CS5154/6054 Quiz 1 Key, 8/23/2022

Exercise 1.2 [⋆]

Consider these 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.

	Doc 1	Doc 2	Doc 3	Doc 4
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
schizophrenia	1	1	1	1
new	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
of	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1

b. Draw the inverted index representation for this collection, as in Figure 1.3 (page 7).

Dictionary	Postings
breakthrough	1
drug	1, 2
for	1, 3, 4
schizophrenia	1, 2, 3, 4
new	2, 3, 4
approach	3
treatment	3
of	3
hopes	4
patients	4

Exercise 1.3 [⋆]

For the document collection shown in Exercise 1.2, what are the returned results for these queries:

a. schizophrenia AND drug
b. for AND NOT(drug OR approach)

a. 1, 2 b. drug OR approach: 1, 2, 3, NOT(drug OR approach) 4, (1, 3, 4) AND 4 is 4.

CS5154/6054 Quiz 2 Key, 8/25/2022

1. In an inverted index, a term is associated with a set of
   a. documents.
   b. tokens.
   c. types.
   d. lemmas.

2. In Boolean retrieval, suppose query "A" returns x documents and query "B" returns y documents. "A AND NOT B" returns at most
   a. x documents.
   b. y documents.
   c. x - y documents.
   d. x + y documents.

3. Let N be the number of documents in the collection, M be the vocabulary size, L be the number of terms in an average document. What is the average length of the posting lists in the inverted index?
   a. L
   b. N/L
   c. NL/M
   d. NM/L

4. How many zeros are there in the term-document?
   a. (M-N)L
   b. (M-L)N
   c. (N-L)M
   d. (L-N)M

5. With an inverted index, what is the closest time needed to rank all documents for a query made of K terms? (None is correct. The right answer is O(KNL/M). No deduction of points.
   a. O(KL)
   b. O(KN)
   c. O(MN)
   d. O(LN)

6. Which text emoticon is not matched by the regular expression r’[:;<]\-?[\)\(3]’?
   a. ;-)
   b. :-3
   c. :(
   d. :\

7. The special sequence \w in regular expression matches
   a. a word.
   b. a word character.
   c. a sequence of words.
   d. the character 'w'.

8. The regular expression r'\[[^\[\]]*\]' matches (either a or c is correct).
   a. [].
   b. [[]].
   c. [^^^].
   d. \[*\].

9. What may be returned as a token with re.findall(r'\w\w+', text)?
   a. 2021-10-14
   b. 9:30am
   c. I
   d. remove_accents

CS5154/6054 Quiz 3 Key, 8/30/2022

Describe using a pseudo algorithm how the inverted index can be used to find the intersection sizes of all documents containing at least one word in the query, when both query and documents are sets of words.

Given: the inverted index as a set of documents (or docIDs) for each term (in a vocabulary) and a query as a set of terms.

Output: an integer for each document as the size of the intersection between the set of terms representing the query and the set of terms representing the document

Step 1: Initialize zero for all intersection sizes.

(Notice when a map, dictionary, or Counter is used, this means stating with an empty one.)

Step 2: For each term in the query, get the associated set of documents from the inverted index

and for each document in this set, add one to the intersection size of the document.

(When a map, dictionary, or Counter is used, start 1 for the document if it is not in it, increase it by 1 if it is already in it.)

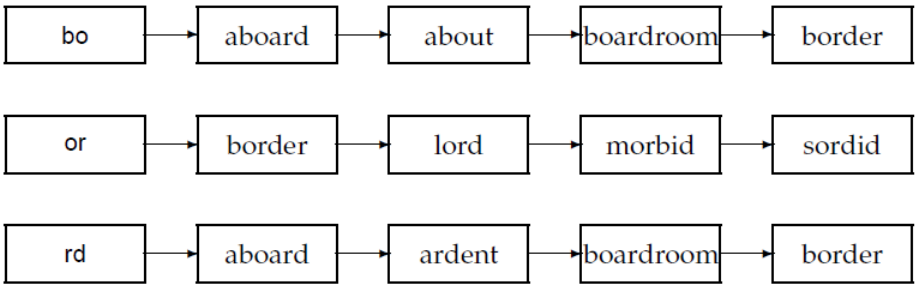
Step 3: The intersection sizes have been computed. Output the positive ones.

