

Information Search and Retrieval Exam Example

Exercise 1

Consider these documents:

Doc1: solution found for laziness

Doc2: old laziness found

Doc3: old approach for treatment of laziness

Doc4: old hopes for laziness patients

A: Draw the term-document incidence matrix for this document collection.

B: Draw the positional inverted index representation for this collections.

Exercise 2

Recommend a query processing order for the following Boolean query:

(bush OR apricot) AND (pudding OR brown) AND (phones OR ears)

Assume that the document frequencies of the terms in the above query are:

ears	213312
phones	87009
pudding	107913
brown	271658
bush	46653
apricot	316812

Explain the rationale of the order that you found.

Exercise 3

Why are skip pointers not useful for queries of the form $x \text{ OR } y$?

Exercise 4

Assume a biword index. Give an example of a document which will be returned for a query of 'open university israel' but is actually a false positive which should not be returned.

Exercise 5

Write down the entries in the permuterm index dictionary that are generated by the term 'game'.

Exercise 6

Describe the MapReduce approach for generating the inverted index of a web collection.

Exercise 7

Imagine to use Jaccard coefficients for computing a query document score: $\text{jaccard}(Q,D) = \frac{|Q \cap D|}{|Q \cup D|}$, where Q and D represent the set of terms included in a query and a document. Let Q='open source'. Identify a document D1 that is relevant for Q and a document D2 not relevant for Q such that $\text{jaccard}(Q,D1) < \text{jaccard}(D,D2)$. This example should illustrate why jaccard is not suited to score documents.

Exercise 8

How does the base of the logarithm affect idf calculation?

Exercise 9

Compute the vector space similarity between the query "film cameras" and the document "film cameras and digital cameras". Assume that the document frequencies of the terms film, camera and digital are: 100, 300, 1,000. Assume the number of documents is $N = 10,000$, logarithmic term weighting for query and document, idf weighting for the query only and cosine normalization for the document only. Treat and as a stop word. What is the final similarity score?

Exercise 10

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 R R

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

- What is the MAP (Mean Average Precision) of each system? Which has a higher MAP?
- Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
- What is the R-precision of each system? (Does it rank the systems the same as MAP?)

Exercise 11

Consider the following training set

D1: "American Boston American" -> Class=Y

D2: "American American Chicago" -> Class=Y

D3: "American Washington" -> Class=Y

D4: "Rome Italy American" -> Class=N

Estimate $P(x_k | c_j)$ and $P(c_j)$ using the following formulas:

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

$$P(x_k | c_j) \leftarrow \frac{n_{jk} + 1}{n_j + |\text{Vocabulary}|}$$

Where: docs_j is the subset of documents for which the target class is c_j ; x_k is a word of the vocabulary; n_{jk} is the number of occurrences of x_k in all the documents in class j , n_j is the total number of words in the documents in class j .

What is the predicted class of the following test document:

D5: "American American American Rome Italy"

Exercise 12

Consider an item-based Collaborative Filtering recommender system. If an active user u_0 has rated four items (p_1, p_2, p_3, p_4) with ratings $(1, 2, 4, 5) = (v_{01}, v_{02}, v_{03}, v_{04})$. Assume further that there are two additional items p_5 and p_6 , which has not been rated by u_0 , and the similarities of p_5 and p_6 with the items (p_1, p_2, p_3, p_4) are given by the following two vectors $(0.8, 0.9, 0.7, 0.3)$ and $(0.7, 0.4, 0.9, 0.9)$ respectively. If CF must recommend just one item, what should it be?

Explain the rationale of the reply by computing the predicted rating for p_5 and p_6 as the weighted and normalized average of the ratings of the products (p_1, p_2, p_3, p_4) (i.e., the usual formula for item-based CF).

Exercise 13

Define the Mean Absolute Error, and compute it for a test set of 200 predictions where the system predicts in 60% of the cases a rating that is larger than the true rating by 1, in 30% a rating that is lower than the actual rating by 2, and in 10% of the cases the exact rating.