October 25th, 2018

German University in Cairo
Media Engineering and Technology
Lecturer: Mervat AbuElkheir
TAs: Hadeel Mostafa

CSEN1076 Natural Language Processing and Information Retrieval

Winter term 2018
Midterm Exam

Bar Code

Instructions: Read carefully before proceeding.

Please indicate your group and major in the tables below.

Major	
CSEN	**DMET**

1) Duration of the exam: 2 hours (120 minutes).

2) (Non-programmable) Calculators are allowed. No books or other aids are permitted for this test.

3) This exam booklet contains 10 pages, including this one. **Note that if one or more pages are missing, you will lose their points. Thus, you must check that your exam booklet is complete.**

4) Write your solutions in the space provided. If you need more space, write on the back of the sheet.

5) Include any assumptions that you need to make.

6) Follow the instructions of your proctors fully. Failure to do so will result in severe disciplinary actions.

Good Luck!

Do not write anything on this page ☺

Question	1	2	3	4	5	Σ
Marks	8	10	12	15	5	50
Final Marks						

Formulas you may need:

- *tf-idf* weight of a term → $\boldsymbol{w_{t,d} = tf \times idf = \left(1 + \log_{10} tf_{t,d}\right) \times \log_{10} \frac{N}{df_t}}$
- Cosine similarity → $\boldsymbol{\cos(q, d) = \frac{\sum_{i=1}^{|V|} q_i \times d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}}$

  where q_i is the *tf-idf* weight of term i in the query, and d_i is the *tf-idf* weight of term i in the document.

- Average Precision → $\boldsymbol{AP(Q) = \frac{1}{m} \sum_{k=1}^{m} Precision(R_k)}$

  where k is the rank position for a relevant doc, and R_k is a ranked result

Question 1 [2 Marks each]:

a) Imagine you have a collection of a million documents (N) with an average of 1,000 words per document and a total of M=500,000 terms (unique words). Which of the following statements is **false** regarding the collection's Term-Document Incidence Matrix?

1. The matrix would be extremely sparse (most entries would be 0).	☐
2. The matrix would consist of a distribution of 0 and 1 with dimension M by N.	☐
3. The matrix shows the term frequency (*tf*) of each term in each document.	■
4. Each column (vector) shows which terms are present in each document.	☐

b) Which of the following statements is **false** with regards to Boolean Retrieval model?

1. It answers queries based on Boolean expressions (AND, OR and NOT).	☐
2. It views documents as a set of terms.	☐
3. It is very precise, as its queries need to meet a very specific condition.	☐
4. It cannot combine two operators, such as "AND NOT" or "OR NOT".	■

c) Find the Jaccard coefficient (JC) for the query and documents below.

Query: top university (set q)
Doc 1: university of California (set d_1)
Doc 2: best university in USA (set d_2)

1. $JC(q, d_1) = 1/4\ , JC(q, d_2) = 1/5$	■
2. $JC(q, d_1) = 1/5,\ JC(q, d_2) = 1/6$	☐
3. $JC(q, d_1) = 0, JC(q, d_2) = 1/6$	☐
4. $JC(q, d_1) = 1/5, JC(q, d_2) = 0$	☐

d) Mark the **false** statement with regards to the term frequency (*tf*)?

1. The *tf* is the number of times that a term occurs in a document.	☐
2. Relevance of a term in a document increases proportionally with its *tf*.	■
3. The *tf* of a query is the sum of the *tf* of each of the terms in the query.	☐
4. The *tf* of a query is 0 if none of the query terms is present in the document.	☐

Question 2:

Consider the following documents:

doc1	phone ring person happy person
doc2	dog pet happy run jump
doc3	cat purr pet person happy
doc4	life smile run happy
doc5	life laugh walk run run

a) Construct the term–document matrix. Assume that no stemming or stop-word removal is required.

b) Construct the inverted index required for ranked retrieval for these five documents. Assume that no stemming or stop-word removal is required.

Answer 2:

a)

	phone	ring	person	happy	dog	pet	run	jump	cat	purr	life	smile	laugh	walk
doc1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
doc2	0	0	0	1	1	1	1	1	0	0	0	0	0	0
doc3	0	0	1	1	0	1	0	0	1	1	0	0	0	0
doc4	0	0	0	1	0	0	1	0	0	0	1	1	0	0
doc5	0	0	0	0	0	0	1	0	0	0	1	0	1	1

b)

phone → 1
ring → 1
person → 1, 3
happy → 1, 2, 3, 4
dog → 2
pet → 2, 3
run → 2, 4, 5
jump → 2
cat → 3
purr → 3
life → 4, 5
smile → 4
laugh → 5
walk → 5

Question 3:

Below is a table showing how two human judges rated the relevance of a set of documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {2, 5, 6, 7, 8} and assume the documents are ranked (document 2 is 1st document in results, document 5 is 2nd document, etc.).

docID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Judge 1	0	0	1	1	1	1	1	1	0	0	0	1
Judge 2	0	0	1	1	0	0	0	0	1	1	1	0

a) Calculate the precision and recall of your system if a document is considered relevant only if the two judges agree it is relevant.
b) Calculate the precision and recall of your system if a document is considered relevant if either judge thinks it is relevant.
c) Calculate the average precision of your IR system based on both relevance scenarios in (a) and (b).

