CMPE 493 – Introduction to Information Retrieval
Sample Exam

100 minutes; closed book/notes; you can use a calculator

Problem 1:
Below is a portion of a positional index in the format:
term: doc1: <position1, position2,...>; doc2: <position1, position2,...>; etc.

Gates: 1: <3>; 2: <6>; 3: <2,17>; 4: <1>;
IBM: 4: <3>; 7: <14>;
Microsoft: 1: <1>; 2: <1,21>; 3: <3>; 5: >16,22,51>;

The /*k* operator, word1 /*k* word2 finds occurrences of word1 within *k* words of word2 (on either side), where *k* is a positive integer argument. Thus *k* = 1 demands that word1 be adjacent to word2. Describe the set of documents that satisfy the query Gates /2 Microsoft.

Problem 2:
What is the Levenshtein edit distance between these two words:

GOOGLE ←→ YAHOO

Show all your work. What is the edit sequence that transforms one of the words to the other?

Problem 3:
Assume that you have a collection consisting of 1 million documents. The document frequencies of some of the words in the collection are given below:

important 1,000
is 950,000
nice 10,000
the 1,000,000
today 10,000
very 100,000
weather 100

Consider the following two documents with document IDs D1 and D2.

D1: today the weather is very very nice
D2: nice weather is very very very important

Assume the only stopwords are: “is”, “am”, “are”, and “the”.
Compute the similarity between D1 and D2 after stopword removal using:
(a) Jaccard coefficient
(b) Cosine similarity with length normalization using TF-IDF weighting. Use the raw term frequencies for TF, rather than the log-scaled term frequencies. Use log-scaled frequency for IDF.

Problem 4

Suppose that two information retrieval systems represent the queries and the documents as vectors using ***TF-IDF weighting with length normalization***. Given a query the first system ranks the documents in order of increasing Euclidean distance, whereas the second system ranks them in order of decreasing cosine similarity. Which of the two systems produces a better ranking? Explain.

Problem 5:

Aylin is a new user of WebMovies (a web site that streams movies on demand). So far, she has seen and rated two movies (on a scale from 1 to 5).

Score (Aylin, Braveheart) = 4
Score (Aylin, Halloween) = 2

What is the next movie that WebMovies will recommend to her given the following scores in its database:

Score (Mert, Braveheart) = 5
Score (Ada, Braveheart) = 2
Score (Alp, Braveheart) = 1
Score (Mine, Braveheart) = 4
Score (Mert, Halloween) = 3
Score (Ada, Halloween) = 5
Score (Alp, Halloween) = 1
Score (Ceyda, Halloween) = 5
Score (Mert, Gladiator) = 4
Score (Ada, Gladiator) = 1
Score (Ceyda, Gladiator) = 5
Score (Ada, Scream) = 4
Score (Alp, Scream) = 1
Score (Mine, Scream)= 2
Score (Mert, Rainman) = 5
Score (Ada, Rainman) = 4
Score (Alp, Rainman) = 3
Score (Mine, Rainman) = 5
Score (Ceyda, Rainman) = 2

Make sure that you describe your algorithm very carefully: first in general and then, using the specific data set above