Faculty of Engineering and Technology

Electrical and Computer Engineering Department

Second Semester 2015/2016

Course 539: Special Topics: Information Retrieval and Web Search

Instructor: Dr Adnan Yahya.

Midterm Exam

Question	Q1	Q2	Q3	Q4	Q5	Q6	Total
ABET outcome	e	e	a	a	b		
Max grade	18	22	15	15	15	17	102/100
Earned							

Please answer the following questions using the exam sheets only.

Question 1 (18%): Consider the following small part of a positional index with the format:

```
word: doc#: <position, position,...>; doc#: <position,...>.

Gates:     1: <1>; 2: <6>; 3: <2,15>; 4: <1>.
IBM:       3: <1,11>; 4: <3>; 7: <14,89>.
Microsoft: 1: <2>; 2: <12, 16,21>; 3: <13>; 5: <21,25>.
```

The `/k` operator, with the format: `term1 /k term2`
finds occurrences of `term1` within `k` words on either side of `term2`, where `k` is a positive integer argument. Thus, `k = 1` demands that `term1` is adjacent to `term2`.

a. Describe the set of documents that satisfy the query: **`Gates /2 Microsoft`**.

Answer:
{1,3}
(documents having **Gates, Microsoft a distance at most 2 from each other: D1 has Gates in position 1 and Microsoft in position 2, D3 has Gates in position 15 and Microsoft in position 13)**

b. Describe the set of values for **k** for which the query: **`Gates /k Microsoft`** returns the set of documents {1,3} as the answer.

Answer:

{2,3,4,5}
1 is no good as it doesn't include document 3. 6 is no good as it will include document 2 which is not in {1,3}.

c. Describe the set of values for k for which the query **`Microsoft /k Microsoft`** returns a non-empty set of documents as the answer.

Answer:
{4,5,…}
Will yield documents 2 and 5 which have multiple occurrences of **`Microsoft`**

d. Reconstruct Document #3 in the proper order .

Answer:
IBM Gates IBM Microsoft Gates

1 2 11 13 15 (checking document 3 in all terms and postings clearly some missing words: maybe stop words!).

Question 2 (22%): Given the following document collection:

```
D1: Speed the High speed
D2: Speed  and the Car accidents
D3: Accidents and tragedies
```

Assume that the stop word list contains the word set `{the, and}`.

a. Show the dictionary and the postings list including all the relevant statistics computed such as: tf, idf, tf-idf values shown explicitly with each document in the postings list (no normalization). Arrange terms alphabetically.

Answer:

Term↓ Postings→df, idf	Posting1 DocID(tf,tf.idf)	Posting2 DocID(tf,tf.idf)	Posting3 DocID(tf,tf.idf)
accidents-2,3/2	D2(1,3/2)	D3(1,3/2)	
car-1,3/1	D2(1,3)		
high-1,3/1	D1(1,3)		
speed-2,3/2	D1(2,3)	D2(1,3/2)	
tragedies-1,3/1	D3(1,3)		

Calculations are for documents with those terms only (not for all docs: NO postings for terms with 0 tf).

b. What are the relevance scores and the "relative" ranking of the documents for the query Q= "speed and accidents" using cosine measure based on tf.idf?

Answer: Vector=(accidents,car,high,speed,tragedies)

Q= (3/2,0,0,3/2,0);

D1=(0, 0,3,3,0); D2=(3/2,3,0,3/2,0); D3=(3/2,0,0,0,3)

Compute Cosine Similarity then rank. E.G.

Sim(Q,D1)= $(9/2)/(9/4+9/4)^{1/2}.(9+9)^{1/2}$= 4.5/2.12*4.24=0.50

Sim(Q,D2)= $(9/4+9/4)/(9/4+9/4)^{½}.(9/4+9+9/4)^{1/2}$= 4.5/2.12*3.67=0.57

Sim(Q,D3)= $(9/4)/(9/4+9/4)^{1/2}.(9/4+9)^{1/2}$=2.12*3.35 =0.31

Rank: D2-> D1->D3

Variations (log, weights, ..) are accepted also.

