OLLSCOIL NA hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

SUMMER EXAMINATIONS 2014

CS4611: Information Retrieval

Professor Barry O'Sullivan
Professor Michel Schellekens
Professor Ian Gent (Extern)

Answer all questions

Total marks: 80

One and a half hour

For your information: 1.125 minutes per mark

PLEASE DO NOT TURN THIS PAGE UNTIL INSTRUCTED TO DO SO
ENSURE THAT YOU HAVE THE CORRECT EXAM PAPER

Question 1

(20 marks)

a) [*4 marks*] How does stemming typically affect recall? Why?

b) [*4 marks*] Give two main differences between database management and information retrieval

c) [*6 marks*] Assuming Zipfs 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 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words).

d) [*6 marks*] Give an example of a sentence that falsely matches the wildcard query rb*rb if the search were to simply use a conjunction of bigrams.

Question 2

(20 marks)

Consider the query: cork floors, where you search a corpus of only 8 documents, with the following retrieval results:

1. Cork Floors and Hardwood Floors
2. Best shop in Cork for all your carpet needs
3. Movie star popped the cork to celebrate victory
4. Cork city of culture

And the following non-retrieved documents:

1. Cork and Kerry draw in hurling contest
2. Cork tiles for home and business
3. Evening Echo voted best Cork paper
4. Simple inline skating tricks

Assume that your user was trying to find information about floors made of cork, not information about Cork city.

a) [*4 marks*] What is the precision and recall for this query and the given results?

b) [*4 marks*] What is the F-measure if we pay twice more attention to precision than recall?

c) [*12 marks*] The relevance of the retrieval has been evaluated by 2 judges the following way (+ means relevant, - means non-relevant):

Judge a: doc 1+ doc 2 + doc 3 + doc 4 -

Judge b: doc 1- doc 2 + doc 3 + doc 4 -

Is there a good agreement between judges? Use the typical measure of agreement we have seen in the course and state your conclusion based on the outcome of the measure.

(*20 marks*)

Question 3

a) [*10 marks*] To compute the cosine similarity formula, use term frequencies instead of the IDF values, and ignore stop words. Assume that the only stop-words are: is, am and are. Compute the cosine similarity between the following two documents, i.e. document 1 and document 2.

document 1: precision is very very high
document 2: high precision is very very very important

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.

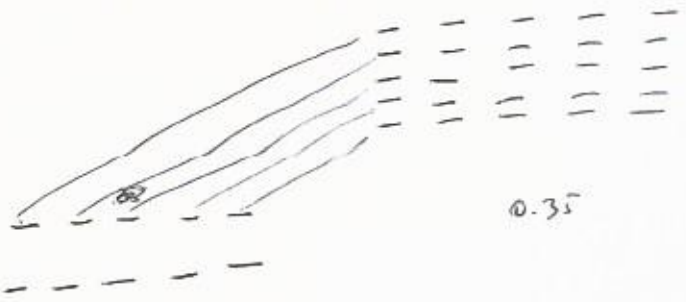
(*20 marks*)

Question 4

Consider the following web graph.

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

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 0.5
c) [*1 marks*] Say a websurf is definitely starting on page C. Determine a probability vector for this situation.
d) [*10 marks*] Use the probability vector obtained under C) to compute an approximation of the page rank score of these pages, using two power iterations only.



PLEASE DO NOT TURN THIS PAGE UNTIL INSTRUCTED TO DO SO

THEN ENSURE THAT YOU HAVE THE CORRECT EXAM PAPER