Team 12 – Šproty

- Indexovanie a vyhľadávanie dokumentov
- Nástroj Lucene

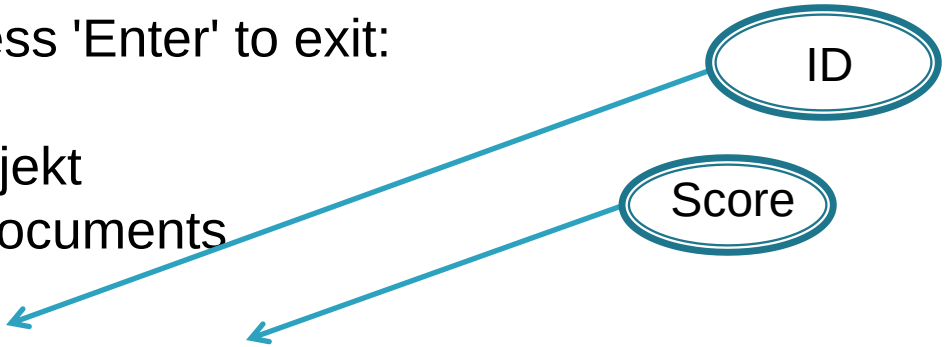
Bc. Ľubomír Eľko

Triedy v Lucene

- **Document -** represents a document in Lucene.
- **Field -** will contain a name for the section and the actual data.
- **Analyzer -** an abstract class that used to provide an interface that will take a Document and turn it into tokens that can be indexed. SimpleAnalyzer, StopAnalyzer and StandardAnalyzer class.
- **IndexWriter -** used to create and maintain indexes.
- **IndexSearcher -** used to search through an index.
- **QueryParser -** used to build a parser that can search through an index.
- **Query -** an abstract class that contains the search criteria created by the QueryParser.
- **Hits -** contains the Document objects that are returned by running the Query object against the index.

Príklad vyhľadania

Enter query or press 'Enter' to exit:
projekt
Searching for: projekt
2 total matching documents



1. [11.doc] (0,16309) C:/Users/lubes.e/test/docs/o_nas.html
Title: Sproty Team Page
Author: team c.12 - sproty
Keywords: sproty, fiit, java, prefuse, foaf, lucene
Description: stranka timu c. 12 (sproty). tema: graficka podpora vyhladavania znalosti v dokumentoch.

<title>Sproty Team Page</title>
<meta name="author" content="Team c.12 - Sproty" />
<meta name="keywords" content="sproty, fiit, java, prefuse, foaf, lucene" />
<meta name="description" content="Stranka timu c. 12 (Sproty). Tema: Graficka podpora vyhladavania znalosti v dokumentoch." />

Problémy I.

Čo keď metadáta chýbajú? Priority:

1. **Riešenie – zapisovanie do metadát**

2. **Riešenie – šablóna a macro pri zavretí**

3. **Riešenie – vlastný parser**
   - keywords - počítanie výskytov slov (najviac slov však bude typu „softvér“, „projekt“, atď., teda nulová inf. hodnota)
   - ale ako zistíme autora alebo názov? [Author: Einstein]

Problémy II.

Diakritika

a) **Konvertor textu do ASCII (prvých 127 znakov)**
   trieda *ASCIIFoldingFilter*

b) **Odstraňovanie diakritiky (Accented character)**
   trieda *ISOLatin1AccentFilter*

Použijeme ***ASCIIFoldingFilter*** nakoľko *ISOLatin1AccentFilter* sa už neodporúča používať a v novej verzii Lucene sa už ani nebude nachádzať.

Problémy III.

Analyzátor na SVK slová

- existuje CzechAnalyzer – prispôsobenie na slovenčinu
- používanie dvoch analyzátorov sa nedoporučuje!

Ako zistiť či ide o SVK alebo ENG dokument?

- prítomnosť smerodajných slov v texte

Ako zistiť či ide o SVK alebo ENG query?

- dá sa z jedného slova určiť o aký jazyk ide a teda aký analyzátor použiť?
- „MySQL“, „network“, „index“, „server“, „open source“, „problem“, „Einstein“

Zdroje a odkazy

Lucene – Wikipédia

http://en.wikipedia.org/wiki/Lucene

Lucene – Dokumentácia

http://wiki.apache.org/lucene-java

Introduction to Text Indexing with Apache Jakarta Lucene

http://onjava.com/pub/a/onjava/2003/01/15/lucene.html

The Lucene Search Engine - Adding search to your applications

http://www.javaranch.com/journal/2004/04/Lucene.html

Lucene – How to accomplish simple tasks?

http://wiki.apache.org/lucene-java/HowTo

Lucene as a Ranking Engine

http://www.wortcook.com/pdf/lucene-ranking.pdf