If stop words are not removed: can produce different values, still accepted.

c. What are the relevance scores and the "relative" ranking of the documents for the query ~~Q=~~ Q= "speed and accidents" using Jaccard measure?

Answer:

Q ∩D1={speed}, Q ∧ D1={speed, accidents, high}; Jaccard: 1/3

Q ∩D2={speed,accidents}, Q ∧ D2={speed, accidents, car}; Jaccard: 2/3

Q ∩D3={accidents}, Q ∧ D3={speed, accidents, tragedies}; Jaccard: 1/3

Relevance Order: D2, {D1,D3}

d. Generally, how does stemming, stop word removal 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**	**Recall**	**Precision**
Stemming	(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)

Question 3 (15%):

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.

Ceiling of (Log_2(160000000)) = 28 bits ~ 4 Bytes

(ii) 2-b. If the vocabulary size is 300,000, and the average dictionary word length is 10 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:

Ceiling of (Log_2(3,000,000)) = 24 bits ~ 3 Bytes

(iii) Compute the γ-code for the decimal number 1022.

Answer:

1111111110111111110

(iv) Recover the gap value in decimal for the following string representing a sequence of gaps in a posting list. 101 1110111 111101010 111110 10101

Answer:

3,15,26,53,

Question 4 (15%) Assuming Zipf's law with a corpus independent constant $A = 0.1N$, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words).

3:

F1 =0.1N

F2= 0.1N/2

F2= 0.1N/3

F1+F2+F3=.183N>.18 of words.

Question 5 (15%) Assume that an IR system returns a ranked list of 10 documents for a given query **Q**. Assume that according to a gold-standard labelling there are 5 relevant documents for this query in the collection, and that the only relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th positions in the ranked results. Calculate and clearly show the precision value for each of the following (11) recall levels: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 for this individual query.

Recall	**Precision**	**R**	**iP**
0.0	100	0.0	0.75
	0	0.1	0.75
0.2	**0.5**	**0.2**	0.75
		0.3	0.75
0.4	**0.67**	**0.4**	0.75
		0.5	0.75
0.6	**0.75**	**0.6**	0.75
	0.43	0.7	0.50
0.8	**0.5**	**0.8**	0.50
	0.44	0.9	0.0
	0.40	1.0	0.0

1	2	3	4	5	6	7	8	9	10	11	12
N	R	R	R	N	N	N	R	N	N	N	N

Will the result be any different if the search returned 12 documents but the same relevant documents in the same positions?

□ NO; explain. The same computation recall has to do with relevant elements returned and they are the same in both cases.

What is the precision at 5 (P@5) for this query?

3/5=0.6

Question 6 (17%) True or False: Place √ in the right square: If in doubt you can add some explanatory words (not recommended if sure about the answer). All True except 1 and 11.

1- □ **True** □ **False** Precision at 5 (P@5) is always better than precision at 10 (P@10).

2- □ **True** □ **False** The Levenshtein distance between "Research" and "Resaerch" is 2.

3- □ **True** □ **False** Zipf's law implies that stop words have higher ranks than other words (higher rank means closer to rank 1: top ranked).

4- □ **True** □ **False** Using dictionaries/thesaurus in search improves recall but may reduce precision.

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

6- □ **True** □ **False** The sentence "He was born before the state was declared" has more tokens than types/terms.

7- □ **True** □ **False** Using skip pointers requires more space for the postings.

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** In ranked retrieval the absolute similarity may be sacrificed (not exact) but relative similarity cannot be sacrificed.

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** In general, for Arabic texts stemming gives better recall and worse precision compared with Rooting (basing search and indexing on word roots).

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** Relevance feedback may try to reformulate the user query to minimize the distance to relevant documents and maximize the distance to non-relevant documents.

14- □ **True** □ **False** In multi-Tier indexing the higher tier index is generally much smaller than lower tier index.

15- □ **True** □ **False** Huge **Query Logs** is one of the **most important** assets search engine companies have and its size may give **major** advantage to one over another.

16- □ **True** □ **False** Document normalization by length tends to favor shorter documents and the fix is to use **pivot normalization**.

17- □ **True** □ **False** Pseudo-relevance feedback requires **no user intervention** to modify the query for better results.