National University of Computer & Emerging Sciences, Karachi
Spring-2018 CS-Department
Midterm II (Sol)
April 04, 2018, 11AM – 12Noon

Course Code: CS317	**Course Name: Information Retrieval**
Instructor Name: Dr. Muhammad Rafi	
Student Roll No:	**Section No: GR1**

- Return the question paper.
- Read each question completely before answering it. There are **3 questions and 2 pages.**
- In case of any ambiguity, you may make assumption. But your assumption should not contradict with any statement in the question paper.
- All the answers must be solved according to the sequence given in the question paper.
- Be specific, to the point and illustrate with diagram where necessary.

Time: 60 minutes. **Max Marks**: 40 points

Question No. 1 **[Time: 25 Min] [Marks: 20]**

Answer the following questions briefly using 4-5 lines of answer book. Be precise, accurate and to the point, only answer genuine query in the question. Each question is of 2 marks.

a. How tf*idf weighting scheme assigned weights to most frequent, less frequent and rear terms of documents in vector space model for IR?

The tf-idf weighting scheme assigns term t a weight in document d that is
1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

b. What are the two assumptions for relevance feedback? Explain.

1.The user has to have sufficient knowledge to be able to make an initial query which is at least somewhere close to the documents they desire.
2.The relevant documents are similar to each other's; these documents have high overlapping terms. Ideally the term distribution is similar to relevant examples marked by the user.

c. Discuss the pros and cons of implicit (indirect) vs. explicit (direct) feedbacks.

Implicit(Indirect) feedback	Explicit (Direct) feedback
Implicit (indirect) feedback does not bother user for explicit actions. It is fast and can be possible for large IR system. It is less reliable and possibly introduce a problem of query drift.	Explicit (Direct) feedback requires user to marks document relevant. It is slow process and does not scale to large systems. It is more reliable and generally save from query drift.

d. Why relevance feedback is not popular among web users? Explain.

Relevance Feedback is mainly used to increase recall, but web users are mainly concerned about the precision of the top few results.
Relevance Feedback slows down returning results as you need to run two sequential queries, the second of which is slower to compute than the first. Web users hate to be kept waiting.
Relevance Feedback is one way of dealing with alternate ways to express an idea (synonymy), but indexing anchor text is commonly already a good way to solve this problem.
Relevance Feedback complicates the user interface.
Relevance Feedback is difficult to explain to a common user.

e. Explain how Rocchio's algorithm is used to fetch pages similar to a user's sample provided page (pages like this options)?

In order to use Rocchio's algorithm for retrieving "Find pages like this one". There is only on relevant feedback item and that is the sample page. We can completely ignore the query (q_0), and we do not want to apply negative judgment and hence alpha (α)= gamma(γ) =0; This implies beta (β)= 1 for the modified query.

f. Define what do we mean by "Search Result snippets". How this concept is related to relevance feedback?

A result snippet is a short summary of the retrieved document, which is designed so as to allow the user by reading a small portion of text summary, He can give feedback on the relevance of the document.

g. What is the value of NDCG for a perfect ranking algorithm?

NDCG = DCG/ IDCG is a ratio between discount cumulative gain and ideal discount cumulative gain. For a perfect ranking DCG=IDCG so NDCG would be 1.

h. What is probability ranking principle in IR? Explain.

The Probability Ranking Principle (PRP) state that for a fixed query q, the obvious relevance order of the documents with respect to the query q is to order the documents with decreasing probability of P (R / di,q) from a collection. {Probability of relevance given the query and document}.

i. Every term in the document is a random variable. What does this assumption signify in probability ranking principle?

Every term is coming from a sample space Σ* over a finite symbol set λ. The probabilities are computed using a random variable x=t for a term. The assumption signify that are terms in a language are equal likely and hence it is treated as a random variable.

j. What is meant by (i) empty-document and (ii) by assumption P(empty-doc/R=1) = P(empty-doc/R=0) =0?

