Question 1 **20 marks**

a) Give two main differences between database management and information retrieval. *(4 marks)*

b) Assuming Zipf's law holds, where the collection frequency of the ith-most common term equals one tenth of the inverse of i. What is the fewest number of most common words that together account for more than 20 % of word occurrences (i.e. the minimum value of m such that at least 20 % of word occurrences are one of the m most common words). *(8 marks)*

c) Assume that Zipf's law holds perfectly in a collection of 5000 tokens. By perfectly we mean that the constant c involved is 1. Given a collection with exactly 4 words alpha, beta, gamma, delta. The frequency order is *(8 marks)*

frequency(alpha) > frequency(beta) > frequency(gamma) > frequency(delta).

What are the frequencies for these four words?

Question 2 **20 marks**

a) [8 marks] Compute the edit distance between the words HA and THE. Clearly display all values in the table. For each cell include all four values computed. *(8 marks)*

b) When you obtain the edit distance between these two words, backtrack through the table to form the possible shortest-edit distance transformations between HA and THE. Clearly mark the backtracking path on the table. Present the operations involved in the possible transformations from HA to THE. *(6 marks)*

c) The Huffman file compression program produces prefix codes. Apply the Huffman file compression program to compute a prefix code for the following alphabet and frequencies in multiples of thousands: a : 4.2, b: 1.4, c: 4.4 *(6 marks)*

Question 3 **20 marks**

Assume that simple term frequency weights are used (not IDF factors), and the only stop-words are: "is", "am" and "are". Use term frequencies instead of the IDF factors in the formula for cosine similarity, and ignore stop words.

a) Compute the cosine similarity between the following two documents: *(10 marks)*

document 1: morale is truly truly low
document 2: low morale is truly truly truly demoralizing

b) Show the 3-gram (inverted) index constructed for the small dictionary containing only the words "gram", "spam", "cram", and "scram". List the 3-grams alphabetically in a table assuming the word-boundary character ($) is alphabetized after "z" and show the posting lists for each. *(10 marks)*

Question 4 **20 marks**

Consider the following web graph:

Page A points to pages C and E
Page B points to A and E
Page C points to B
Page D points to E

a) Compute the adjacency matrix corresponding to this graph. (*3 marks*)

b) Compute the probability matrix for this graph, where teleporting has a probability of 0.1. (*6 marks*)

c) Say a websurf is definitely starting on page E. Determine a probability vector for this situation. (*1 marks*)

d) Use the probability vector obtained under Question 4 c) to compute an approximation of the page rank score of these pages, using two power iterations only. (*10 marks*)