National University of Computer & Emerging Sciences, Karachi
Spring-2020 CS-Department
Final Examination (Sol)
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)?

   The main limitation of BM is exact retrieval based on exact matching of query terms. Other problems in BM is query formation it is very hard to come up with a Boolean expression for some complex query. The results of BM are flat that is no ranking in it.

2. In practical Implementation of Vector Space Model (VSM), what is the major problem you have observed? illustrate.

   In VSM for large documents we have possible a large number of features. Representing these features with weighting like (tf*idf) gives very smaller values, thus finding similarity via cosine is also very small values (underflow of values may occur). It is one of the very basic implementation challenge for VSM.

3. What is the important factor that can result in false negative match in a VSM?

   Semantic sensitivity- that is document with similar context but different term vocabulary might result in a “false negative” match.

4. From a Human Language standpoint, what is the major drawback of VSM for IR?

   From a Human Language standpoint – the word order is very important and VSM lost this order in representation.

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

In PIR, the assumption about a word that is not in the query, that this work has an equal likely to appears in relevant and non-relevant documents. Simplify the derivation for the formula as it will cancel the numerator and denominator of the equal likely odds from the formula.

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 **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

First getting the documents vectors with tf*idf weighting as below:
d1 =< 0.04,0.04,0.11,0.06,0.09,0.09,0.03>
d2 =< 0.02,0.11,0.16,0.06,0.39,0.07,0.06>
d3 =< 0.01,0.07,0.07,0.05,0.09,0.05,0.03>
q =(0; 0; 1; 1; 0; 0; 1> no tf*idf weighting for query.

Cosine(d1,q) = 0.213
Cosine(d2,q) = 0.303
Cosine(d3,q) = 0.155

Ranking d2,d1, and d3

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]

An information retrieval system can be built with either high precision and high recall value, if the system always return only 1 relevant document it precision will be 1. On the other hand, if the system returns all documents for every query it recalls will be 1. Hence both precision and recall not good to report on information retrieval systems.

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]

In a system, Precision(P) = Recall® 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 pf 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

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?

From the given information, we can see that tp=5; fp=7 and fn=7-5=2 so for precision we have Precision = tp / (tp+fp) = 5/12 =0.416

2) What is the F1 on the top 12?

Let’s find recall for F1: we know Recall = tp/ (fn+tp) = 5/7= 0.714 hence
F1= 2 X (Precision * Recall) / (Precision + Recall)
F1= (2X0.416X0.7146) / (0.416+0.741) = 0.616 / 1.157 = 0.532

3) Assume that these 12 documents are the complete result set of the system. What is the MAP for the query?

The MAP possible the current ranking from the system is:
MAP = 1/5 * (1/1+2/2+3/5+4/9+5/12) = 0.6922

4) What is the largest possible MAP that this system could have?

The maximum MAP possible when the remaining two relevant documents retrieved next to these 12 documents.
MAP = 1/8 * (1/1+2/2+3/5+4/9+5/12+6/13+7/14+8/15) = 0.619

5) What is the smallest possible MAP that this system could have?

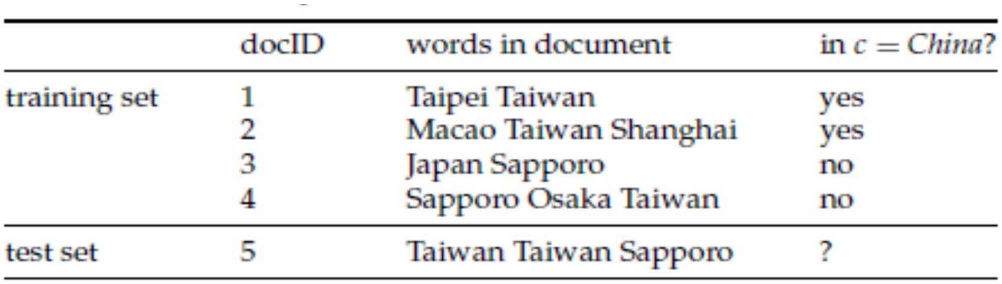
The minimum MAP possible when the remaining four relevant documents found as the last documents from the collection.

