BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

First Semester 2016/2017

Course 539: Special Topics: Information Retrieval and Web Search

Instructor: Dr Adnan Yahya. **Midterm Exam**

Time: 90 minutes max

Please answer the following questions using the exam sheets only.

Question	Q1	Q2	Q3	Q4	Q5	Q6	Total
ABET outcome	e	a	e	a	B		
Max grade	20	16	15	18	16	18	103/100
Earned							

Question 1 (20%): Consider that our document collection S has the following documents: D1;D2,D3;D4:

D1: `Information retrieval is an important subject.`

D2: `The Johnson family has got a golden retriever.`

D3: `Information theory uses plenty of theorems from mathematics.`

D4: `It provides a golden opportunity for information sharing.`

Our dictionary DICT consists of 8 words: w1 = information, w2 = retrieval, w3 = subject, w4 = Johnson, w5 = golden, w6 = theory, w7 = mathematics, w8 = sharing. By stemming, “retrieval" and “retriever" are regarded as the same word, and so are “theory" and “theorem".

Problem1 10%: Fill in the following table for these documents:

Answer: tf,tf.idf, No Normalizations (by do length, max frequency in doc) , NO log for idf: both can be added and the answers should be the same.

Word	**df**	**Idf=4/df**	**D1 tf↓**	**D2 tf↓**	**D3 tf↓**	**D4 tf↓**	**Q**
W1	3	1.33	1,1.33	0	1,1.33	1,1.33	0.415
W2	2	2	1,2	1,2	0	0	1
W3	1	4	1,4	0	0	0	2
W4	1	4	0	1,4	0	0	0
W5	2	2	0	1,2	0	1,2	0
W6	1	4	0	0	2,8	0	0
W7	1	4	0	0	1,4	0	0
W8	1	4	0	0	0	1,4	0

Problem2 10%: Each document can be viewed as an 8 dimensional vector of the 8 words (W1-W8) using tf.idf. Assume that a query **Q** (which is a sequence of words) has been converted to a point (0.415; 1; 2; 0; 0; 0; 0; 0). What is the score of D1 - D4 with respect to this query according to the cosine metric? List the document according to their retrieval rank for **Q**.

Answer: Using Cosine Similarity as required:
Sim(Q,D)= Sum(qixDi)/(|Q|x|D\). |Q|=SQRT(0.415^2+1^2+2^2)= SQRT(5.17)=2.27

Sim(Q,D1)= (1.33*0.415+2*1+4*2)/((2.27*SQRT(1.33^2+1+4))=0.995 Rank 1
Sim(Q,D2)= (1*2)/((2.27*SQRT(4+16+4))=0.18 Rank 2
Sim(Q,D3)= (1.33*0.415)/((2.27*SQRT(1.33^2+64+16))=0.027 Rank 4
Sim(Q,D4)= (1.33*0.415)/((2.27*SQRT(1.33^2+4+16))=0.052 Rank 3
Order D1,D2,D4,D3
Comment: can normalize and the ranking should be quite similar, through not necessarily the values.

Question 2 (16%):

a. A search engine has a collection of 160,000,000 pages (documents) with 400 words per page, on average.

(i) What is the minimal length for document IDs for the postings? In bits and in full bytes.

Answer:

Ceiling of log2(160,000,000)= 28bits
or 4bytes

Comment: cannot say 32 bits or 4 bytes: if we can work with bits then 28 bits is enough otherwise we may work with bytes in which case we need 4 bytes.

(ii) If the vocabulary size is 400,000, and the average dictionary word length is 8 characters
How many **bits** do you need for pointers if one is to store the dictionary as a single string with pointers to the start of each **word** (what is the length of each pointer).

Answer:

400,000x8=3,2000,000
For that we need 22 bits.

(iii) Compute the γ-code for the decimal number 1021.

Answer: 1021 = 1111111101, removing the most significant bit gives: 111111101 (9 digits).

1111111110111111101

(iv) Recover the gap values (in decimal) for the following string representing γ-encoding of a sequence of gaps in a posting list.

1011110111111101010111110 10101

Answer: 11, 1111, 11010, 110101=3, 15, 26, 53 (in decimal, in binary is not enough)

Question 3 (15%)

1. 10% Generally, how does stemming, stop word removal and Hamza Normalization (make all forms of Hamza the same) affect the overall dictionary size, term index size for each dictionary term and search recall and precision (I): Increase, (D) decrease, (NE): no effect.

Answer: circle as needed in the following table

Effect on: ➔	**Overall Dictionary size**	**Term Index Size**	**Search Recall**	**Search Precision**
Stemming	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)
Hamza Normalization (treat ء ؤ ئ إ أ أsame)	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)
Stop Word Removal	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)	(I), (D), (NE)

2. 5% Compute the edit distance between the following strings. Remember that the edit distance is the minimum number of deletions, insertions and substitutions needed to transform the first string into the second.
   How would you normalize the score? Why is the normalization needed?
   String 1: abracadabra
   String 2: nabucodor

Answer:7 (more details are needed) as explained in class.

By the size of the larger word. To account for variations in distances resulting from word size.

Question 4 (18%)

1- Assuming Zipf's law holds for the AP89 corpus with the below characteristics.

General:

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Some word rankings incollection:

Word	***Frequency f***	***Rqnk r***
assistant	5,095	1,021
sewers	100	17,110
toothbrush	10	51,555

1- Find the parameters of Zipf law for this collection (Zipf law constants). Test the values found to make sure they are obeyed everywhere.

