HW 2

CS 582 Information Retrieval
Spring 2020
Total points: 100
Issued: 02/03/2020 Due: 02/18/2020

Three day late submission is possible, but not encouraged. A late submission incurs a 10% penalty per day. No credit is given to submission more than three days late.

Your task is to implement a basic vector space retrieval system. You will use the Cranfield collection to develop and test your system.

The Cranfield collection is a standard IR text collection, consisting of 1400 documents from the aerodynamics field, in SGML format. The dataset, a list of queries and relevance judgments associated with these queries are available from Blackboard.

Tasks: To complete this assignment, you need to use the pre-processing tools implemented during assignment 1. Note that you also need to eliminate the SGML tags (e.g., <TITLE>, <DOC>, <TEXT>, etc.) - you should only keep the actual title and text.

For text pre-processing, remove stopwords, perform stemming (note: if a word becomes a stopword after stemming, please remove it), remove punctuation and numbers (replace them with “”), split on whitespace, and remove words with one or two characters in length. Perform the same text processing operations on both the documents and the queries.

1. Implement an indexing scheme based on the vector space model, as discussed in class. The steps pointed out in class can be used as guidelines for the implementation. For the weighting scheme, use and experiment with:
   - TF-IDF
2. For each of the ten queries in the queries.txt file, determine a ranked list of documents, in descending order of their cosine similarity with the query. The output of your retrieval should be a list of (query_id, document_id) pairs.

   Determine the average precision and recall for the ten queries, when you use:
   - top 10 documents in the ranking
   - top 50 documents in the ranking
   - top 100 documents in the ranking
   - top 500 documents in the ranking

Note: A list of relevant documents for each query is provided to you, so that you can determine precision and recall.

Submission instructions:

1. write a README file including:

- a detailed note about the functionality of each of the above programs,
- complete instructions on how to run them
- answers to the questions above

2. make sure you include your name in each program and in the README file.
3. make sure all your programs run correctly on the CS machines. You will lose 40 points if your code is not running on these machines. The path to the data should be an input parameter, and not hardcoded.
4. submit your assignment through Blackboard.