OLLSCOIL NA hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

SUMMER EXAMINATIONS 2012

CS4611:Information Retrieval

Professor James Bowen
Professor Michel Schellekens
Professor Ian Gent (Extern)

Answer all questions

One and a half hour
The use of non-programmable calculators is permitted.

Question 1 [*6 marks*]

Write down the entries in the permuterm index and bigram index that are generated by the term **cork**.

Question 2 [*4 marks*]

Are the following statements true or false?

a) [*1 mark*] In a Boolean retrieval system, stemming never lowers precision.

b) [*1 mark*] In a Boolean retrieval system, stemming never lowers recall.

c) [*1 mark*] Stemming increases the size of the vocabulary.

d) [*1 mark*] Stemming should be invoked at indexing time but not while processing a query.

Question 3 [*10 marks*]

Compute the Levenshtein distance matrix for BEAR → BEEF. For this computation, only produce the first two rows of the Levensthein distance (the rows corresponding to the letter B and the letter E, not counting the initialization row) and display for each cell in the matrix, the 3 values computed.

Question 4 [*6 marks*]

Imagine using the Jaccard coefficient for computing a query document score:

$$jaccard(Q, D) = \frac{|Q \cap D|}{|Q \cup D|}$$

where Q and D represent the set of terms included in a query and a document.

a) [*4 marks*] Illustrate that a non relevant result can be ranked higher by the Jaccard coefficient than a relevant result. Let Q =**'open source'**. Identify a document D1 that is relevant for Q and a document D2 not relevant for Q such that

$$jaccard(Q, D1) < jaccard(Q, D2)$$

b) [*2 marks*] Explain why the Jaccard coefficient is not suited to score documents.

Question 5 [*15 marks*]

Consider these documents:

Doc1 a a a e c
Doc2 b b a c c
Doc3 e e d

Compute the tf-idf weights for a, b and c in each document.

Question 6 [*15 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 = nonrelevant, 1 = relevant). Let us assume that you have written an IR system that for this query returns the set of documents $\{4, 5, 6, 7, 8\}$.

docID	Judge1	Judge2
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) [*5 marks*] Calculate the kappa measure between the two judges.

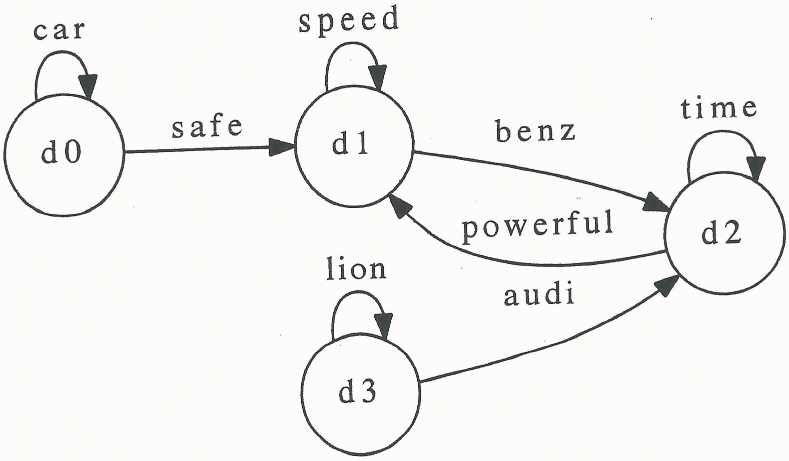
b) [*5 marks*] Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.

c) [*5 marks*] Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

Question 7 [*24 marks*]

Consider the following directed graph, representing hyperlinks on the internet between three given documents, d0, d1, d2 and d3. Construct the probability transition matrix (with teleportation probability $\alpha = 0.1$) for this graph based on the markov chain model. 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). The first power iteration is simply the initialization vector.

We initialize as follows: the document from which we start the model is d2.


car
speed
safe
d0
d1
benz
time
powerful
d2
lion
audi
d3