3-5 Tools for Content-Based Filtering

Introduction

- You probably don't want to build from scratch
  - Especially for prototyping
- Many tools are available for building recommenders
- This lecture: a brief survey of a few of them (particularly open-source ones)

LensKit

- Provides high-level recommendation framework
- Does not currently provide content-based recommender implementations
- Helps with the plumbing for small- or medium-scale applications

Search Software

- Many content-based filtering algorithms are search algorithms (incl. TF-IDF)
- Existing search software useful to build recommenders

Apache Lucene

- Open-source Java search package
- Very widely used
- Provides document indexing w/ arbitrary fields and fast search
- Several relevance and ranking algorithms
- Good, out-of-the-box performance

Using Lucene

1. Create an index
2. Add ‘document’ representations of items
3. Construct queries
4. Ask for results (will be scored)

Building the Index

```
IndexWriterConfig config = /* configure */ ;
Directory dir = FSDirectory.open(indexFile);
IndexWriter w = new IndexWriter(dir, config);
for (ItemInfo item: getItems()) {
    Document doc = new Document();
    doc.add(new Field("title", item.title));
    doc.add(new Field("tags", item.tags));
    w.add(doc);
}
w.close();
```

Finding Similar Items

```
IndexSearcher idx = getIndexSearcher();
IndexReader reader = idx.getIndexReader();
Document itemDoc = findItemDoc(idx, item);
MoreLikeThis mlt = new MoreLikeThis(reader);
String[] fields = {"title", "tags"};
mlt.setFieldNames(fields);
Query q = mlt.like(docid);
TopDocs results = idx.search(q, n + 1);
```

Related Projects

- SOLR provides search server (with REST API) on top of Lucene
- PyLucene is Python implementation
- Lucy is in C w/ bindings for other langs
- Lucene.NET
- Semantic Vectors provides LSI for Lucene

Other Search Software

- Lemere (C++, Univ. of Maryland)
- Xapian (C++)
- Any search package is a valuable tool

Case-based Reasoning

- jCOLIBRI is a Java-based CBR framework

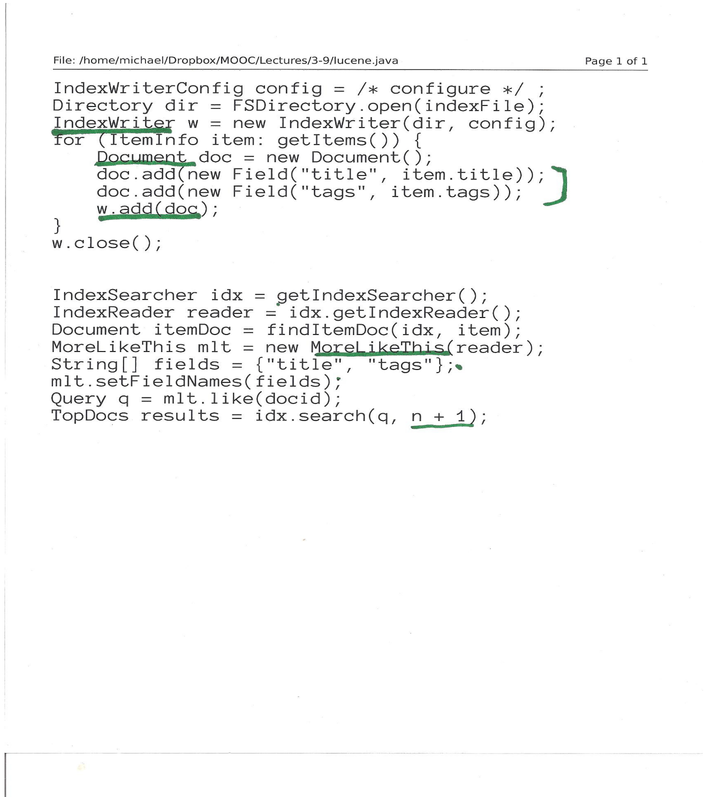
Machine Learning Toolkits

- Recommendation algorithms can be seen as a special-case of machine learning
- Machine learning libraries can be useful to implement recommenders
  - Weka
  - Apache Mahout
  - Milk, SciPy/NumPy (Python)
  - R, Matlab

Just Scratch the Surface

- There are many toolkits
  - Some general-purpose
  - Some special-purpose
- These are just a few, that may be helpful particularly for CBF recommenders
- Discuss others in the forums

3-5 Tools for Content-Based Filtering



File: /home/michael/Dropbox/MOOC/Lectures/3-9/lucene.java Page 1 of 1

```
IndexWriterConfig config = /* configure */ ;
Directory dir = FSDirectory.open(indexFile);
IndexWriter w = new IndexWriter(dir, config);
for (ItemInfo item: getItems()) {
    Document doc = new Document();
    doc.add(new Field("title", item.title));
    doc.add(new Field("tags", item.tags));
    w.add(doc);
}
w.close();


IndexSearcher idx = getIndexSearcher();
IndexReader reader = idx.getIndexReader();
Document itemDoc = findItemDoc(idx, item);
MoreLikeThis mlt = new MoreLikeThis(reader);
String[] fields = {"title", "tags"};
mlt.setFieldNames(fields);
Query q = mlt.like(docid);
TopDocs results = idx.search(q, n + 1);
```