National University of Computer & Emerging Sciences, Karachi

Spring-2019 CS-Department
Final Examination (Sol)
May 17, 2019- Time(9AM – 12Noon)

[redacted]

Student Roll No:	Section No:

□ Return the question paper.
□ Read each question completely before answering it. There are 8 questions and 3 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: 180 minutes. Max Marks: 100 points

Basic IR Concepts

Question No. 1 [Time: 20 Min] [Marks: 10]

a. What is stemming? What are the advantages of having stemming in IR pipeline? [2]

Stemming is a heuristic- rule based approach, generally fast and use a single term to generate the equivalent tokens although unreadable. It can reduce the size of vocabulary hence index size is reduced.

b. What do we mean by Query Expansion? What are Local and Global approaches to query expansion? [2]

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations.

Query expansion involves techniques such as:
Finding synonyms of words, and searching for the synonyms as well.
Finding all the various morphological forms of words by stemming each word in the search query.
Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results.
Re-weighting the terms in the original query.

Local Query Expansion: When the query is expanded by using local information, like the results returned to the user, it feedback to it. we called it local query expansion

Global Query Expansion: When query is expanded by using external resources like query log, external thesaurus, and external knowledge-base we call it Global query expansion.

c. Give a logical situation where user initiate a Wild Card Query? [2]

In information retrieval often users are uncertain about the spelling of a query term e.g. Sydney or Sidney which can leads to a wildcard query of the form S*dney.

d. What is a major disadvantage of having Permuterm Index? [2]

The main disadvantage of having permuterm index for processing wildcard queries is the size of the index it is often very large for some reasonable collections.

e. How do we estimate the index size for an IR Collection? [2]

In Information Retrieval system the index size is often estimated by using Heaps’ Law. The law state that the size of collection is approximately equal to the number of token is a log-log space. The equation is �� = ����$^{\square}$

Question No. 2 [Time: 20 Min] [Marks: 10]

Consider a term document matrix with frequency of each term in the document below:

[redacted]

Query (q)	1	1	1	1	1

Using Vector Space Model (VSM) find the similarity of each document with the given fixed query. You can use term frequency as coefficient of vectors. Cosine (di, q) = (di . q) / (|di| x | q|)

The document vectors of the collection using term frequency can be given as below: d1 = <10,8,0,2,1> assuming the term subscripts dictate the dictionary order of the terms. |d1| =13 d2 = <0,0,9,9,8> and |d2| = 15.03

d3 = <2,2,4,4,6> and |d3| = 8.71

q = <1,1,1,1,1> and |q| = 2.23

Cosine (d1,q) = 21 / (13 * 2.23) = 0.72

Cosine (d2,q) = 27 / (15.03 * 2.23) = 0.80

Cosine (d3,q) = 18 / (8.71 * 2.23) = 0.92

Question No. 3 [Time: 20 Min] [Marks: 15]

a. Why Precision and Recall are not considered a good evaluation metric for IR systems? [5]

An IR system can alter the values of precision and recall. We can get 100% precision while returning only one related document to a given query. Similarly, we can always get 100% recall by just returning all documents for a given query. Hence, precision and recall alone cannot be a good measure for IR system.

b. A user presented a collection of document against a given fixed query q in the following order to system relevance {D4, D3, D6, D5, D2, D1}. The gold standard of relevance on scale of 0-3, where 0 means not relevant, 3 means highly relevant, 1 and 2 means marginally relevant in between, described as { D1, D3, D4, D6, D2, D5 } □ { 3,3,2,1,1,0}. You are required to compute Cumulative Gain (CG) and Discount Cumulative Gain (DCG). Use logarithmic scale (Relevance $_i$ / log$_2$ (i+1)) [5+5]

The cumulative gain of the given query is computed as ∑ ������ �� □ ◌1◌1৭ so 2+3+1+0+1+3 = 10 hence the CG$_6$ = 10

For discount cumulative gain

i	Rel $_i$	Log$_2$(i+1)	Rel $_i$ / log$_2$(i+1)
1	2	1	2
2	3	1.58	1.89
3	1	2	0.5
4	0	2.32	0
5	1	2.58	0.38
6	3	2.80	1.07

Now discount cumulative gain is computed as ∑ఠ◌π□□ □ □ ◌1◌1৭ □□□ଣ (◌।◌।৭) = 2+1.89+0.5+0+0.38+1.07 = 5.84

hence DCG$_6$ = 5.84

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

Suppose that a user's initial query is q= w1 w3 w2 and IR systems return four documents. User selected d1= w2 w3 w4, d2= w3 w3 w4 w5 and d4= w1 w3 w4 w1 as relevant. While d3= w2 w4 w5 w3 as non-relevant to her query. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback algorithm to get modify query vector (optimal) after relevance feedback? Rocchio equation is given below.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

q= < 1, 1, 1, 0, 0>
d_1= < 0, 1, 1, 1, 0>
d_2= < 0, 0, 2, 1, 1>
d_3= < 0, 1, 1, 1, 1>
d_4= < 2, 0, 1, 1, 0>

Using the given equation, we will get,

q_m= α * < 1, 1, 1, 0, 0> + β* 1/3 *{ < 0, 1, 1, 1, 0> + < 0, 0, 2, 1, 1> + <2,0,1,1,0>} – γ *{< 0, 1, 1, 1, 1> }

q_m= α * < 1, 1, 1, 0, 0> + β* 1/3 *{ < 2, 1, 4, 3, 1> } – γ *{< 0, 1, 1, 1, 1> } q_m= α * < 1, 1, 1, 0, 0> + β* 1/3* <2,1,4,3,1> – γ * {< 0, 1, 1, 1, 1> } q_m= < α + 2β/3 , α + β/3- γ , α + 4β/3- γ, β- γ, β/3- γ >

We need to put zero on all the dimensions where we have identifiable negative values:

q_m= < α + 2β/3, α + β/3- γ, α + 4β/3- γ, β- γ, β/3- γ >

There is no need to choose values for α, β and γ.