(Notice: The reader of the pseudo code is not expected to understand Python or other coding languages. You lose some points if you assume the opposite. You also lose points if the algorithm is not using the inverted index.)

CS5154/6054 Quiz 4 Key, 9/1/2022

Exercise 3.10

Compute the Jaccard coefficients between the query bord and each of the terms in Figure 3.7 that contain the bigram or.



► **Figure 3.7** Matching at least two of the three 2-grams in the query bord.

JACCARD COEFFICIENT

quently, we require more nuanced measures of the overlap in k-grams between a vocabulary term and q. The linear scan intersection can be adapted when the measure of overlap is the *Jaccard coefficient* for measuring the overlap between two sets A and B, defined to be $|A \cap B|/|A \cup B|$. The two sets we consider are the set of k-grams in the query q, and the set of k-grams in a vocabulary term. As the scan proceeds, we proceed from one vocabulary term t to the next, computing on the fly the Jaccard coefficient between q and t. If

The 2gram set of the query **bord** is A = {bo, or, rd}. Find that (B) for each of the terms containing **or** and calculate the Jaccard coefficient between A and B.

Word for B	B	A ∩ B	|A ∩ B|	|A u B|	j.c.=|A ∩ B|/|A u B|
aboard	ab, bo, oa, ar, rd	bo, rd	2	3+5-2	2/6
border	bo, or, rd, de, er	bo, or, rd	3	3+5-3	3/5
lord	lo, or, rd	or, rd	2	3+3-2	2/4
morbid	mo, or, rb, bi, id	or	1	3+5-1	1/7
sordid	so, or, rd, di, id	or, rd	2	3+5-2	2/6

CS5154/6054 Quiz 5 Key, 9/6/2022

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

► **Figure 6.8** Example of idf values. Here we give the idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

► **Figure 6.9** Table of tf values for Exercise 6.10.

Exercise 6.10

Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in Figure 6.9. Compute the tf-idf weights for the terms car, auto, insurance, best, for each document, using the idf values from Figure 6.8.

	Doc1	Doc2	Doc3
car	27 x 1.65 = 44.55	4 x 1.65 = 6.6	24 x 1.65 = 39.6
auto	3 x 2.08 = 6.24	33 x 2.08 = 68.64	0
insurance	0	33 x 1.62 = 53.46	29 x 1.62 = 46.98
best	14 x 1.5 = 21	0	17 x 1.5 = 25.5

Exercise 6.9

What is the idf of a term that occurs in every document? Compare this with the use of stop word lists.

When df = N, log (N/df) = log 1 = 0. This is equivalent to have a stop word list for all words occurring in all documents in the collection. When df is close to N but not N, the word will be included but with very little weight.

CS5154/6054 Quiz Key, 8/8/2022

Exercise 6.18

One measure of the similarity of two vectors is the *Euclidean distance* (or L_2 distance) between them:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

Given a query q and documents $d_1, d_2, \ldots$, we may rank the documents d_i in order of increasing Euclidean distance from q. Show that if q and the d_i are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.

Hints or lemmas that you can use in your proof.

1. ranking by Euclidean distance is the same as ranking by squared Euclidean distance:
2. If $|q - d1| < |q - d2|$, then $\sum_i (q_i - d1_i)^2 < \sum_i (q_i - d2_i)^2$ and vice versa.
3. q and d1, d2 are all normalized to unit vectors means $\sum_i q_i^2 = \sum_i d1_i^2 = \sum_i d2_i^2 = 1$.
4. $(q_i - d1_i)^2 = q_i^2 - 2\, q_i\, d1_i + d1_i^2$.
5. When q and d1 are normalized to unit vectors, the cosine similarity between them is $\sum_i q_i\, d1_i$.
6. need to show that $|q - d1| < |q - d2|$ if and only if $\sum_i q_i\, d1_i > \sum_i q_i\, d2_i$.

