

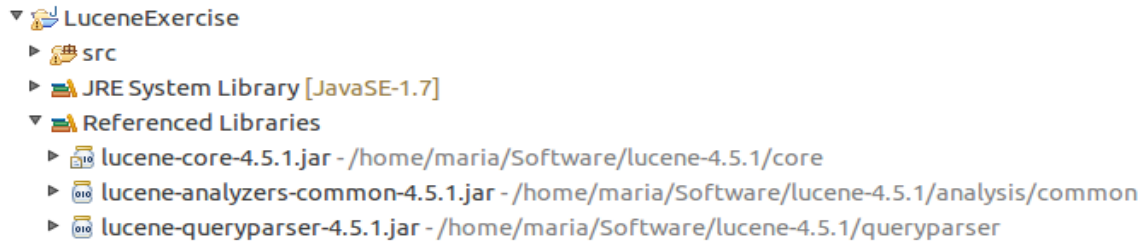
Упражнение върху Lucene

0. Свалете Lucene 4.5.1 от <http://lucene.apache.org/core/downloads.html> и го разархивирайте.

1. Създайте нов java проект в eclipse; добавете като външни библиотеки(external libraries/jars) в build path(давате Configure build path... на проекта) следните 3 библиотеки:

- lucene-4.5.1/queryparser/lucene-queryparser-4.5.1.jar
- lucene-4.5.1/analysis/common/lucene-analyzers-common-4.5.1.jar
- lucene-4.5.1/core/lucene-core-4.5.1.jar

Накрая трябва да получите:



2. Прочетете и проследете наръчника

<http://www.lucenetutorial.com/lucene-in-5-minutes.html>

Разгледайте класа HelloLucene.java

<http://www.lucenetutorial.com/code/HelloLucene.java>

Обърнете внимание на класовете

IndexSearcher(http://lucene.apache.org/core/4_5_1/core/org/apache/lucene/search/IndexSearcher.html) и

IndexReader(http://lucene.apache.org/core/4_5_1/core/org/apache/lucene/index/IndexReader.html)

3. В Lucene документите са фрагментирани на полета(fields). За всеки тип поле се създава отделен индекс(за всички документи, които го съдържат).

Добавете в *addDoc* ново текстово поле в документите, в което се записва "content"(подобно на "title")

4. Създайте нов тип поле – FieldType *type*, на който и добавете като минимум

```
type.setIndexed(true);  
type.setStoreTermVectors(true);  
type.setStored(true);  
type.setIndexOptions(IndexOptions.DOCS_AND_FREQS);
```

5. Изчислете метриката *TF-IDF(term, document)* – тя показва колко думата *term* е определяща за документа *doc* в рамките на колекцията от документи, които са индексирани.

Най-простата формула за *TF-IDF*:

$$TF-IDF(t, doc) = TF(t, doc) * IDF(t)$$

където:

t = term/терм/дума, *doc* = документ

TF(t, doc) е колко пъти се среща *t* в *doc*; term frequency

IDF(t) = $\log_2(N/N_t)$ - намаляващ фактор, при често-срещан в колекцията *t*.

N - броят на всички документи

N_t - броят на документите, които съдържат *t*

За целта добавете метод, който прави това, с примерна сигнатура

```
public double getTFIDF(int docId, String term, String field)
```

Разковничето на решението е класът IndexReader и методи като getTermVector; вижте горния линк и документацията на Lucene.

Внимавайте с версията на Lucene!

Примерни решения:

Мария Матева + Стефан Дойчев: <https://github.com/miaaaooow/LuceneExercise>

Васил Георгиев: <https://github.com/mcrius/lucene/tree/master/src/lucenetest>

ноември, 2013год.

Мария Матева