Dr. Nastaran FATEMI

Information Retrieval Exercises

Part 1: Boolean Model and Vector Space Model

A. For a conjunctive query (AND), is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isn't.

B. How should the Boolean query x AND NOT y be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.

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

D. Consider the following collection of four documents:

- Doc1: Shared Computer Resources
- Doc2: Computer Services
- Doc3: Digital Shared Components
- Doc4: Computer Resources Shared Components

Assuming each word is a term:

1. What documents are retrieved, with the Boolean model, with the query "Computer AND NOT Components"?
2. Compute the idf value of the terms "Computer" and "Components" (consider we have only these 4 documents in the collection).
3. Compute the vector model representation of Doc4 using tf-idf weights (logarithmic tf * idf).
4. Compute the vector model representation of the query "Computer Components" using only raw tf and no idf.
5. Compute the similarity between the query "Computer Components" and Doc4 with the cosine similarity measure.

E. Compute the vector space similarity between:

a. the query**: "digital cameras"** and

b. the document: **"digital cameras and video cameras"**

by filling out the empty columns in the table bellow.

- Assume total number of documents in the collection **N = 10,000,000**
- use logarithmic term weighting (wf columns) for query and document,
- idf weighting for the query only, and
- cosine normalization for the document only.

Treat **and** as a stop word. Enter term counts in the **tf** columns.

What is the final similarity score?

	query					document			
word	tf	wf	df	idf	q_i = wf-idf	tf	wf	d_i = normalized wf	$q_i \cdot d_i$
digital			10,000						
video			100,000						
cameras			50,000						

Part 2: Evaluation in Information Retrieval

F. Consider an information need for which there are 6 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):

System1	NNNNR RRRRN
System2	NRRNR RNNNN

a. What is the Precision, Recall, and F-Measure of each system for the top 10 documents? Comment on your results.

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

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

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

G. The following list of R's and N's 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.

RRNNN NNNRN RNNNR NNNNR

a. What is the precision of the system on the top 20?

b. What is the F-Measure 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) to (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?