Lucene

0.1 Documentation

by Michael Wechner

1. Generic Search

URL:

```
/lenya/$PUB_ID/search-$AREA/lucene
```

Indices and Excerpts:

```
src/webapp/lenya/pubs/$PUB_ID/work/search/index/$AREA/index
src/webapp/lenya/pubs/$PUB_ID/work/search/htdocs_dump/$AREA
```

Configuration:

```
src/webapp/global-sitemap.xmap
```

2. Customizing/Overwriting Generic Search Interface

XSLT:

```
src/webapp/lenya/pubs/$PUB_ID/lenya/xslt/search/search-and-results.xsl
```

URL:

```
/lenya/$PUB_ID/osearch-$AREA/lucene
```

3. Crawling a website

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml -Dcrawler.xconf=/home/username/src/coco
```

4. Creating an index from the command line

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml -Dlucene.xconf=/home/username/src/cocoo
```

or

```
/usr/local/j2sdk1.4.2/bin/java -classpath build/lenya/webapp/WEB-INF/classes:build/leny
```

5. Extract text from a PDF document

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml -Dhtdocs.dump.dir=/home/username/src/co
```

Also see the targets `pdfbox` and `pdfadobe`.