IndexingOtherLanguages

How To Index non-English Languages using Lucene

Lucene is a Java based, UNICODE-compatible library for integrating searching into applications.

With a little extra effort, it is quite easy to index and search non-English language based documents (and even search non-English based documents using English!)

This document will not go into the details of how to setup Lucene to index and search (using readers, etc.), those are best covered in other pages such as I ntroductionToLucene and other HowTo tutorials, as well as many excellent articles available online. It is also assumed the reader understands how the Lucene Analyzer works (if not, see IntroductionToLucene and AnalysisParalysis.)

There are several key items you will need to consider when indexing

1. Know the encoding of the documents you wish to index. Java assumes the native encoding when reading in files unless you tell it otherwise. To create a Reader that supports reading in other encodings, see InputStreamReader. I find it easiest to convert all of my files to UTF-8 before indexing, and then I read them in by doing:

```
Reader reader = new InputStreamReader(new FileInputStream("path to file"), "UTF-8");
```

2. Identify the Analyzer you will use or write your own if none exists. There are many great analyzers available that will index a wide variety of languages. See Sandbox for some. Otherwise, look around the web. If you are writing your own, consider donating it to the Lucene Sandbox so that others can benefit from your brilliance. See item 3. below for what is needed in a custom analyzer.
'Put example of writing an Analyzer here'
3. The key to proper analysis is to identify what you want your final tokens to be. Do you want them tokenized, stemmed, lowercased, all stop words removed, etc. With non-English languages, many people have a hard time finding tokenizers and stemmers for the language they are interested in. There are many great sites out there that provide solutions to these problems, one just needs to look. Often times, a simple google search for something like "arabic tokenizer" will do the trick. Other times, you may need to dig into some academic papers to find a description of the problem. Another great resource is the Lucene User mailing list archives. Chances are you aren't the first one to tackle the language.

Once you have your Analyzer setup and your documents indexed, take a look at the Index using Luke

Searching is just as in the English case. Make sure you use the same analyzer you did for indexing when analyzing your search.