A proof for 6:
By 2, $|q - d1| < |q - d2|$ if and only if $\sum_i (q_i - d1_i)^2 < \sum_i (q_i - d2_i)^2$.
By 4, this is $\sum_i q_i^2 - 2\sum_i q_i\, d1_i + \sum_i d1_i^2 < \sum_i q_i^2 - 2\sum_i q_i\, d2_i + \sum_i d2_i^2$.
By 3, this is $1 - 2\sum_i q_i\, d1_i + 1 < 1 - 2\sum_i q_i\, d2_i + 1$.
By cancelling out the 1's and removing the 2's, this is $-\sum_i q_i\, d1_i < -\sum_i q_i\, d2_i$.
By removing the negative sign and changing order, this is $\sum_i q_i\, d1_i > \sum_i q_i\, d2_i$.

CS5154/6054 Quiz 7 Key, 9/13/2022

Exercise 8.1 [⋆]

An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

true positives tp = 8, false positives fp = 10, precision P = tp /(tp + fp) = 8/(8 + 10) = 8/18 = 4/9 = 0.444
20 = tp + fn and recall R = tp /(tp + fn) = 8/20 = 2/5 = 0.4

Exercise 8.10 [⋆⋆]

Below is a table showing how two human judges rated the relevance of a set of 12 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 {4, 5, 6, 7, 8}.

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

a. Calculate the kappa measure between the two judges.
b. Calculate precision, recall, and F_1 of your system if a document is considered relevant only if the two judges agree.
c. Calculate precision, recall, and F_1 of your system if a document is considered relevant if either judge thinks it is relevant.

a. Observed probability of agreement P(A) = 4/12 = 1/3, Pooled marginals P(0) = (6 + 6)/(12 + 12) = ½, P(1) = (6 + 6)/(12 + 12) = ½. Probability that the two agree by chance P(E) = $P(0)^2 + P(1)^2$ = ¼ + ¼ = ½. Kappa statistic κ = (P(A) – P(E)) / (1 – P(E)) = (1/3 – ½)/(1 – ½) = 2(1/3 – ½) = 2/3 – 1 = -1/3
b. (two judges agree it is relevant) {4,5,6,7,8} is the retrieved, {3,4} is the relevant, tp = 1, fp = 4, fn = 1, P = tp/(tp + fp) = 1/5, R = tp/(tp + fn) = ½, F1 = 2/(1/P + 1/R) = 2/(5 + 2) = 2/7 = 0.286.
c. Now {3,4,5,6,7,8,9,10,11,12} is the relevant. tp = 5, fp = 0, fn = 5, P = tp/(tp + fp) = 1, R = tp/(tp + fn) = ½, F1 = 2/(1/P + 1/R) = 2/(1 + 2) = 2/3 = 0.667.

CS5154/6054 Quiz 8 Key, 9/15/2022

Exercise 8.8 [⋆]

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1 R N R N N N N N R R

System 2 N R N N R R R N N N

a. What is the MAP of each system? Which has a higher MAP?

b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?

c. What is the R-precision of each system? (Does it rank the systems the same as MAP?)

	P1	P2	P3	P4	MAP	R-precision
System 1	1/1	2/3	3/9	4/10	54/90 = 0.6	2/4 = 0.5
System 2	1/2	2/5	3/6	4/7	69/140 = 0.49	1/4 = 0.25

System 1 has the higher MAP. Get most R's as early as possible. The R-precision ranks like MAP.

Exercise 8.9 [⋆⋆]

The following list of Rs and Ns represents relevant (R) and nonrelevant (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 F_1 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?

a. P = 6/20 = 0.3

b. R = 6/8 = 0.75, 1/F1 = ½(1/P + 1/R) = ½(20+8)/6 = 14/6 = 7/3, F1 = 3/7 = 0.429.

c. 25% recall occurs when we have retrieved 2 and the precision is 2/2 = 1.

d. 33% recall occurs when we retrieved >= 33% of 8 or 2.64 or at the third R with precision 3/9, or 0.33, which is smaller than the precision at the fourth R, 4/11, or 0.36. The interpolated precision at the third R is then 4/11, or 0.36.

e. MAP = (1/1 + 2/2 + 3/9 + 4/11 + 5/15 + 6/20)/6 = (1 + 1 + 0.33 + 0.36 + 0.33 + 0.3)/6 = 3.32/6 = 0.55.