatic shift in the natural ranges for US Bird species in response to climate change
- Query q : climate change

Due date: Tuesday, July 14, 2009, 15:45 (turn assignments in before class in V38.03 or by email)