Student Name:

University of Texas at Dallas
Department of Computer Science
CS6322 – Information Retrieval
Spring 2011
Instructor: Dr. Sanda Harabagiu

Take-Home Final Exam

Issued: November 26th 2011
Due: December 5th 2011 –in class

Problem 1 (40 points) :

Consider the following web graph:

D1 → D2, D1 → D4, D1 →D6, D1 →D15
D2 → D1, D2 → D3
D3 → D4, D3 → D8, D3 → D9, D3 →D10
D4 → D5, D4 → D11, D4 → D15
D5 → D1, D5 → D7, D5 → D11
D6 → D5, D6 → D15
D7 → D4, D7 → D10
D8 → D1, D8 → D3
D9 → D12, D9 → D13, D9 → D14
D10 → D7, D10 → D9, D10 → D13
D11 → D2, D11 → D6
D12 → D13, D12 → D14
D13 → D2, D13 → D8
D14 → D3, D14 → D12
D15 → D6, D15 → D12

The content of the Web documents are:

D1

house health wealth happiness family wealth Rome Italy health Paris France

D2

medicine biology cells child health

D3

science knowledge wise family

D4

mother girl child family London Rome Italy

D5

singing dancing shopping shopping shopping shopping

D6

fitness Australia gym Italy shoes beach

D7

computers TV internet football

D8

chemistry substance science nature

D9

museum opera singing dancing painting

D10

physics nature Malibu Italy fashion art

D11

mathematics calculus probabilities science

D12

Sydney Australia Milan Italy Paris France wealth health

D13

Malibu Australia Hawaii TV fitness fresh air palm trees

D14

waves surfing physics nature beach

D15

house park beach ocean Sydney Australia

A. (10 points) Compute the page ranks of each of the Web pages from the graph.

B. (10 points) Use the HITS algorithm to compute the hub and authority score of each Web page.

C. (10 points) Use K-Means to cluster the collection of web document in k=3 clusters. List the final clusters and their centroids. How did you decide to stop the clustering process?

D. (10 points) Consider that each cluster obtained at step B represents a different topic. Compute the topic-sensitive ranks of each Web page.

Problem 2 (30 points) :

Use the same text collection as in Problem 1 and generate clusters based on the following hierarchical methods:

A. *(10 points)* Single-link agglomerative clustering
B. *(10 points)* Complete-link agglomerative clustering
C. *(10 points)* Group-average link agglomerative clustering.

Use a similarity measure based on cosine similarity and for each clustering method show: (1) the clusters; (2) the centroids; and (3) the medoids.

Problem 3 (30 points) :

Consider the same document collection as in Problem 1.

A. *(10 points) For the query Q0= "kid Europe" you are told that the following document are relevant: D2, D3, D4, D12 and D15. Use the Rocchio method to expand the query.*
B. *(10 points) Use automatic local analysis to expand the query.*
C. *(10 points) Use automatic global analysis to expand the query.*