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). From the example in the javadoc[4]:

```
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();
```

So, as I build the documents myself, I don't need to read them off disk! Additionally, "fieldname" is set manually, and isn't some random/automatic id, so I can stick source URLs there.

What is all that?

- What is `Analyzer`?

[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

An Analyzer builds TokenStreams, which analyze text. ***It thus represents a policy for extracting index terms from text.*** [5]

`StandardAnalyzer` appears to be general purpose, but there are also one that specialize in, for example, Japanese text.

- What is `IndexWriterConfig`?

  Holds all the configuration that is used to create an IndexWriter. [6]

- What is `IndexWriter`?

  An IndexWriter creates and maintains an index. [7]

- What is `Document`?

  Documents are the unit of indexing and search. [8]

- What is `TextField.TYPE_STORED`?

  A field that is indexed and tokenized, without term vectors. For example this would be used on a 'body' field, that contains the bulk of a document's text. [9]

  It seems I may be able to index pages *without* storing the contents by using `TextField.TYPE_NOT_STORED`, though I'm not sure how that works. Look into this. I don't particularly need or want to store the pages themselves

How could I get PDFs to index as well?

Store [10]

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?

- 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. Though, I imagine there are some more things to look at (or there wouldn't be a storage package)

Querying [11]

More from the example on the index[12]:

```
// Now search the index:
DirectoryReader ireader = DirectoryReader.open(directory);
IndexSearcher isearcher = new IndexSearcher(ireader);
// Parse a simple query that searches for "text":
QueryParser parser = new QueryParser("fieldname", analyzer);
Query query = parser.parse("text");
ScoreDoc[] hits = isearcher.search(query, 10).scoreDocs;
assertEquals(1, hits.length);
```

[5] https://javadoc.io/static/org.apache.lucene/lucene-core/9.4.1/org/apache/lucene/analysis/Analyzer.html
[6] https://javadoc.io/static/org.apache.lucene/lucene-core/9.4.1/org/apache/lucene/index/IndexWriterConfig.html
[7] https://javadoc.io/static/org.apache.lucene/lucene-core/9.4.1/org/apache/lucene/index/IndexWriter.html
[8] https://javadoc.io/doc/org.apache.lucene/lucene-core/latest/org/apache/lucene/document/Document.html
[9] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/document/TextField.html
[10] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/store/package-summary.html
[11] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/search/package-summary.html
[12] https://javadoc.io/doc/org.apache.lucene/lucene-core/latest/index.html

```
// Iterate through the results:
for (int i = 0; i < hits.length; i++) {
  Document hitDoc = isearcher.doc(hits[i].doc);
  assertEquals("This is the text to be indexed.", hitDoc.get("fieldname"));
}
ireader.close();
directory.close();
IOUtils.rm(indexPath);
```

What is all that?

- What is `IndexSearcher`?

  > Implements search over a single IndexReader. Applications usually need only call the inherited search(Query,int) method. For performance reasons, if your index is unchanging, you should share a single IndexSearcher instance across multiple searches instead of creating a new one per-search. If your index has changed and you wish to see the changes reflected in searching, you should use DirectoryReader.openIfChanged(DirectoryReader) to obtain a new reader and then create a new IndexSearcher from that. Also, for low-latency turnaround it's best to use a near-real-time reader [13]

  May require some work here, as my index will receive updates pretty regularly (every page load)

- What is `QueryParser`?

  > QueryParser parses the user query string and constructs a Lucene Query object [...] The first parameter to the QueryParser constructor specifies the default search field, which is content field in this case. This default field is used if the query string does not specify the search field. The second parameter specifies the Analyzer to be used when the QueryParser parses the user query string. [14]

- What is `Query`? Representation of a user query.
- What is `ScoreDoc`? List of documents and their scores. It appears that you can choose the retrieval model used for scoring [15] Doesn't actually return the documents themselves, but contains the documents "number" in the index, score, etc The `10` it's taking is the number of results we want to get back.
- What is being returned in `hitDoc`? Gets the actual document referred to in ScoreDoc

PyLucene

Having trouble setting up a java environment?

[13] https://javadoc.io/static/org.apache.lucene/lucene-core/9.4.1/org/apache/lucene/search/IndexSearcher.html
[14] http://web.cs.ucla.edu/classes/winter15/cs144/projects/lucene/index.html
[15] https://lucene.apache.org/core/9_4_1/core/org/apache/lucene/search/package-summary.html