Zipf Law: Y= K X^C, where Y is frequency and X is rank.

Log(Y)=Log(K)+C *log(C).

Two variable, thus need two equations: for assistant and sewers (can be different).

Log(5095)=Log(K)+C *log(1021)….(1)

Log(100)=Log(K)+C *log(17110) ….(2)

Solve then check if you get the results for toothbrush; then comment: little correspondence!!

2- Estimate the frequencies of the most frequent 3 words (at rank 1, 2, 3).

e.g. assistant has rank 1021 thus first word should be 1021 times more frequent than assistant which means 5095 x 1021= 5,201,995

Comment: Can use other parameters, can use a combination, can average them, NO Problem

3- If we add another collection to this one with similar characteristics: exactly the same number of words and vocabulary size (but no duplicate documents): In your opinion:

3-1. What will happen to the new vocabulary size?

They will have a large number of common words but still some new words, according to heap’s law and thus the vocabulary should grow (usually not double, though).

3-2. What will happen to the frequency of the 3 most frequent words?

They will have higher frequency, need not stay the same as in either collections. Most likely to be high frequency words in both collections (stop words in both)

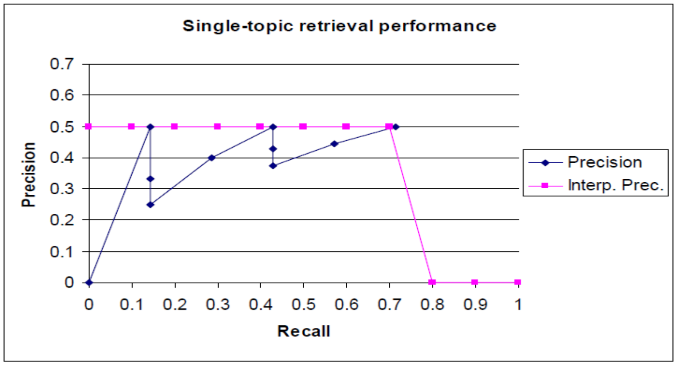
Question 5 (16%) Given a query q, where the relevant documents are d1, d3, d6, d7, d10, d12, d13, an IR system retrieves the following ranking: d2, d6, d5, d8, d3, d12, d11, d14, d7, d13.

1. What are the precision and recall for this ranking at each retrieved document?

Doc	**Recall**	**Precision**
D2	0/7 = 0.00	0/1 = 0.00
D6	1/7 = 0.14	1/2 = 0.50
D5	1/7 = 0.14	1/3 = 0.33
D8	1/7 = 0.14	1/4 = 0.25
D3	2/7 = 0.28	2/5 = 0.40
D12	3/7 = 0.42	3/6 = 0.50
D11	3/7 = 0.42	3/7 = 0.42
D14	3/7 = 0.42	3/8 = 0.37
D7	4/7 = 0.57	4/9 = 0.44
D13	5/7 = 0.71	5/10 = 0.50

2- Interpolate the precision scores at 11 recall levels. Remember that the interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level between the *j*-th and (*j* + 1)-th level:

Recall	**Precision**
0.0	**0.5**
0.1	**0.5**
0.2	**0.5**
0.3	**0.5**
0.4	**0.5**
0.5	0.5
0.6	**0.5**
0.7	**0.5**
0.8	**0.0**
0.9	**0.0**
1.0	**0.0**



Question 6 (18%) True or False: Place √ in the right square: If in doubt you can add some explanatory words (not recommended if sure about the answer).

1- □ **True** √ **False** In search results, precision at 5 (P@10) is always higher than precision at 10 (P@20).

2- √ **True** □ **False** With Positional indexing it is possible to recover the original document from the index something not possible for nonpositional index.

3- □ **True** √ **False** According to Heap's law the vocabulary keeps growing with the increase of the corpus size until it reaches the max for the given language then it remains the same.

4- √ **True** □ **False** Positional indexing can double or even triple the space needs of an index.

5- √ **True** □ **False** Boolean search requires more advanced skills on part of the user.

6- √ **True** □ **False** The phrase “قد قيل ما قيل إن صدقا وإن كذبا ... فما اعتذارك من قول إذا قيل” has more tokens than types/terms.

7- √ **True** □ **False** Overall, using skip pointers requires more space for the posting lists.

8- √ **True** □ **False** In the “bag of words” model of the document word order and word co-occurrence patterns are NOT important.

9- □ **True** √ **False** Pseudo-relevance feedback is based on user judgement on relevance to revise the query while Relevance feedback blindly assumes that the first N documents are relevant.

10- √ **True** □ **False** The most important measure of search engine quality is user **happiness** and the most important factor in user happiness is **relevance** of results

11- □ **True** √ **False** Using Jacqard Similarity on letter bigrams the word “trik” is closer (more similar) to “Trick” than “tric” is to “Trick”.

12- √ **True** □ **False** The vector space model of IR assumes that the order in which terms occur in a document is not important for retrieval.

13- □ **True** √ **False** We can easily get the number of unique terms in a particular document from an inverted index of postings.

14- √ **True** □ **False** If the **third** most frequent word in a document collection has a frequency of 300,000 then the word with frequency 100,000 is the one with **rank 9** (9th most frequent word).

15- □ **True** √ **False** Two documents D1 and D2 both have the word “Palestine” twice (2 times). D1 and D2 will always have the same rank in any web search.