National University of Computer & Emerging Sciences, Karachi
Spring-2021 Department of Computer Science
Midterm Exam-II (Sol)
April 19, 2021, 08:15 AM –09:15 AM

Course Code: CS317	**Course Name:** Information Retrieval
Instructor Name / Names: Dr. Muhammad Rafi, Zeshan Khan	
Student Roll No:	**Section:**

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are **3 questions** on **2 pages**.
- In case of any ambiguity, you may make assumption. But your assumption should not contradict any statement in the question paper.

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

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. Give three reasons why relevance feedback has been little used in web search.

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.

b. Describe the assumptions for Rocchio's algorithms for relevance feedback? Why these assumptions are not correct? Explain.

Assumption 1: User know the query. It is absolutely incorrect as we have seen sometimes user do not know the query term or do not know how to spell it (type it).
Assumption 2: Relevance prototype is well-behaved (separate vector spaces for relevant and non-relevant documents. It is also very incorrect as we cannot guarantee that a unique term can only appears in either collection.

c. Define Navigational Query on the web? Give an example.

Navigational queries seek the website or home page of a single entity that the user has in mind, say Lufthansa airlines. In such cases, the user's expectation is that the very first search result should be the home page of Lufthansa.

d. Does negative feedback is of any importance in web-search? Explain.

Negative feedback is very hard to implement in a retrieval system. In any collection there are several classes of documents that do not relate to a given query. Hence there can be many confusing vectors for such feedback, converging it to a relevant space is hard.

e. Illustrate any two limitations of Normalized Discount Cumulative Gain(NDCG)?

Normalized DCG does not penalize for missing documents in the result.
Normalized DCG metric does not penalize for (non-relevant) bad documents in the result.

f. Outline at least two challenges of a modern industry standard web-crawler and suggest one solution to overcome each of them.

The web-sites generally block access through automatic crawlers, the crawler must be polite in accessing these websites.
There can be several issues with physical access, network access and application layer of the web-host, a crawler must be robust to all these problem.
Crawler should be distributed and scalable as it need to access millions of pages per unit time.

g. Define what do we mean by Near Duplicate web-pages? Why this is problematic for a crawler?

Near Duplicate web-pages are those pages which contains more than fifty-percent contents from some other web-page. Computing a duplicate page is straight forward but finding a near duplicate is more computationally expensive. Most industry standard crawler try to use storage optimally and these duplicate pages are rapid consumer of storage but do not give an benefit for the users of information.

h. Discuss some good characteristics of Dynamic Summary of the search results?

Dynamic Summary are query dependent and a user can easily understand whether a document is related or not by skimming the summary text. There are some rapid dynamic summary generation algorithms like KWICS, which are rapid and scalable.

i. What are the three important components of information retrieval system's evaluation?

To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair.

j. Must there always be a break-even point between precision and recall? Either show there must be or give a counter-example.

In a system, Precision(P) = Recall(R) if and only if, False Positive (FP)= False Negative(FN) or True Positive(TP) =0, If in a rank retrieval system if the highest document is not relevant then TP=0 and that is a trivial break-even point. On the other hand, if it is not the case, then the number of FP increases as you go down and the number FN false negative decreases. The at the start of the list fp<fn and at the end of the list fp>fn. Thus there has to be a break-even point the rank list. At the break-even point F1=P=R

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

Consider the partial document collection D= {***d1***: w4 w5 w6 w1; ***d2:*** w3 w2 w1; ***d3:*** w7 w2 w1} and ***q:*** w4 w3 w7; if the following table gives the **tf** and **idf** scores of each term, compute the score of each document against the given query, using cosine of angle between query vector and document vector. Also produce the ranking of the documents against this query. No need to transform vectors into unit vectors.

Word	tf-d1	tf-d2	tf-d3	idf
W1	0.10	0.17	0.12	0.34
W2	0.17	0.21	0.19	0.78
W3	0.23	0.34	0.14	0.81
W4	0.26	0.28	0.29	0.54
W5	0.15	0.65	0.55	0.90
W6	0.31	0.22	0.36	0.62
W7	0.23	0.45	0.27	0.45

First getting the documents vectors with tf*idf weighting as below:

d1 = < (0.10x0.34); 0; 0; (0.26x0.54); (0.15x0.9); (0.31x0.62); 0>
d2 = < (0.17x0.34);(0.21x0.78); (0.34x0.81); 0 ; 0; 0; 0>
d3 = < (0.12x0.34);(0.19x0.78); 0; 0; 0; 0; (0.27x0.45)>
now for query: q =(0; 0; 1; 1; 0; 0; 1> no tf*idf weighting for query. It is most suitable weighting in the absence of query terms tf scores.

Now,
Cos(d1,q) = d1 x q / (|d1|) x |q| = (0.26x.54) / (0.275x 1.7321) = 0.294
Cos(d2,q) = d2 x q / (|d2|) x |q| = 0.342
Cos(d3,q) = d3 x q / (|d3|) x |q| = 0.261

Hence, Ranking is d2, d1, d3

Solution #2

Q=<0x0.34;0x0.78;1x0.81;1x0.54;0x0.90;0x0.62;1x0.45>

Cos(d1,q)=d1xq/(|d1|x|q|)=(0.26x0.54)x(1x0.54)/(0.275x 1.088)=0.253

Cos(d2,q)=d2xq/(|d2|x|q|)=(0.34x0.81)x(1x0.81)/(0.325x 1.088)=0.630

Cos(d3,q)=d3xq/(|d3|x|q|)=(0.27x0.45)x(1x0.45)/(0.196x 1.088)=0.256

Hence, Ranking is d2, d3, d1

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

The following list of Rs and Ns represents relevant (R) and non-relevant (N) returned documents in a ranked list of 10 documents retrieved in response to a query from a collection of 1,000 documents. The top of the ranked list (the documents the system thinks are most likely to be relevant) is on the left of the list. This list shows 4 relevant documents. Assume that there are 6 relevant documents in total in the collection for the given query.

R R N N R N R N N N

a. What is the precision of the system on the top 6?

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Precision = \frac{3}{3+3} = 0.5$$

b. What is the F1 on the top 6?

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Recall = \frac{3}{3+3} = 0.5$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$F1 - Score = 2 * \frac{0.5 * 0.5}{0.5 + 0.5} = 0.5$$

c. Illustrate the first point where precision drop.

$$Precision_1 = \frac{1}{1} = 1$$

$$Precision_2 = \frac{2}{2} = 1$$

$$Precision_3 = \frac{2}{3} = 0.67$$

The precision drops at 3rd point.

d. What is the largest possible MAP that this system could have? Justify your answer.

$$Largest\ MAP = \frac{\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{7} + \frac{5}{11} + \frac{6}{12}\right)}{6} = 0.68$$

The next two relevant documents that are not listed in the top 10 documents can occur at 11 and 12th rank in the best case.

e. What is the smallest possible MAP that this system could have? Justify your answer.

$$Smallest\ MAP = \frac{\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{7} + \frac{5}{999} + \frac{6}{1000}\right)}{6} = 0.53$$

The next two relevant documents that are not listed in the top 10 documents can occur at 999 and 1000^{th} rank in the worst case.

BEST OF LUCK