Tech Review

Overview

As Lucene is a 20 year old project, widely known and used, I don't think I could really offer anything new or insightful on the topic. I will, instead, focus simply on what I need to complete my project and the features of Lucene that I might personally use for that purpose.

My project is essentially just a search engine, but the specific things that I need from Lucene are:

- Process new pages/documents as they come in, index them with their url as the key
- Store/load this index
- Query this index

In this document I'll go through the Lucene documentation [1], picking out things that are likely to be useful. This should all (hopefully) prove rather simple, considering Lucene *is* a search library.

Indexing

Starting from the top, I need to be able to append documents to the index in real time (something I couldn't find any mechanism for within MeTA [2], which I initially thought to use, leading me to Lucene). As the user browses, I want to be able to index the contents of all pages they open. Ideally, I wouldn't need to save the pages to disk before indexing them.

> IndexWriter is used to create an index, and to add, update and delete documents. The IndexWriter class is thread safe, and enforces a single instance per index. Creating an IndexWriter creates a new index or opens an existing index for writing, in a Directory, depending on the configuration in IndexWriterConfig. A Directory is an abstraction that typically represents a local file-system directory (see various implementations of FSDirectory), but it may also stand for some other storage, such as RAM. [3]

Promising! (though I'm not seeing any mention of a possibility to stream in files one-by-one)

```
  Analyzer analyzer = new StandardAnalyzer();

Path indexPath = Files.createTempDirectory("tempIndex");
Directory directory = FSDirectory.open(indexPath);
IndexWriterConfig config = new IndexWriterConfig(analyzer);
IndexWriter iwriter = new IndexWriter(directory, config);
Document doc = new Document();
String text = "This is the text to be indexed.";
doc.add(new Field("fieldname", text, TextField.TYPE_STORED));
iwriter.addDocument(doc);
iwriter.close();
```

[4]

So, as I build the documents myself, I don't need to read them off disk!

Store [5]

I obviously need the index to persist beyond a single browsing session. What index storage mechanisms are available in Lucene? What is the most efficient way to do this?

[1] https://lucene.apache.org/core/9_4_1/index.html
[2] https://meta-toolkit.org/
[3] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/index/package-summary.html
[4] https://javadoc.io/doc/org.apache.lucene/lucene-core/latest/index.html
[5] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/store/package-summary.html

- Store document indicies
  - not necessarily in any human readible format
  - How do I make this persist on disk?

This appears to happen by default. See the `indexPath` and `directory` lines above.

Querying [6]

Not much to add right now. Need to be able to search the index or it's useless.

- query index
  - return matching urls
  - context would be nice, but not necessary

PyLucene

- PyLucene
  - I can't be assed to get all this java shit working
  - PyLucene should work the same
  - Wrapper around lucene, not a port

[6] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/search/package-summary.html