National University of Computer & Emerging Sciences, Karachi
Spring-2020 CS-Department
Final Examination
June 22, 2020 (Time: 9AM – 12:30PM)

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

- This is an offline exam. You need to produce solutions as a single pdf and need to upload at slate.
- Read each question completely before answering it. There are **6 questions and 4 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/code where necessary.

Time: 210 minutes+ 30 min. for submission **Max Marks**: 100 points

Basic IR Concepts / IR Retrieval Models	
Question No. 1	**[Time: 30 Min] [Marks: 20]**

a. Answer the following questions to the point. Not more than 5 lines of text. [2x5]

1. What are some of the limitations of Boolean Retrieval Model in information Retrieval(IR)?
2. In practical Implementation of Vector Space Model (VSM), what is the major problem you have observed? illustrate.
3. What is the important factor that can result in false negative match in a VSM?
4. From a Human Language standpoint, what is the major drawback of VSM for IR?
5. In Probabilistic Information Retrieval- what do we mean by the assumption "If the word is not in the query, it is equally likely to occur in relevant and non-relevant"? – explain

b. Consider the partial document collection D= {***d1***: w1 w2 w4 w1; ***d2:*** w3 w2 w6; ***d3:*** w1 w2 w7} and ***q:*** w4 w3 w7; if the following table gives the **tf** and **idf** score of each term, compute the score of each document against the given query ***q*** (assume query use simple $\mathbf{tf}_q$ scores), using cosine of angle between query vector and document vector. Give vector representations of documents vector and query. Also produce the ranking of the documents against this query. [10]

Word	**tf-d1**	**tf-d2**	**tf-d3**	**idf**
W1	0.34	0.17	0.12	0.14
W2	0.12	0.29	0.19	0.38
W3	0.23	0.33	0.14	0.51
W4	0.26	0.28	0.22	0.24
W5	0.15	0.66	0.15	0.60
W6	0.31	0.22	0.16	0.32
W7	0.23	0.45	0.21	0.15

Evaluation in IR / Relevance Feedback	
Question No. 2	**[Time: 30 Min] [Marks: 20]**

a. Why Precision and Recall together not a very good evaluation scheme for IR? Justify in term of system development of IR perspective. [5]
b. What is a Break-Even Point in IR Evaluation? Must there always be a break-even point between precision and recall? Either show there must be or give a counter-example. How break-even point related to the value of F1? [5]
c. The following list of Rs and Ns represents relevant (R) and non-relevant (N) returned documents in a ranked list of 12 documents retrieved in response to a query from a collection of 1000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 5 relevant documents. Assume that there are 8 relevant documents in total in the collection. [10]

   R R N N R N N N R N N R

   1) What is the precision of the system on the top 12?
   2) What is the F1 on the top 12?
   3) Assume that these 12 documents are the complete result set of the system. What is the MAP for the query?
   4) What is the largest possible MAP that this system could have?
   5) What is the smallest possible MAP that this system could have?

Text Classification	
Question No. 3	**[Time: 30 Min] [Marks: 15]**

Consider the following examples for the task of text classification [5+5+5]

	docID	words in document	in $c = China$?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

i. Using the k-Nearest Neighbors (KNN) with k=3 identify the class of test instance docID=5?
ii. Using the Rocchio's algorithm, classify the test instance docID=5?
iii. Using the Multinomial Naïve Bayes to estimate the probabilities of each term (feature) that you use to classify the test instance docID=5?

Text Clustering	
Question No. 4	**[Time: 30 Min] [Marks: 15]**

a. Consider a collection of overly simplified documents d1(1,4); d2(2,4); d3(4,4); d4(1,1); d5(2,1) and d6(4,1). Apply k-means algorithm using seeds d2 and d5. What are the resultant clusters? What is the time complexity? How do we know that this result is optimal or not? [5]
b. Consider a collection of overly simplified documents d1(1,4); d2(2,4); d3(4,4); d4(1,1); d5(2,1) and d6(4,1). Apply HAC using single link. What are the resultant clusters? What is the time complexity? How do we know that this result is optimal or not? [5]
c. Would you expect the same results in part (a) and part (b) of this question? Why these results are different (if they are)? What they represent from the possibility of clustering arrangements? [5]

Web Search & Crawler	
Question No. 5	**[Time: 30 Min] [Marks: 15]**

a. What are the different types of users queries on the web? Give example of each type of the query (Note: other than the textbook example). [5]
b. Identify why these properties are essential (must / should) for a web crawler? Give one problem and one solution for each one: [5]
    i. Politeness
    ii. Freshness
    iii. Extensible
c. Why it is better to partition hosts (rather than individual URLs) between nodes of a distributed crawl system? Suggest an architecture for handing both URLs and Host in a crawler. (draw the diagram and explain its working). [5]

Link Analysis	
Question No. 6	**[Time: 30 Min] [Marks: 15]**

a. Consider a subgraph of web represented by a collection of 6 pages (namely A, B, C, D, E, and F), first draw a pictorial representation of this graph along with adjacency matrix. Is it always possible to follow directed edges (hyperlinks) in the given web graph from any node (web page) to any other? Justify it. [5]

b. Using the adjacency matrix from part (a), Using A and H as column vectors for Hub and Authority, apply HITS algorithm for two iterations to identify at least one hub and one authority page from the collection. [5]

c. Using the adjacency matrix from part (a), Assume that the PageRank values for any page p_i at iteration 0 is $\mathbf{PR(p_i)} = 1$ and that the damping factor for iterations is d = 0.85 Perform the PageRank algorithm and determine the rank for every page after 2 iterations. [5]

Closing Remarks:

You need to prepare a pdf file of all the question as per the question ordering. The orientation should be portrait for each page. It should be clearly visible for each and every text written on the page. You suppose to upload it on Slate as an assignment submission. You have good 30 minutes for it. Wish you all the best.