MAP = 1/8 * (1/1+2/2+3/5+4/7+5/997+6/998+7/999+8/1000) = 0.399

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?

Dictionary Order = < japan,macao,osaka,sapporo,shanghai,taipei, taiwan>
d1= <0,0,0,0,0,1,1> d2=<0,1,0,0,1,0,1> d3=<1,0,0,1,0,0,0> d4=<0,0,1,1,0,0,1>
d5=<0,0,0,1,0,0,2>
distance |d1,d5| = √3 distance |d2,d5| = √4 distance |d3,d5| = √5
distance |d4,d5| = √2

d5 closest neighbours are d4,d1,d2,d5 hence d5 belong to class **c**.

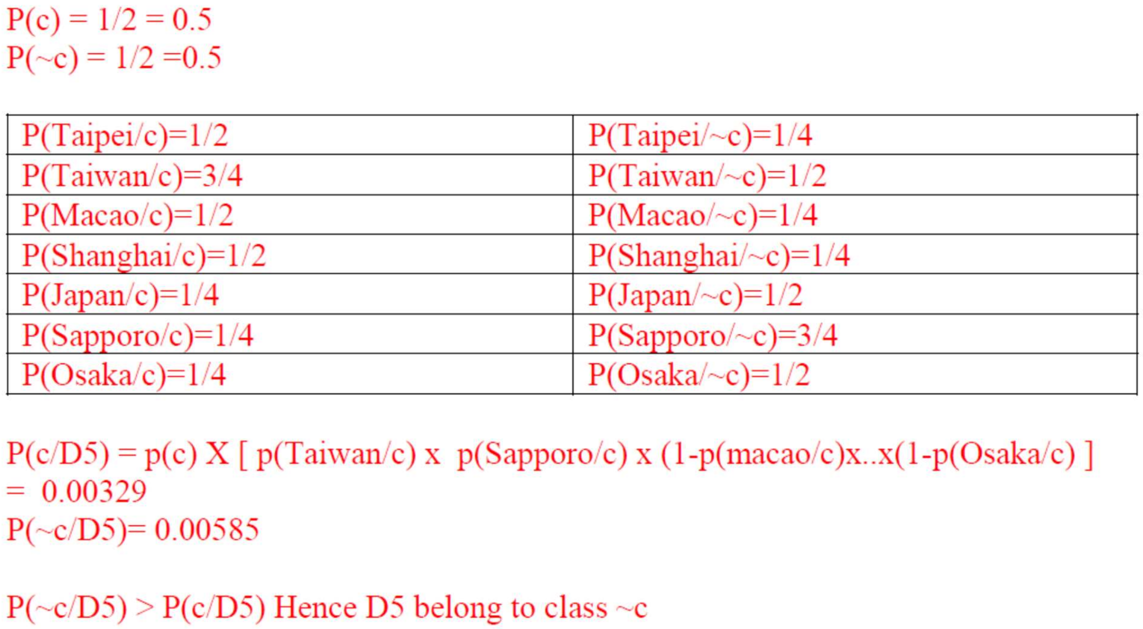
ii. Using the Rocchio's algorithm, classify the test instance docID=5?

Consider again

d1= <0,0,0,0,0,1,1> d2=<0,1,0,0,1,0,1> d3=<1,0,0,1,0,0,0> d4=<0,0,1,1,0,0,1>
d5=<0,0,0,1,0,0,2>
class c = ½ |d1 d2| = <0,1/2,0,0,1/2,1/2,1/2,1>
class ~c = ½ |d3 d4| = <1/2,0,1/2,1,0,0,1/2>

distance |d5,c| = γ2
distance |d5,~c| =√2.75 hence d5 belong to class **c**.

iii. Using the Multinomial Naïve Bayes to estimate the probabilities of each term (feature) that you use to classify the test instance docID=5?