An empty document is a document that contains no selected vocabulary word hence it contains zero feature of interest. An empty document has an equal likely chances of part of a relevant or non-relevant collection for a fixed query. The assumption to P(empty-doc/R=1) = P(empty-doc/R=0) is to simplify the computation and the model.

Question No. 2 **[Time: 15 Min] [Marks: 10]**

Consider the following given document-collection and a query.

d1: w1 w2 w4 w6
d2: w1 w2 w7 w3
d3: w8 w5 w4 w5 w6
q: w2 w5 w6

Using the Vector Space Model (VSM) for IR, compute, document vectors using tf*idf weighting scheme, where tf is the term frequency in a document, and idf is log(N/df) {N is total number of documents and df is the document frequency of term t} with * as multiplication of the two factors. Rank the given document w.r.t. query q using cosine function as a distance. Show all intermediates steps of calculations.

	idf	tf-D1	tf-D2	tf-D3	tf-q	tf*idf D1	tf*idf D2	tf*idf D3	tf*idf q
w1	0.18	1	1	0	0	0.18	0.18	0	0
w2	0.18	1	1	0	1	0.18	0.18	0	0.18
w3	0.48	0	1	0	0	0	0.48	0	0
w4	0.18	1	0	1	0	0.18	0	0.18	0
w5	0.48	0	0	2	1	0	0	0.96	0.48
w6	0.18	1	0	1	1	0.18	0	0.18	0.18
w7	0.48	0	1	0	0	0	0.48	0	0
w8	0.48	0	0	1	0	0	0	0.48	0

Vector d1 =<0.18,0.18,0,0.18,0,0.18,0,0> and |d1| = 0.36

Vector d2 =<0.18,0.18,0.48,0,0,0,0.48,0> and |d2| = 0.73

Vector d3 =<0,0,0,0.18,0.96,0.18,0,0.48> and |d3| = 1.10

Vector q =<0,0.18,0,0,0.48,0.18,0,0> and |q| = 0.54

Sim(d1,q) = cos(d1,q) = (d1 X q) / (|d1| x |q|) = 0.73

Sim(d2,q) = cos(d2,q) = (d2 X q) / (|d2| x |q|) = 0.08

Sim(d3,q) = cos(d3,q) = (d3 X q) / (|d3| x |q|) = 0.83

Thus ranking order will be D3, D1 and D2.

Question No. 3	**[Time: 15 Min] [Marks: 10]**

Consider the following given document-collection and a query.

d1: w1 w2 w4 w6
d2: w1 w2 w7 w3
d3: w8 w5 w4 w5 w6
q: w2 w5 w6

Using the probabilistic model for IR, Probability Ranking Principle (PRP) without given relevance, rank these documents using PRP. Show all intermediates steps of calculations in tabular form.

Using the Probability Ranking Principle (PRP) without given relevance, we will rank the document collection is decreasing order of P(R=1/d,q). Each term P(wi) = (N(w) + 0.5) / (N+1)

Words	W1	W2	W3	W4	W5	W6	W7	W8
N(wi)	2	2	1	2	1	2	1	1
P(wi)	0.625	0.625	0.375	0.625	0.375	0.625	0.375	0.375

Now, we want to rank all documents with the probability of relevance with the query. P(R=1/ di,q) for i=1,2 and 3.

RSV for d1= P(R=1/ d1,q) $=^{rank}$ { common terms w2 and w6} = (0.625/0.375) X (0.625/0.375)

$=^{rank}$ 2.78

RSV for d2= P(R=1/ d2,q) $=^{rank}$ { common terms w2 } = (0.625/0.375)

$=^{rank}$ 1.67

RSV for d3= P(R=1/ d3,q) $=^{rank}$ { common terms w5 and w6 } = (0.375/0,625) x (0.375/0,625) x (0.625/0.375)

$=^{rank}$ 0.6

Thus ranking order will be D1, D2 and D3.