Information Retrieval: Practical 1

Hilary Term 2012

The aim of this practical session is to implement a vector space document retrieval model, and to evaluate it.

Data Sets

The document set you have available consists of documents taken from two sources: the New York Times (NYT) newspaper and the Xinghua English (XIE) corpus. There are 2,631 documents in total. The documents have already been tokenised, and there is no need for any further lexical processing of the documents. You can access the documents here:
/usr/local/practicals/ir/practical1/docs/
There is also a gzipped tar file containing the complete document collection if you wish to copy this onto your own machine.

Inverted File Index

An inverted file index for the complete collection can be found here:
/usr/local/practicals/ir/practical1/data/index.txt .
The format of each line is:
term, document_frequency, doc_1, term_frequency_1, doc_2, term_frequency_2, . . ., doc_n, term_frequency_n
So, each word is followed by its document frequency, and then the list of documents in which it occurs together with the corresponding term frequency for that document. You will need to read in this index and store it in a suitable data structure. One possibility is to use two hash tables: one mapping from terms onto document frequencies, and the other mapping from terms onto lists of (doc, freq) pairs.

Euclidean Lengths

Another file you will find useful is the following:
/usr/local/practicals/ir/practical1/data/doc_lengths.txt

This contains the Euclidean lengths of all the documents in the collection, where the weight for each vocabulary term is simply term_frequency × inverse_document_frequency. The vocabulary used for calculating the lengths (and for building the index) was created by taking all words which occur between 6 and 1,600 times in the collection.

Information Need 1

The information need for this practical is information relating to "financial instruments being traded on the American stock exchange". You should form a query by taking all the words from this description which appear in the index. For evaluation purposes, the relevant set of documents has been identified in advance and the document ids are listed in the file:
/usr/local/practicals/ir/practical1/data/relevant.txt

TF-IDF Weights

You are required to develop a vector space model in which a document is represented by a vector in the normal way, with the basis vectors given by the terms in the index. The weights should be simply the product of the term frequency and the inverse document frequency (rather than a more elaborate function). As for the query, you should use weights of 1 and 0 for the query terms, i.e., either a term is in the query or it is not.

For the purpose of these practicals, TF is just the term frequency, i.e., the number of times the term appears in the document; there is no need to use a more complex expression involving logs. Similarly for IDF: this is just the inverse of the number of documents in which a term appears.

Similarity Measure

You should use the cosine measure to compare the query and document vectors. You may find the slide entitled "Evaluation of Cosine Measure" in the lecture on "Document and Text Representation" a helpful resource for implementing the retrieval algorithm.

Output and Evaluation

The output should be a list of ranked document ids together with the document's score. This output can then be compared against the list of relevant documents:
/usr/local/practicals/ir/practical1/data/relevant.txt
You should evaluate your system using precision-at-recall values. For example, if

Rank	Precision	Recall	Recall	Precision
1	1.00	0.1	0.1	1.00
2	1.00	0.2	0.2	1.00
3	0.67	0.2	0.3	0.60
4	0.50	0.2	0.4	0.57
5	0.60	0.3	0.5	0.63
6	0.50	0.3	0.6	0.50
7	0.57	0.4	0.7	0.54
8	0.63	0.5	0.8	0.57
9	0.56	0.5	0.9	0.60
10	0.50	0.5	1.0	0.59
11	0.45	0.5		
12	0.50	0.6		
13	0.54	0.7		
14	0.57	0.8		
15	0.60	0.9		
16	0.56	0.9		
17	0.59	1.0		

Table 1: Example of precision and recall values at rank; and corresponding precision-at-recall values

your evaluation script produces precision and recall values at rank, as in the left part of Table 1, the precision-at-recall values are as in the right part of Table 1. Finally, your new evaluation script should produce an average precision value, which in this example would be 6.6 / 10 = 0.66.

Information Need 2

As a final test, consider the situation where the user has a better defined information need, and more knowledge of the system and document collection. The query the user creates is: { *stocks, shares, stock, market, exchange, New, York, traded, trading* }. Evaluate your system with this query against the same set of relevance judgements.

Summary of What You Need to Do

1. Read the instructions for Practical 2; this may affect your design;
2. Build a document retrieval system using TF-IDF and the similarity measure as instructed;
3. Retrieve the documents from the given data set which are relevant to Information Need 1, using the supplied inverted file index;

4. Evaluate your system, as instructed;
5. Retrieve the documents from the given data set which are relevant to Information Need 2, and evaluate as before;
6. You may find that on the corpus used in this practical, normalizing by document length may actually hurt your retrieval results. Provide some explanations of why this might happen, and whether, in this circumstance, it would still make sense to normalize.
7. Write a short report.

Your Retrieval System and Assessment

The languages we recommend for building your retrieval system are Java, Perl or Python (these will be supported by the practical demonstrators). However, you can implement the assignment in any programming language you choose. Before you design your system, you should read the instructions for Practical 2; this may affect your design.

You should expect to have completed this practical by the end of session 3, and have it signed-off by one of the demonstrators. When checking your work the demonstrator will want to see a working version of the program in action, as well as appropriate commenting of the code. You will write a combined report for both practicals that will be due at the end of the 5th session.

The practical report does not carry any weight towards the end-of-term assignment, but counts towards the requirement that you achieve an overall pass in your practicals.