P(c) = 1/2 = 0.5
P(~c) = 1/2 =0.5

P(Taipei/c)=1/2	P(Taipei/~c)=1/4
P(Taiwan/c)=3/4	P(Taiwan/~c)=1/2
P(Macao/c)=1/2	P(Macao/~c)=1/4
P(Shanghai/c)=1/2	P(Shanghai/~c)=1/4
P(Japan/c)=1/4	P(Japan/~c)=1/2
P(Sapporo/c)=1/4	P(Sapporo/~c)=3/4
P(Osaka/c)=1/4	P(Osaka/~c)=1/2

P(c/D5) = p(c) X [p(Taiwan/c) x p(Sapporo/c) x (1-p(macao/c)x..x(1-p(Osaka/c)]
= 0.00329
P(~c/D5)= 0.00585

P(~c/D5) > P(c/D5) Hence D5 belong to class ~c

Text Clustering

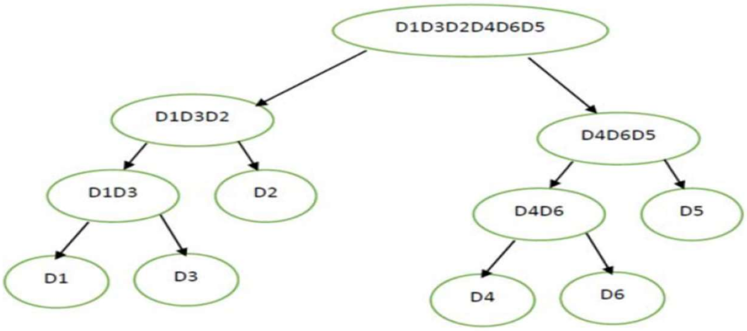
Question No. 4 **[Time: 30 Min] [Marks: 15]**

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]

Let C1=d2 and C2 = d5
Starting C1 and C2 as initial clusters, we need to decide about the membership for each of the documents.
For d1: Dist(C1, d1) < Dist(C2,d1) => d1 belongs to C1.
For d3: Dist(C1, d3) < Dist(C2,d3) => d3 belongs to C1.
For d4: Dist(C1, d4) > Dist(C2,d4) => d4 belongs to C2.
For d6: Dist(C1, d6) > Dist(C2,d6) => d6 belongs to C2.
Hence d1, d2 and d3 are in C1 cluster, and d4, d5 and d6 are in C2 cluster.

K-mean coverage to a local minimum, in order to find optimal clustering one need to produce all possible clustering arrangement. The running time of K-mean is O(n*k*d), where n is the number of documents, k is the number of clusters and d is the number of features.

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]



HAC produced overlapping (hierarchical)n clusters. The resultant clusters are overlapping sub-clusters as above. It time complexity in O (n^3), if carefully implemented it can be kept as quadratic. It cannot guarantee he optimal solution.

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]

No. The results must be different as one is partition clustering and other is hierarchical clustering. They only represent a fraction of the possibility of number of clustering arrangements through Catalan Number.

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]

Informational queries seek general information on a broad topic, such as leukemia or Provence. There is typically not a single web page that contains all the information sought; indeed, users with informational queries typically try to assimilate information from multiple web pages. Query: what is open wound?

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. Query: Home page setting

Transactional query is one that is a prelude to the user performing a transaction on the Web – such as purchasing a product, downloading a file or making a reservation. In such cases, the search engine should return results listing services that provide form interfaces for such transactions. Query: purchase air ticket for air blue

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

Politeness is a must property for modern crawlers, Web servers have both implicit and explicit policies regulating the rate at which a crawler can visit them. These politeness policies must be respected. The problem is when the crawler is not obeying these policies there are many stalls, and it further delays the crawl processes.