Answer 3:

a)

docID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Judge 1	0	0	1	1	1	1	1	1	0	0	0	1
Judge 2	0	0	1	1	0	0	0	0	1	1	1	0
Final decision	0	0	1	1	0	0	0	0	0	0	0	0

Precision = 0 (of the retrieved documents, none was relevant.)

Recall = 0 (of the relevant documents, none was retrieved.)

b)

docID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Judge 1	0	0	1	1	1	1	1	1	0	0	0	1
Judge 2	0	0	1	1	0	0	0	0	1	1	1	0
Final decision	0	0	1	1	1	1	1	1	1	1	1	1

Precision = 4/5 (# retrieved documents that are relevant from total # of retrieved documents.)

Recall = 4/10 (# of relevant documents retrieved from total # of relevant documents.)

c)

docID	**2**	**5**	**6**	**7**	**8**
Final decision (a)	0	0	0	0	0
Final decision (b)	0	1	1	1	1
Precision	0	0.5	0.67	0.75	0.8

$AP(Q) in\ (b) = 0$

$AP(Q) in\ (b) = \frac{(0.5+0.67+0.75+0.8)}{4} = 0.68$

Question 4:

1) Two retrieval systems, **X** and **Y**, are being compared. Both are given the same query, applied to a collection of 2500 documents. System **X** returns 700 documents, of which 90 are relevant to the query. System **Y** returns 300 documents, of which 120 are relevant to the query. Within the whole collection there are in fact 160 documents relevant to the query.

   a) What are the values of True Positives, False Positives, True Negatives, False Negatives for system **X**?
   b) Calculate the precision and recall for the 2 systems, showing the details of your calculations.

2) For the following search tasks, choose the most appropriate evaluation metric (precision, recall, F-measure). Justify your answer.

   a) A lawyer searching for all relevant evidence to one of his cases. The lawyer is evaluated by whether he could win the case and he bills his client by hours. Therefore he does not mind to read through all the documents that are returned by a search engine.
   b) An American basketball fan searching for information and history for NBA. Some of the returned pages provide a lot of relevant details, for example, team rankings, match scores, the latest news, etc. Some pages are just marginally relevant. Others are less interesting or irrelevant.

Answer 4:

1)

a) TP = 90, FP = 610, TN = 1730, FN = 70

b) Precision (sys. X) = 90/700 = 0.13

Precision (sys. Y) = 120/300 = 0.4

Recall (sys. X) = 90/160 = 0.56

Recall (sys. Y) = 120/160 = 0.75

2)

a) Recall. Lawyer needs all relevant documents and can tolerate noise.

b) Precision. Fan needs relevant results to appear first, and they do not mind missing on some information.

Answer for part 2 will be given 4 each.

For (a), 4 for recall, 2 for F-measure, and 1 for precision. One mark is deducted if recall is the answer but with no justification. Otherwise half the mark is deducted if there is no justification.

For (b), 4 for precision, 3 for F-measure, and 1 for recall. One mark is deducted if precision is the answer but with no justification. Otherwise half the mark is deducted if there is no justification.

Question 5:

Calculate cosine similarity between the query and the list of documents below, given the vectors provided. Each vector provides *tf-idf* weights for the terms: *city*, *Spain*, *United States*, *paella*, *gumbo*, *shellfish*, *beer*. Note that paella is a Spanish rice dish. What will be the answer to the query?

Vectors

- Vector for Query "Which Spanish city has the best paella?" → vector: ⟨10,7,0,20,0,5,0⟩
- Wikipedia Entry for Valencia, Spain → vector: ⟨7,30,1,25,0,5,0⟩
- Wikipedia Entry for Seville, Spain → vector: ⟨7,25,3,0,0,6,0⟩

Wikipedia Entry for New Orleans, USA → vector: ⟨7,5,30,0,20,10,0⟩

Answer 5:

tf	city	Spain	United States	paella	gumbo	shellfish	beer
Query	10	7	0	20	0	5	0
Valencia Entry	7	30	1	25	0	5	0
Seville Entry	7	25	3	0	0	6	0
New Orleans Entry	7	5	30	0	20	10	0

cos(Valencia, Query) = (10×7+7×30+0×1+20×25+0×0+5×5+0×0)/(23.96×40) = 0.84 → 1st

cos(Seville, Query) = (10×7+7×25+0×3+20×0+0×0+5×6+0×0)/(23.96×26.8) = 0.43 → 2nd

cos(NO, Query) = (10×7+7×5+0×30+20×0+0×20+5×10+0×0)/(23.96×38.4) = 0.17 → 3rd

The Spanish city that has the best paella is Valencia.