University of Texas at Dallas
CS 6322 : Information Retrieval
Spring 2014
Instructor: Dr. Sanda Harabagiu
Grader: Ramon Maldonado

Issued: March 30th 2015
Due April 20th 2015 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 queries from the file:

/people/cs/s/sanda/cs6322/hw3.queries

The retrieval system must read a query, parse it, discard stop-words, generate the lemmas for the content words and then determine scores for documents against the queries by summing the weights for every matching query-document term pair. Implement and compare two 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, discounting stop-words, - you may use the same stopword list as in the previous homework;

avgdoclen: the average document length in the collection, considering the doclen of each document, 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.
FOR each query:

1. Turn in the indexed form of the query, and the **top 5 documents** for the query under both weighting schemes (you may build two different systems if you think that's simpler). (*70 points)*
2. Indicate the rank, score, external document identifier, and headline, for each of the top 5 documents for each query. (*5 points*)
3. Identify which documents you think are relevant and non-relevant for each query. (*10 points*)
4. Describe why the top-ranked non-relevant document for each query did not get a lower score. (*5 points*)
5. 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. For example, you can point out that the weighting scheme seems to be working for this query as well as a list of other queries, but not for some other queries you have noticed. Try to explain why it works and why it does not work.(*5 points)*
6. Describe the design decisions you made in building your ranking system. (*5 points*)