Cmpe 493 Introduction to Information Retrieval, Fall 2021
Assignment 1 - A Simple Search System for Boolean Queries
Due: 17/11/2021 (Wednesday), 17:00

In this assignment you will implement a document retrieval system for simple boolean queries using the **non-positional** inverted indexing scheme. You will use the Reuters-21578 data set, which is available on Moodle (reuters21578.zip). Reuters-21578 contains 21578 news stories from Reuters newswire. There are 22 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 including case-folding, stopword removal, and punctuation removal. Please use the stopword list on Moodle (stopwords.txt) for stopword removal and you can use the list in "string.punctuation" for punctuation removal for Python. Please use "latin-1" encoding while you read the .sgm files due to the corruption in one file.

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. Then, you need to create an inverted index consisting of the dictionary and the postings lists. You should store your inverted index as a file(s) and during query processing you should only use the inverted index, not the original Reuters-21578 dataset. We will delete the "reuters21578" data set while testing the query processor. Note that the inverted index construction and query processor should be designed as two separate modules. That is, the query processor should NOT construct the inverted index each time it is run. It should just use the inverted index file(s) generated by the indexing module.

3. Implementing a query processor: You should implement a query processor for Boolean queries constructed using the AND, OR or NOT operators by using the postings merge algorithm. That is, the queries will be of the following four types (here w_i is a single-word keyword):

   (i) Conjunction: w_1 AND w_2 AND w_3...AND w_n
   (ii) Disjunction: w_1 OR w_2 OR w_3...OR w_n
   (iii) Conjunction and Negation: w_1 AND w_2...AND w_n NOT w_{n+1} NOT w_{n+2} ...NOT w_{n+m}
   (iv) Disjunction and Negation: w_1 OR w_2...OR w_n NOT w_{n+1} NOT w_{n+2} ...NOT w_{n+m}

   The query processor should take as input a query and return the IDs of the matching documents sorted in **ascending order**. Here are some example input queries:

   - price AND oil
   - petroleum OR oil OR gas
   - price AND oil NOT vegetable
   - petroleum OR oil NOT price

You should use Python to implement your search system. We should be able to run your program by following the instructions in your readme file. You have to state the exact commands to run the indexing module and to run the query processing module. You should NOT use any third party libraries.

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.

   (ii) Describe the data structures (hash, b-tree, linked list etc.) that you used for representing the inverted index (i.e., the dictionary and the postings lists).

   (iii) Provide a screenshot of running the indexing module of your system.

   (iv) Provide four screenshots of running your system for each of the four types of queries.

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

3. Readme: Detailed readme describing how to run your program including the Python version.

Honor Code: You should work individually on this assignment and all the source code should be written by you. You are NOT allowed to use any available libraries or any code written by other people. Violation of the Honor Code will be strictly penalised, not only by a zero grade from the homework, but also by filing a petition to the Disciplinary Committee.

Late Submission: You are allowed 5 late days (until 22 November 17:00 o'clock) for this assignment with no late penalty. After 22 November 17:00 o'clock, 1 point will be deducted for each late hour (unless you have a serious excuse).