Cmpe 493 Introduction to Information Retrieval, Fall 2015
Assignment 1 - A Simple Document Retrieval System for Boolean Queries, Due: 06/11/2015 (Friday), 17:00

In this assignment you will implement a document retrieval system for simple boolean queries using the inverted indexing scheme. You will use the Reuters-21578 data set, which can be download from Moodle (reuters21578.zip). Reuters-21578 contains 21578 news stories from Reuters newswire classified under one or more of 118 categories. There are 21 SGML files, each containing 1000 news articles, except the last file, which contains 578 articles.

You should perform the following steps:

1. Pre-processing the Data Set : The text of a news story is enclosed under the <TEXT> tag. You should use the <TITLE> and the <BODY> fields to extract the text of a news story. Implement your own tokenizer to get the tokens from the news texts and perform normalization operations like case-folding, stopword removal, and stemming. You can use the stopword list on Moodle (stopwords.txt). You can stem each token using the Porter Stemmer (http://tartarus.org/martin/PorterStemmer/). Note that you should perform the same preprocessing steps for the queries as well.

2. Building the Inverted Index: Instead of taking each SGML file as a document unit, you should index each news article as a separate document and use the $NEWID$ field as document IDs.

3. Implementing a Boolean query processor: You should implement the postings merge algorithm for conjunctive and disjunctive queries. Assume that each query consists of either a conjunction or a disjunction of terms. You do NOT need to handle phrase queries, proximity queries, wildcard queries, or queries containing the NOT operator or parenthesis. That is, the queries will be of the following two types:
   (i) w_1 AND w_2 AND w_3... AND w_n
   (ii) w_1 OR w_2 OR w_3... OR w_n
   where each w_i is a single-word keyword.

You may use any programming language of your choice. However, we should be able to run your program by following the instructions in your readme file.

Submission: You should submit a *".zip"* file named as YourNameSurname.zip containing the following files using the Moodle system:

1. Report:
   (i) Describe the steps you have performed for data preprocessing and provide answers for the following questions.

   (a) How many tokens does the corpus contain before stopword removal and stemming?

   (b) How many tokens does the corpus contain after stopword removal and stemming?

   (c) How many terms (unique tokens) are there before stopword removal, stemming, and case-folding?

(d) How many terms (unique tokens) are there after stopword removal, stemming, and case-folding?

(e) List the top 20 most frequent terms before stopword removal, stemming, and case-folding?

(f) List the top 20 most frequent terms after stopword removal, stemming, and case-folding?

(ii) Describe the data structures that you used for representing the inverted index (dictionary and postings).

(iii) Provide a screenshot of running your system for a conjunctive query.

(iv) Provide a screenshot of running your system for a disjunctive query.

2. Source code and executable: Commented source code and executables of your document retrieval system.

3. Readme: Detailed readme describing how to run your program.

Late Submission: You are allowed a total of 5 late days on homeworks with no late penalties applied. You can use these 5 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 3 days late. In that case you will have to submit the remaining homeworks on time. After using these 5 extra days, 10 points will be deducted for each late day.