OLLSCOIL NA hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

Autumn Examinations 2012

CS4611:Information Retrieval

Professor Michel Schellekens

Dr XXX (Extern)

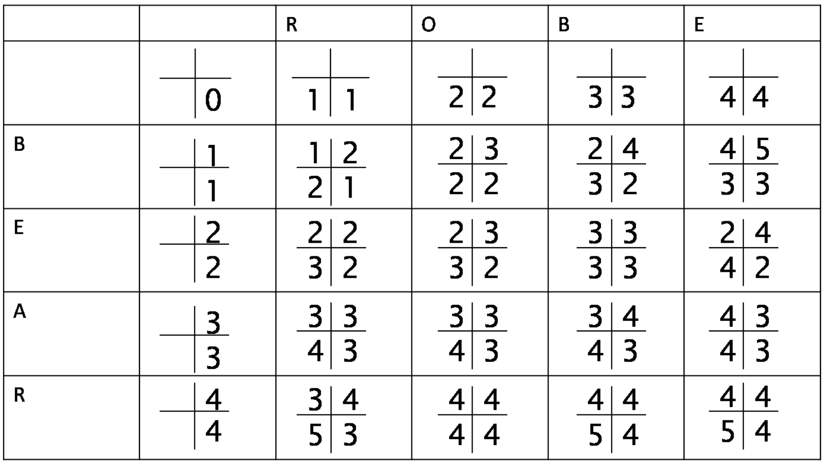
One and a half hour

Question 1 Give an example of a sentence that falsely matches the wildcard query ORA*TOR if the search were to simply use a conjunction of bigrams. (*5 marks*)

Solution: Query: O, OR, RA, TO, OR, R
Then use: OPERATOR (all the bigrams fit it, but the word does not match the template. Note that OR occurs at the end of OPERATOR).

Question 2 Compute Levenshtein distance matrix for BEAR → ROBE. For this computation, display for each cell in the matrix, the 3 values computed. (*15 marks*)

Solution:



		R	O	B	E
	0	1 1	2 2	3 3	4 4
B	1 1	1 2 2 1	2 3 2 2	2 4 3 2	4 5 3 3
E	2 2	2 2 3 2	2 3 3 2	3 3 3 3	2 4 4 2
A	3 3	3 3 4 3	3 3 4 3	3 4 4 3	4 3 4 3
R	4 4	3 4 5 3	4 4 4 4	4 4 5 4	4 4 5 4

Question 3 Question: consider the following measure to evaluate an information retrieval system: the inaccuracy measure is defined as (fp + fn)/(tp + tn + fp + fn), i.e. the sum of false positives and false negatives over the total number of documents.

Is this a good measure of the way the system performs (w.r.t. user satisfaction). In other words, if we keep the inaccuracy of the system (as measured by this measure) low, will we guarantee user satisfaction? (*5 marks*)

Solution: This is not the case. Note that a system in which 99% of the documents are non relevant (which is a typical case), will be deemed to perform well under this measure in case it deems all documents non relevant. In that case, we get that fp = tp = 0, so (fp + fn)/(tp + tn + fp + fn) = fn/(tn + fn). Note: in this scenario: fn is relatively low compared to tn. Indeed, since we do not retrieve any documents, the false negatives are the non retrieved "relevant documents". Most non retrieved documents are not relevant, hence tn is much higher than fn. In other words, the fraction is close to zero in general. On the other hand, user satisfaction is not guaranteed at all.
Hence inaccuracy by this measure is as low as possible, while user satisfaction is not guaranteed at all.

Question 4 Given a document containing terms A, B and C with given frequencies:
A (2), B (4), C (5)
Assume collection contains 20,000 documents and document frequencies of these terms are: A (100), B (880), C (75)
Calculate tf-idf weight for A,B,C in this document. (*10 marks*)

Solution: A: $(1 + log_{10}(2)) * log_{10}(\frac{20000}{100}) = 2.994$
B: $(1 + log_{10}(4)) * log_{10}(\frac{20000}{880}) = 2.173$

C: $(1 + log_{10}(5)) * log_{10}(\frac{20000}{75}) = 4.122$

Question 5 Looking at a collection of web pages, you find that there are 8000 different terms in the first 30,000 tokens and 25,000 different terms in the first 7,000,000 tokens. Assume a search engine indexes a total of 60,000,000,000 (6×10^{10}) pages, containing 400 tokens on average. What is the size of the vocabulary of the indexed collection as predicted by Heaps' law? (*10 marks*)