Text Clustering

Question No. 5 [Time: 15 Min] [Marks: 10]

a. Give at least 4 ways in which, you can possibly terminate K-Means algorithm. [5]

We can apply one of the following termination conditions for K-Means:

1. A fixed number of iterations.
2. Assignment of documents to clusters (the partitioning function γ) does not change between iterations.
3. Centroids ~μ_k do not change between iterations.
4. Terminate when RSS falls below a threshold.

b. Consider the following set of data points D ={d1(1,2), d2(2,2), d3(4,2), d4(1,1), d5(2,1), d6(4,1)} taking d2 and d5 what will be the final clusters using K-Means. Is this solution optimal? [5]

Starting d2 and d5 as initial seeds, we need to decide about the membership for each of the documents.
For d1: Cos(d2,d1) > Cos(d5,d1) => d1 belongs to d2.
For d3: Cos(d2,d3) < Cos(d5,d3) => d3 belongs to d5.
For d4: Cos(d2,d4) > Cos(d5,d4) => d1 belongs to d2.
For d6: Cos(d2,d6) < Cos(d5,d6) => d6 belongs to d5.

Hence d1,d2 and d4 are in one cluster, and d5,d3 and d6 are in other. K- mean is a greedy algorithm it is never guarantee that an optimal solution is produced.

Text Classification

Question No. 6 [Time: 15 Min] [Marks: 10]

Consider the following examples for the task of text classification

Dataset DocID Features- Words in documents Class Fruit=Yes/No

[redacted]

7	Orange, Red, Lemon	?

a. Using the training data first calculate the class prior probabilities? [4]

P(Fruit=Yes) = 3/5 = 0.6
P(Fruit=No) = 2/5 = 0.4

b. Using Multinomial Naïve Bayes to estimate the probabilities of each term (feature), that you will be using for doing part c? [4]

P(Orange/Fruit) 1/6 P(Orange/~Fruit) 4/15
P(Mango/Fruit) 1/9 P(Mango/~Fruit) 1/15
P(Melon/Fruit) 1/9 P(Melon/~Fruit) 1/15
P(Red/Fruit) 1/18 P(Red/~Fruit) 2/15

P(Lemon/ Fruit)	1/18	P(Lemon/~Fruit)	2/15

c. Apply the Multinomial Naïve Bayes to classify the given test instance? [2]

P(d6/Fruit) 0.6 * (1/6) * (1/9) * (1/9) = 0.001

P(d6/~Fruit)	0.4 *(4/15)*(1/15)*(1/15) = 0.0004

Document d6 belongs to class Fruit=Yes.

P(d7/Fruit) 0.6 * (1/6) * (1/18) * (1/18) = 0.0003

P(d7/~Fruit)	0.4 *(4/15)*(2/15)*(2/15) = 0.001

Document d7 belongs to class Fruit=No.

Web Search & Crawling

Question No. 7 [Time: 25 Min] [Marks: 15]

a. What are the different types of users queries on the web? Give example of each type of the query. [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.

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.

A 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

b. Differentiate between Precision critical tasks vs Recall critical tasks. Give an example of each. [5]

Precision Critical Tasks	Recall Critical Tasks
Time matters a lot in these tasks	Time matter less
Tolerance to missed documents	Non-tolerance to missed documents
Redundant resources	Less redundant very few resources
Example: web search	Example: legal/patent search

c. What do we mean by Robustness of an Industry Scale Web Crawler? Give an example. [5]

The Robustness of a crawler is to deal with all the different types of traps that web servers generally deployed against crawling of their contents. The Web contains servers that create spider traps, which are generators of web pages that mislead crawlers into getting stuck fetching an infinite number of pages in a particular domain. Crawlers must be designed to be resilient to such traps. Not all such traps are malicious; some are the inadvertent side-effect of faulty website development.

Link Analysis

Question No. 8 [Time: 30 Min] [Marks: 15]

a. Outline at least 4 differences between HITS and PageRank algorithms for Link Analysis. [5]

b. HITS PageRank

It gives two scores Hub and Authority for

It gives one score per page.

each page.

It is executed at query time. It is precomputed at indexing time. It is query dependent. It is independent from query. It is not robust against web/link spams It is robust against web-spams

Never favours pages, but can be manipulated for higher scores.	It favours old pages. It can also be manipulated.

c. Consider a web graph with five nodes 1, 2, 3, 4 and 5. The links are as follows: 1 → 2,1 → 4, 1 →5, 3 → 2, 3 → 5,4 → 5, 4 → 2. Using a and h as column metrics. Produce two iterations of HITS algorithm and updates on a and h. Identify one page as the best hub and authority. Use L2 normalization for both vectors of a and h. A is the adjacency matrix for the given web graph. [10]

$$\begin{aligned} \vec{h} &\leftarrow A\vec{a} \\ \vec{a} &\leftarrow A^T\vec{h}, \end{aligned}$$

Let A be the connectivity matrix for the given graph. We know that $h^1 = A \cdot a^0$

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

A =	0 10 00 10 10 0 0	$A^T =$	0 0 1 1 1 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1	

a^1 <0,3,0,1,3> normalizing we get <0, 0.15,0, 0.05, 0.15> h^1 <3,0,2,2,0> normalizing we get <0.27, 0, 0.18,0.18, 0> a^2 <0, 0.63,0,0.27,0.63>

h^2	<0.4,0,0.33,0.33,0>

Best Hub is n1; Best Authority is n5

< The end>