CSL-554
Text and Web Intelligence Analytics

ASSIGNMENT 1 - (Solutions)

Submission Instructions:

1. **The deadline to submit assignment-1 is 17 August 2018 by 3:00 pm (no late submissions will be allowed).**
2. **The weightage for assignment-1 is 10% in assignment/ internal assessment component.**
3. **Each question in Assignment-1 is of 12.5 marks.**

Question 1: Consider an index for 1 million documents each having a length of 1,000 words. Consider there are 100K distinct terms in total. We do not wish to keep track of term-frequency information. What is the space requirement for a term-document incidence matrix that does not exploit sparsity (Answer in terms of Bytes required)?

Solution:

n = 1M

m = 100K

The full bit-matrix including all zeros would need cn X m entries:

10^6 X 10^5 bits = 10^11 bits = 12.5 GB

Question 2: Two retrieval systems, X and Y, are being compared. Both are given the same query, applied to a collection of 1500 documents. System X returns 400 documents, of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are a total of 50 documents relevant to the query.

Tabulate the results for each system, and compute the precision and recall for both X and Y. Show your working.

Solution:

X	Relevant	Not relevant	Total
Retrieved	40	360	400
Not retrieved	10	1090	1100
Total	50	1450	1500

Y	Relevant	Not relevant	Total
Retrieved	15	15	30
Not retrieved	35	1435	1470
Total	50	1450	1500

System X precision $P = \frac{40}{400} = 0.1$

System Y precision $P = \frac{15}{30} = 0.5$

System X recall $R = \frac{40}{50} = 0.8$

System Y recall $R = \frac{15}{50} = 0.3$

Question 3: Following is a term-document incidence matrix. Which documents will be retrieved if we query the system as T1 AND T3 AND NOT T6. Show the working.

	D1	D2	D3	D4	D5	D6
T1	1	0	0	1	1	1
T2	0	1	0	1	0	1
T3	1	0	0	1	0	1
T4	1	0	0	1	1	1
T5	0	1	1	1	0	0
T6	0	0	1	0	1	1

Solution:

T1 AND T3 AND NOT T6

T1: 1 0 0 1 1 1
T3: 1 0 0 1 0 1
---------- (Performing AND Operation)
1 0 0 1 0 1
NOT T6: 1 1 0 1 0 0
---------- (Performing AND Operation)
1 0 0 1 0 0 (Result)

Thus resultant documents retrieved are D1 and D4

Question 4: Suppose we have a topic (i.e. a query) with a total of 5 relevant documents in the whole collection. A system has retrieved 6 documents whose relevance status is:

[+, +, -, -, -, +]

in the order of ranking. A "+" (or "-") indicates that the corresponding document is relevant (or non- relevant). For example, the first two documents are relevant, while the third is non-relevant, etc. Compute the precision and recall for this result.

Solution:

Total Relevant=5
Total Retrieved=6
Relevant Retrieved=3
Precision= Relevant Retrieved/ Total Retrieved = 3/6= 0.5
Recall= Relevant Retrieved/ Total Relevant= 3/5= 0.66

Question 5: Write out a postings merge algorithm for an "*x* AND (NOT *y)"* query.

a) p_1 AND (NOT p_2)

INTERSECT (p_1, p_2)
answer ← < >
while $p_1 \neq$ NIL and $p_2 \neq$ NIL
do if docID (p_1) = docID (p_2)
then $p_1 \leftarrow next(p_1)$
$p_2 \leftarrow next(p_2)$
else if docID(p_1) < docID(p_2)
then ADD (answer, docID(p_1))
$p_1 \leftarrow next(p_1)$
else $p_2 \leftarrow next(p_2)$
return answer

Question 6: Consider the following documents:

Doc 1 breakthrough drug for schizophrenia
Doc 2 new schizophrenia drug
Doc 3 new approach for treatment of schizophrenia
Doc 4 new hopes for schizophrenia patients

a. Draw the term-document incidence matrix for this document collection.
b. Draw the inverted index representation for this collection.
c. For this document collection what are the returned results for queries:
    a. schizophrenia AND drug
    b. for AND NOT(drug OR approach)

Solution:

a. Term- document incidence matrix

	Doc1	Doc2	Doc3	Doc4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	1	1	1
of	0	0	1	0
patients	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

b.

Word	Frequency	Document
approach	1	3
breakthrough	1	1
drug	2	1,2
for	3	1,3,4
hopes	1	4
new	3	2,3,4
of	1	3
patients	1	4
schizophrenia	4	1,2,3,4
treatment	1	3

c)

a) schizophrenia AND drug
Schizophrenia: 1 1 1 1
Drug: 1 1 0 0
----------(AND operation)
1 1 0 0
Result: Doc-1, Doc-2

b) for AND NOT(drug OR approach) (4)
Drug: 1 1 0 0
Approach: 0 0 1 0
----------(OR operation)
1 1 1 0----Result:1
For: 1 0 1 1
NOT (drug OR approach): 0 0 0 1
----------(AND operation)
Final result of query: 0 0 0 1
Thus it is present only in Doc-4

Question 7: The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

a. abandon/abandonment
b. absorbency/absorbent
c. marketing/markets
d. university/universe
e. volume/volumes

Solution:

c. marketing/market should not be conflated
d. university/universe should not be conflated
These words cannot be conflated as they depict different meaning entirely. Words can be conflated when they share some characteristics together. In all other cases, the words share the same meaning, except cases 'c' and 'd'

Question 8: The following list of Rs and Ns represents relevant (R) and non relevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 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 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N N N N R N R N N N R N N N N R

a. What is the precision of the system on the top 20?
b. What is the F1 on the top 20?
c. What is the uninterpolated precision of the system at 25% recall?
d. What is the interpolated precision at 33% recall?
e. Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned.

f. What is the largest possible MAP that this system could have?
g. What is the smallest possible MAP that this system could have?
h. In a set of experiments, only the top 20 results are evaluated by hand. The result in (e) is used to approximate the range (f)–(g). For this example, how large (in absolute terms) can the error for the MAP be by calculating (e) instead of (f) and (g) for this query?

Solution:

a. Precision = 6/20 = 0.3

b. Recall = 6/8 = 0.75
F1 =(2PR)/(P+R)=3/7 = 0.43

c. Uninterpolated precision could be 1, 2/3, 2/4, 2/5, 2/6, 2/7, 1/4

d. Highest precision found for any recall level larger than 33% is 4/11 = 0.364, Interpolated precision at 33% recall is 4/11 = 0.364.

e. MAP = 1/6 * (1 + 1 + 3/9 + 4/11 + 5/15 + 6/20) = 0.555

f. Largest possible MAP will be achieved when 2 relevant documents (which are remaining as total there are 8 relevant documents in collection) are retrieved at 21 and 22 positions.
MAP(largest)=(1/8)*(1+1+3/9+4/11+5/15+6/20+7/21+8/22)=0.503

g. . Smallest possible MAP will be achieved when 2 relevant documents (which are remaining as total there are 8 relevant documents in collection) are retrieved at 9999 and 10000 positions.
MAP(smallest)=(⅛)*(1+1+3/9+4/11+5/15+6/20+7/9999+8/10000)=0.417

h. Error can be in range
0.555-0.417 = 0.138,
and 0.555-0.503 = 0.052
Error is in [0.052, 0.138]