ii. Freshness

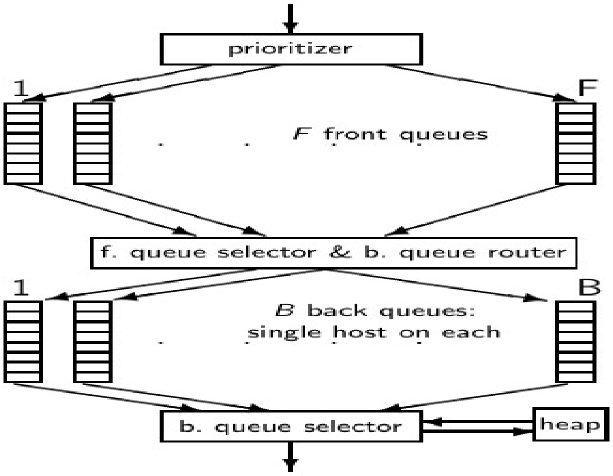
Freshness is a desirable property of modern crawlers. In many applications, the crawler should operate in continuous mode, it should obtain fresh copies of previously fetched pages. A search engine crawler, for instance, can thus ensure that the search engine's index contains a fairly current representation of each indexed web page. For such continuous crawling, a crawler should be able to crawl a page with a frequency that approximates the rate of change of that page. Otherwise it would not be able to get the updated page for indexing.

iii. Extensible

The modern crawlers should be Extensible. Crawlers should be designed to be extensible in many ways – to cope with new data formats, new fetch protocols, and etc. This demands that the crawler architecture be modular and 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]

The crawler must establish connection to Host in order to get the required webpages. It will be efficient to fetch all pages once the connection of the required host is established, hence sorting and maintaining a queue of each Host for the pages that need to be crawl save a good amount of the time. A The URL frontier maintains such an order.

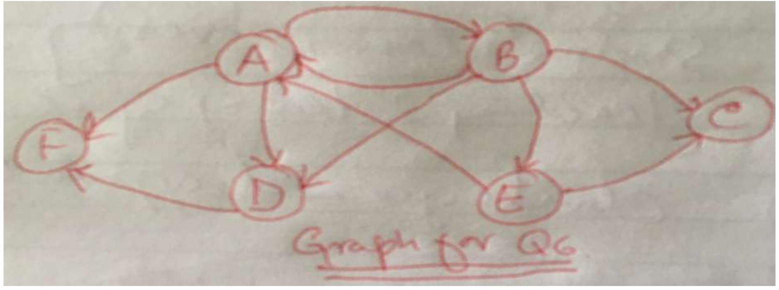


Link Analysis

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 as in Bow-Tie Model? Justify it. [5]

Page A is connected to B, D and F   Page B is connected to A,C,D and E
Page D is connected to F   Page E is connected to A and C



This graph is very similar to Bow-Tie Model and hence all the finding of the model can be validated easily.

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]

Let A be the connectivity matrix for the given graph from part a. We know that

$h^1 = A \cdot a^0$ and $a^1 = A^T \cdot h^0$, similarly $h^2 = A \cdot a^1$ and $a^2 = A^T \cdot h^1$

$$A= \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad a^0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad h^0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$\mathbf{a^2 = < 5/\sqrt{70}, 6/\sqrt{70}, 0, 0, 3/\sqrt{70}, 0>}$ $\mathbf{h^2 = < 3/\sqrt{41}, 2/\sqrt{41}, 3/\sqrt{41}, 1/\sqrt{41}, 3/\sqrt{41}, 3/\sqrt{41}>}$

Best Hub is B; Best Authority is A

c. Using the adjacency matrix from part (a), Assume that the PageRank values for any page p_i at iteration 0 is **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]

Adjacency matrix

A=	0	1	0	0	1	1
	1	0	1	1	1	0
	0	0	0	0	0	0
	0	0	0	0	0	1
	1	0	1	0	0	0
	0	0	0	0	0	0

A=	0	1/3	0	0	1/3	1/3	d	0.85
	1/4	0	1/4	1/4	1/4	0		0.85
	0	0	0	0	0	0		0.85
	0	0	0	0	0	1		0.85
	1/2	0	1/2	0	0	0		0.85
	0	0	0	0	0	0		0.85

Iteration1:	Iteration2:
0.84	0.51
0.85	0.63
0	0
0.85	0
0.85	0.42
0	0

Hence A= 0.51 B=0.63 C=0 D=0 E=0.42 F=0

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. <**The End**>