CS 599(02) - Information Retrieval

Programming Assignment #1

<u>Due date</u>: 10/8/14 14:00 hrs

Statement

The main aims of a text based retrieval system is to retrieve documents based on user query. For the problem of search, it is very important to serve user's query as fast as possible, which also implies that (1) index should be as efficient as possible for all types of queries (2) size of index has to be as small as possible, so that it fits in memory making search fastest possible (3) indexing can take more time, as user is not waiting for it.

For this assignment, we are not going to employ any ranking/scoring, instead we will focus on retrieving all related documents. To achieve that, we will use Boolean Retrieval mechanism. After accuracy of retrieval, credit will be given based on efficiency (smaller the index size and query time).

Details

You have to implement two methods for Boolean Retrieval:

- `buildIndex(String dir)`
    - Build index supposed to go over all files under 'dir'. There will be no subdirectories inside it.
- `Search()`
    - Search supposed to return name of all the documents under 'dir' that satisfies the given query. Note that, only basename of the files are to be returned, not the full path.
    - Query can be of two forms:
        - OR: returned doc should contain at least one term from the query.
        - AND: returned doc should contain all the terms from the query.

Hint: Input documents have been collected from real data-set. That means, one might need to employ some kind of tokenizer to split a doc into a vector of terms.

Code

Download PA1.zip from the Blackboard under Assignments → Programming Assignments → PA1.

Code Structure

`data/` → This directory contains some news data under `test`/, which will be used as corpus for this assignment. `test.txt` is the input search operation type, query and the retrieved documents.
`java/` → This directory contains all the JAVA code. You only need to modify `BooleanRetrievalModel.java`
`run.py` -> A tool to compile, run & test your code.

Extract and Run

Download the file PA1.zip. Extract it. Let's say you extracted into 'PA1' directory.
From terminal:

```
$cd PA1
# To run and evaluate your code
$./run.py
```

Compatibility

Note that, all assignments (including this one) will be tested under Linux environment with Python and Oracle Java is installed. Given code might work on other platforms (like Windows, etc.) but has not been tested. Hence, it is encouraged to develop and test your code in a Linux based environment.

Submission

You should only modify and upload `BooleanRetrievalMOdel.java` to Blackboard. **Any change in other files will not be accepted and you will not be evaluated in that case**.

Evaluation

There is some held out data and query set against which your code will be tested and evaluated. Hence, it is encouraged not to write generic code that work in most of the cases. Your main aim is to write an efficient data structure that should index and retrieve the documents very fast with the query operations. You are free to read different techniques and use them to improve your accuracy. You will be evaluated on the basis of accuracy in retrieval, speed and memory usage.

Honor Code

I encourage students to discuss the programming assignments including specific algorithms and data structures required for the assignments. However, students should not share any source code for solution.

Code exists on the web for many problems including some that we may pose in problem sets or assignments. Students are expected to come up with the answers on their own, rather than extracting them from code on the web. This also means that we ask that you do not share your solutions to any of the homework - programming assignments, or problem sets - with any other students. This includes any sort of sharing, whether face-to-face, by email, uploading onto public sites, etc. Doing so will drastically detract from the learning experience of your fellow students.