University of Texas at Dallas
CS 6322 : Information Retrieval
Fall 2011
Instructor: Dr. Sanda Harabagiu
Grader: Bryan Rink
Issued: November 7th 2011
Due November 28th 2011 before midnight

Problem (100 points)
Ranked Retrieval

In this assignment you will implement a simple statistical retrieval system, using the inverted list index that you built in the last assignment. The system should retrieve the documents that satisfy the natural language queries from the file:

/net/core/export/home/cs/002/s/sanda/cs6322/hw3.queries

The retrieval system must read a query, index it (i.e., parse it, discard stop-words, stem terms, etc), and then determine scores for documents by summing the tf.idf weights for every matching query-document term pair.
Implement and compare two tf.idf term weighting functions.

W1 = (0.4 + 0.6 * log (tf + 0.5) / log (maxtf + 1.0))
* (log (collectionsize / df)/ log (collectionsize))

W2 = (0.4 + 0.6 * (tf / (tf + 0.5 + 1.5 *
(doclen / avgdoclen))) * log (collectionsize / df)/
log (collectionsize))

tf: the frequency of the term in the document,

maxtf: the frequency of the most frequent indexed term in the document,

df: the number of documents containing the term,

doclen: the length of the document, in words, counting stopwords,

avgdoclen: the average document length in the collection, and

collectionsize: the number of documents in the collection.

W1 is a variation of older, but well-known, 'max tf' term weighting. W2 is a variation on Okapi term weighting. Both TW1 and TW2 use a fairly standard scaled idf.

Documents should be presented in ranked order of the total scores. Implement a simple terminal or Web interface for query entry and document display.

For each query, turn in the indexed form of the query, and the top 10 documents for the query under both weighting schemes (you may build two different systems if you think that's simpler). Indicate the rank, score, external document identifier, and headline, for each of the top 10 documents for each query. Identify which documents you think are relevant and non-relevant for each query. Describe why the top-ranked non-relevant document for each query did not get a lower score. Briefly discuss the different effects you notice with the two weighting schemes, either on a query-by-query basis or overall, whichever is most illuminating.

Describe the design decisions you made in building your system.