Solution: $log_{10}(M_1) = log_{10}k + blog_{10}(T_1)$ and $M_1 = 8000$ $T_1 = 30,000$
$log_{10}(8000) = log_{10}k + blog_{10}(30,000)$
$log_{10}(M_2) = log_{10}k + blog_{10}(T_2)$ and $M_2 = 25,000$, $T_2 = 7,000,000$
$log_{10}(25,000) = log_{10}k + blog_{10}(7,000,000)$
thus $log_{10}k \approx 2.9675$,$k \approx 927.897$ and $b = 0.209$
$log_{10}(M) = log_{10}k + 0.209 * log_{10}(60,000,000,000 * 400) = 5.76396$ thus $M = 10^{7.778} \approx 5.8 * 10^5$

Question 6 Compute the variable byte code of 12 and 135 (*4 marks*)
Decode VB code of documents IDs: 000000101001011010010001 (*2 marks*)
Compute the gamma code of 13 (*2 marks*)
Decode gamma code of documents IDs: 11110100011111000101 (*2 marks*)

Solution: $12_2 = 1100$ VB: 10001100 $135_2 = 10000111$ VB: $00000001, 10000111$
$278_2 = 100010110$ and $17_2 = 10001$ thus doc_{278} and doc_{17}
$13_2 = 1101$ gamma code: 1110101
$24_2 = 11000$ gamma code: 111101000 and $37_2 = 100101$ gamma code: 11111000101

Question 7 Consider the following table representing judges'decisions:

		Judge 2		
		Yes	No	Total
Judge 1	Yes	220	50	270
	No	60	23	83
	Total	280	73	353

Compute the Kappa statistic for this judgment (*6 marks*). What is your conclusion on the acceptability of the judgments? (*4 marks*)

Solution:

$$P(A) = \frac{220 + 23}{353} = 0.688$$

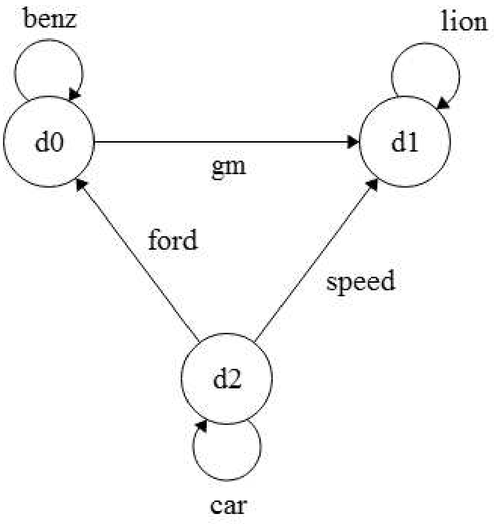
$$P(rel) = \frac{280 + 270}{353 * 2} = 0.779 \quad P(notrel) = \frac{83 + 73}{353 * 2} = 0.221$$

$$P(E) = P(rel)^2 + P(notrel)^2 = 0.779^2 + 0.221^2 = 0.6557$$
$$k = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.688 - 0.6557}{1 - 0.6557} = 0.094$$

Kappa value in the interval $[2/3, 1]$ are seen as acceptable, for this example we need to redesign relevance assessment methodology used etc.

Question 8 Consider the following directed graph, representing hyperlinks on the internet between three given documents, d0, d1 and d2. 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 three power iterations as you would normally compute to reach the steady state (you can stop after three iterations). The first power iteration is simply the initialization vector.

We initialize as follows: the document from which we start the model is d2.(*15 marks*)



Solution:

Link Matrix: $\begin{array}{cccc} & d_0 & d_1 & d_2 \\ d_0 & 1 & 1 & 0 \\ d_1 & 0 & 1 & 0 \\ d_2 & 1 & 1 & 1 \end{array}$

Transition matrix without teleporting: $\begin{array}{cccc} & d_0 & d_1 & d_2 \\ d_0 & 0.5 & 0.5 & 0 \\ d_1 & 0 & 1 & 0 \\ d_2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}$

Transition matrix with teleporting $\alpha = 0.1$: $\begin{array}{cccc} & d_0 & d_1 & d_2 \\ d_0 & 0.4833 & 0.4833 & 0.0333 \\ d_1 & 0.0333 & 0.9333 & 0.0333 \\ d_2 & 0.3333 & 0.3333 & 0.3333 \end{array}$

	$P_t(d_0)$	$P_t(d_1)$	$P_t(d_2)$				
t_0	0	0	1	0.3333	0.3333	0.3333	$\mathrm{d}\vec{P}$
t_1	0.3333	0.3333	0.3333	0.28327	0.58324	0.13329	$\mathrm{d}\vec{P}^2$
t_2	0.28327	0.58324	0.13329	0.200752	0.725669	0.073279	$\mathrm{d}\vec{P}^3$