OLLSCOIL NA hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

2017/2018

Semester 1 - Winter 2017

CS4611 Information retrieval

Professor Omer Rana (extern)
Professor Cormac Sreenan
Professor Michel Schellekens

1.5 hours

Calculators Allowed

Total marks: 80

Answer all Questions

PLEASE DO NOT TURN THIS PAGE UNTIL INSTRUCTED TO DO SO

PLEASE ENSURE THAT YOU HAVE THE CORRECT EXAM PAPER

Question 1 [20 marks]
a) [4 marks] Give two main differences between database management and information retrieval.

b) [8 marks] 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).

c) [8 marks] 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 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 HAS and THIS. Use the table below to compute the values in the last row.

		T	H	I	S
	0	1 \ 1	2 \ 2	3 \ 3	4 \ 4
H	1 \ 1	1 \ 2 2 \ 1	1 \ 3 2 \ 1	3 \ 4 2 \ 2	4 \ 5 3 \ 3
A	2 \ 2	2 \ 2 3 \ 2	2 \ 2 3 \ 2	2 \ 3 3 \ 2	3 \ 4 3 \ 3
S	3 \ 3				

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

c) [6 marks] 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 : 4.6, b: 1.4, c: 4

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) [10 marks] Compute the cosine similarity between the following two documents:

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

b) [10 marks] 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.

Question 4 [20 marks]

Consider the following web graph:

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

a) [3 marks] Compute the adjacency matrix corresponding to this graph.

b) [6 marks] Compute the probability matrix for this graph, where teleporting has a probability of 0.2.

c) [1 mark] Say a websurf is definitely starting on page E. Determine a probability vector for this situation.

d) [10 marks] 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.