CSCI 4130 Information Retrieval

Project 1

Build an inverted index for the given corpus and apply the search algorithms on the given Boolean queries.

High-level solution Steps:

1. Normalize text – address punctuation characters, stemming/lemmatization, and lowercasing. Keep the stop words.
2. Extract tokens and identify vocabulary for the dictionary.
3. Scan the corpus and build a positional index.
4. Implement the intersecting algorithm for queries.
5. Develop a simple interface for users to specify queries.
6. Design test cases and execute them.

A partial solution on tokenization is provided as Ngram.java, please read and try the code thoroughly and pick the part you need for your program.

What to submit:

1. Your source code
2. A report include
    a. Team members
    b. An instruction on how to use your system for searching
        i. Use some screenshots for the instruction
    c. Test cases and test results