LocateInside

Like Unix 'locate', but searches inside files

The Problem

Unix systems have 'locate' command for fast file searching, which used a prebuilt database to speed-up the search. However, it looks like there is no standard command for similar search inside files based on prebuilt index (like in Windows and Mac OS X).

The alternative solution is to use 'grep' (http://stackoverflow.com/questions/16956810/finding-all-files-containing-a-text-string-in-linux and http://www.cyberciti.biz/faq/howto-search-find-file-for-text-string/), but it doesn't use index and have to do a full scan of the content of the files. Much bigger problem is that 'grep' will only look for plain-text string, so it will probably fail for PDF or Word documents.

The Solution

So I decided to create a 'locateinside' utility, which would solve the above problem. It has the following features:

1) The program looks for filed in some directory of interest containing the specified text;

2) The program indexes files in the directory of interest to create a reusable index to speed up the search. This is done using Apache Lucene;

3) It uses Apache Tika and thus indexes not only plain-text files, but also PDF, HTML, OpenDocument, RFT, Word documents (and some more: http://tika.apache.org/1.4/formats.html);

4) It uses one system-wise index (stored in a temp directory at the moment), but during a request it works only with a part of it responsible for the directory of interest;

5) When updating the index for a specified directory, it doesn't just reindex all files, but looks for the changed files. For that it at the same time (while indexing documents) reads the index, extracts the information about indexed files contained in the index before (the modification date), and compares it with the existing files. Deleted files are handled as search time for simplicity.

Usage

The program should work in Windows, but was tested only for Linux.

The program directory contains src folder with sources, and bin folder with executables. Folder bin contains shell script 'locateinside', which should have executable right.
Imagine we are in the directory of LocateInside (~/locateinside). Then we can execute it for the current directory:

bin/locateinside i AND love AND you

This will create the index of the program directory and look for words 'i', 'love' and 'you'. Only file, 'src/main/java/locateinside/LocateInside.java', contains all three words, so the full path of that file should be the result.

When invoked on a machine with an existing index it does not update the index for performance reasons. To force index update (after some files were modified) add '-u' option:

bin/locateinside -u i AND love AND you

JEVGENIJA PANTIUCHINA