Term Exam 1 - Information Retrieval - Fall 2014

Date: 29th Sep 2014 **Marks: 50** **Time: 90 mins.**

Name: RollNo:

Attempt the examination on the question paper and write concise answers. Marks would be deducted for superfluous incorrect answers. You may use extra sheets for rough work but DO NOT attach them to the question paper. ***Extra sheets will NOT be checked****. You are not allowed to ask any questions. In case of any ambiguities, state an appropriate assumption and solve the question using that assumption. Good Luck!*

Question	1	2	3	4	5	6	7	8	Total
Marks	/14	/6	/3	/3	/2	/2	/5	/5	/40

1. Some documents and a stop word list is given below.

   Document 1: How much wood would a woodchuck chuck if a woodchuck would chuck wood?
   Document 2: A woodchuck would chuck all the wood he could chuck, if a woodchuck could chuck wood!
   Document 3: A woodchuck would chuck so much wood he would not know how much wood he chucked.
   Stop words: how, would, if, a, all, the, he, could, so, much, not

   a. Tokenize, case-fold, remove stop words, stem, and then provide an inverted index for the documents using only the term frequency. (6 marks)

Term	Doc1	Doc2	Doc3
chuck	2	3	2
wood	2	2	2
woodchuck	2	2	1
know	0	0	1

   b. Change the inverted index to incorporate the tf-idf weighting scheme. Use the simplest definition of idf (i.e. idf = 1/df). This means you should not normalize or use log scaling for calculating the idf. (3 marks)

Term	df	1/df
chuck	3	1/3 = 0.33
wood	3	1/3 = 0.33
woodchuck	3	1/3 = 0.33
know	1	1

Term	Doc1	Doc2	Doc3
chuck	2/3	1	2/3
wood	2/3	2/3	2/3
woodchuck	2/3	2/3	1/3
know	0	0	1

=

Term	Doc1	Doc2	Doc3
chuck	0.66	1	0.66
wood	0.66	0.66	0.66
woodchuck	0.66	0.66	0.33
know	0	0	1

   c. Given the following query and the index above, provide a ranked result list using cosine similarity. Use the inverted index and idf weights calculated in part b to weight the query. (5 marks)
   Query: Would a woodchuck chuck wood

Term	Doc1	Doc2	Doc3	q
chuck	2/3	1	2/3	1/3
wood	2/3	2/3	2/3	1/3
woodchuck	2/3	2/3	1/3	1/3
know	0	0	1	0

|Doc1| = sqrt(4/9+4/9+4/9+0) = 1.15
|Doc2| = sqrt(1+4/9+4/9+0) = 1.37
|Doc3| = sqrt(4/9+4/9+1/9+1) = 1.41
|q| = sqrt(1/9+1/9+1/9+0) = 0.58

sim(Doc1,q) = (2/3.1/3+2/3.1/3+2/3.1/3)/(||Doc1||q||) ≈ 1
sim(Doc2,q) = (1.1/3+2/3.1/3+2/3.1/3)/(||Doc2||q||) ≈ 0.98
sim(Doc3,q) = (2/3.1/3+2/3.1/3+1/3.1/3)/(||Doc3||q||) ≈ 0.68

Ranking: Doc1, Doc2, Doc3

2. Consider two annotations of a single annotator over 20 documents for relevance (R) and non-relevant (N), where the annotations have been performed by two judges.

Document	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Judge 1	R	R	R	R	R	R	R	R	N	R	R	R	R	N	R	R	N	R	R	R
Judge 2	R	R	N	R	R	R	N	R	N	R	R	R	R	N	R	R	N	R	R	R

What is the intra-annotator agreement between these two annotations as measured by the kappa value? What can you tell about the quality of agreement as indicated by the kappa value? Why? (4+1+1=6 marks)

K = (P(A) – P(E))/ (1-P(E))
P(A) = 18 / 20 = 0.9
P(relevant) = (15+15+0 +2) / 40 = 4/5
P(irrelevant) = (0+2+3+3) / 40 = 1/5
P(E) = P(relevant)P(relevant) + P(irrelevant)P(irrelevant) = 16/25 + 1/25 = 17/25 = 0.68
K = 0.687

Which tells us that there's fair agreement between the two annotations according to Krippendoff's interpretation.

3. Given a document collection D of 2,000,000 documents, when a query Q is run, a result set of 200 documents is returned. In a practical experiment. Suppose you manually evaluate that 150 of the documents returned by the search are relevant. (3 marks)
   a. What is the Precision?
      P = 150/200 = ¾ = 0.75
   b. What is the Recall?
      We can't calculate recall since we don't know the TOTAL number of relevant documents in the collection

4. While the F-measure is the harmonic mean of Recall and Precision, the G-measure is the geometric mean and Information Content (IC) corresponds to the arithmetic mean. Compare all three measures for a system with P = 1.0 and R = 0.2. (3 marks)

   F = 2PR/(P+R) = 0.33
   G = sqrt(P.R) = sqrt(.2) = 0.45
   IC = (P+R)/2 = 0.6

5. Why is accuracy not a good measure for information retrieval problems? (2 marks)
   Because in a practical scenario, the number of documents which are neither relevant nor retrieved is very large which brings accuracy very close to 1 even if we do not return any results

1

6. Are the following statements true (T) or false (F)? (2 marks)
   a. In a Boolean retrieval system, stemming never lowers precision. F
   b. In a Boolean retrieval system, stemming never lowers recall. T
   c. Stemming increases the size of the vocabulary. F
   d. Stemming should be invoked at indexing time but not while processing a query. F

7. We have a two-word query. For one term the postings list consists of the following 16 entries:
   [4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 122, 157, 180]
   For the other it is the one entry postings list: [47]. How many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:
   a. Using standard postings lists. (2 marks)

      Applying MERGE on the standard postings list, comparisons will be made unless either of the postings list end i.e. till we reach 47 in the upper postings list, after which the lower list ends and no more processing needs to be done. Number of comparisons = 11

   b. Using postings lists stored with skip pointers, with a skip length of $\sqrt{N}$, where N is the length of the list. (3 marks)

      Using skip pointers of length 4 for the longer list and of length 1 for the shorter list, the following comparisons will be made: 1. 4 & 47 2. 14 & 47 3. 22 & 47 4. 120 & 47 5. 81 & 47 6. 47 & 47 Number of comparisons = 6

8. An information retrieval system returns the following ranked list for a particular query where R represents a relevant document and a blank cell represents a non-relevant document. From the known relevance judgments, you know that there are 8 relevant documents in total. What is the Mean Average Precision (MAP)? (5 marks)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
R		R					R	R						R					R

(1/1 + 2/3 + 3/8 + 4/9 + 5/15 + 6/20 + 0 + 0) / 8 = 0.3899
When a relevant document is not retrieved at all, the precision value in the MAP is taken to be 0. Division by 6 in the answer (0.52) will earn only 2 marks.