OLLSCOIL NA hÉIREANN

THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

2016/2017

Semester 1 - Winter 2016

CS4611 Information retrieval

Dr. Helen Purchase (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]
Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = non-relevant, 1 = relevant).

docID	*Judge*1	*Judge*2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

a) [10 marks] Calculate the kappa statistic between the two judges. Show your work leading to the kappa statistic outcome. What is your conclusion based on this outcome?

Let us assume that you've written an IR system that for the query (agreed upon for this information need) returns the set of documents 4, 5, 6, 7, 8. Calculate the following measures for your system, if a document is considered relevant only if the two judges agree:

b) [3 marks] precision

c) [3 marks] recall

d) [4 marks] the F-measure (for $\alpha = \frac{1}{2}$)

Show your work for each part b), c) and d).

Question 2 [20 marks]

a) [10 marks] We are given a random sample of 10,000 documents from a collection containing 1,000,000 documents. We count the different words in this sample, and we find 5,000. Supposing that the collection satisfies Heaps' law with exponent 0.5, derive an estimate of the number of different words you expect to find in the whole collection. Show the work leading to your answer.

b) [10 marks] Assuming Zipf's Law with a corpus independent constant A = 0.1, what is the fewest number of most common words that together account for more than 25% of word occurrences (ie. the minimum value of m such that at least 25% of word occurrences are one of the m most common words). Show your work leading to the answer.

Question 3 [20 marks]

777 in binary code is 1100001001.

a) [4 marks] Compute the gamma code of 777. Show your work.

b) [4 marks] Compute the variable byte code of 777. Show your work.

c) [7 marks] State the cosine similarity formula and provide the formula expressing each component involved in the formula.

d) [5 marks] A company has improved their information retrieval system with a new feature. They decide not to hire a team of judges to carry out on-location testing to support statistical analysis as hiring the judges is too costly. What approach could the company take to test the new feature, ensuring that the cost of the testing is less than hiring judges? Stating the name of the test is not sufficient. Describe the process involved in the testing.

Question 4 [20 marks]
Consider five documents, A, B, C, D and E.

The hyperlinks are as follows:

A points to pages C and D, B points to pages A and C, C points to B and E points to A.

Draw the directed graph corresponding to these hyperlinks.

a) [5 marks] Provide the adjacency matrix for this graph.

b) [8 marks] Construct the probability transition matrix where you use teleportation probability $\alpha = 0.5$ for this graph.

c) [7 marks] From this matrix, determine the first two power iterations as you would normally compute to reach the steady state (you can stop after two iterations, where in the first iteration you multiply the initialisation vector with the transition probability matrix). We initialize as follows: the document from which we start the random walk is C.