

Project 2: **Indexer**  
Search Engine Development  
Due: Mon, April 7 by midnight  
100 points

You are to create an indexer for the web pages you have crawled. The pages will all be in the cache directory, as deposited by your web crawler. Each web page will be named using the MD5 hash of the web page's normalized URL. The URLs crawled will be in the file urls.dat. We'll assume the pages are in English.

For simplicity, the indexer should be executed as a command line program. Your program should go through the following process:

1. Iterate through all the HTML files in the cache directory and extract the title, keyword and description metatags, and the body. Any HTML in the body of the web page should be ignored. For simplicity, we will not consider anchor text.
2. All text should be converted to lowercase and tokenized by replacing all non-alphanumeric characters with spaces. Each word/token is then extracted, and the following stop words should be removed: a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with.
3. The Porter stemming algorithm should be applied to the tokenized text. You can find out more about the algorithm and a Java implementation which you may use here: <http://tartarus.org/~martin/PorterStemmer/>
4. A single inverted index should be created for the document corpus which maintains the term frequency, the document ID the term is found in, the position(s) of the term in the document (1 is first word, 2 is second, etc.) , a 1/0 indicating the term's occurrence in the title, 1/0 for occurrence in the metatags, and the term frequency in the body.
5. You should create a special class called InvertedIndex to store the inverted index. The class should have the following public methods: saveToFile(String filename), readFromFile(String filename), and print().
6. After creating the InvertedIndex, your program should print it to the screen (using the print method) save it to a file called index.dat (using the saveToFile method).

Your index will be used in the next project to perform queries using a servlet, so you may want to keep this in mind when writing your code so minimal code changes will be necessary.

Submit your completed **WebIndexer.java** file to Easel before class on the day it is due. (The file must be named correctly.)