
Lucene - die Volltext Suchmaschine

JÖRG KESSENICH

Ziel des Vortrags

Eine kurze Einführung in Lucene

Vorstellen anhand eines kleinen Beispielprogramms

Tools, Suchmaschinen und anderes

Was ist „Lucene“

Eine Volltextsuche

Apache project (java)

Eine No-Sql Datenbank

Eine sehr schnelle Suchmaschine

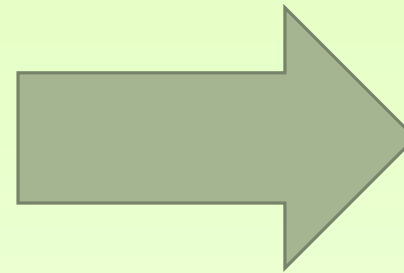
Lucene.Net ist eine Portierung des java codes (dll)

Inverted Index / Wie funktioniert Lucene

1 Mandeln in einer Pfanne goldbraun rösten

2 Gemüse unter dem heißen Grill goldbraun gratinieren

3 Unter dem vorgeheizten Grill goldbraun überbacken



mandeln	1
pfann	1
goldbraun	1,2,3
rost	1
gemuse	2
heis	2
grill	2,3
gratienier	2
vorgeheizt	3
uberback	3

Beispiel Suche von Kochrezepten

Lucene Example Cookbook

File Do Search Clear Filter About

Category

- Auflauf / Überbackenes (12)
- Backen (24)
- Beilage (15)
- Blanchieren (3)
- Braten (43)
- Braten (Fleisch) (2)
- Britisch (1)
- Brot / Brötchen / Toast (6)
- Brunch / Frühstück (6)
- Deutsch (regional) (5)
- Dünsten (12)
- Eier (83)
- Einfach (83)
- Fingerfood / Snack (5)
- Fisch (11)
- Frittieren (2)
- Geflügel (3)
- Gemüse (53)
- Getreide (21)
- Gewürze (34)
- Gratinieren / Überbacken (8)
- Grillen (2)
- Gut vorzubereiten (14)
- Hauptspeise (68)
- Hülsenfrüchte (8)
- Italienisch (2)
- Kalb (1)
- Kalorienarm / leicht (8)
- Kartoffeln (74)

Enter Search Criteria **Kartof*** **Do Search**

Einfach **Eier**

Search Results

Grundrezept Kartoffelteig

- 1 kg vorw. festkochende Kartoffeln waschen, trocknen und rundherum mit einer Gabel mehrmals einstechen. Im heißen Ofen
- Die Kartoffeln kurz abkühlen lassen, pellen und durch eine Kartoffelpresse drücken. Mit Salz, Pfeffer und Muskat kräftig würze

Kartoffelbällchen

- Milch, 200 ml Wasser und 2 Ei weiche Butter in einem Topf erhitzen. Heiße Milchmischung in eine Schüssel geben. Das Püree
- Ei und Schinken mit einem Kochlöffel unter die Kartoffelmasse rühren. Mit Salz, Pfeffer und Muskat würzen.
- Die Kartoffelmasse am besten mit einem Eisportionierer zu 8-10 Bällchen formen und auf ein mit Backpapier belegtes Backblech

Kartoffelsalat mit Hackbällchen

- Kartoffeln waschen und mit Salzwasser bedeckt zugedeckt aufkochen. Bei mittlerer Hitze 20 Min. garen.
- Zwiebel sehr fein würfeln. Gemüsebrühe erhitzen, mit Zwiebelwürfeln, Gurkensud, Senf und Salz verrühren und abschmecken.
- Hackfleisch mit Ei, Semmelbröseln, Ketchup und Salz gut vermischen. Zu walnussgroßen Bällchen formen und auf einen geölt
- 4 Ei Öl in einer beschichteten Pfanne erhitzen und die Hackbällchen darin rundherum hellbraun anbraten. Bei mittlerer Hitze

Kartoffelknödel

- Brötchen würfeln und in einer Pfanne mit der Butter bei mittlerer Hitze goldbraun rösten. Beiseite stellen.
- Kartoffeln waschen und in einem Topf mit kaltem Wasser bedecken. Zugedeckt aufkochen lassen und bei mittlerer Hitze 20-2
- Kartoffeln pellen und noch heiß durch die Kartoffelpresse drücken oder mit einem Kartoffelstampfer sehr fein zerstampfen. V
- Die Kartoffelmasse mit Salz und etwas Muskat würzen. Mehl und Stärke darübersteuern und mit den Händen locker untermisc
- Reichlich Wasser in einem großen Topf zum Kochen bringen und leicht salzen. Die Knödel in das kochende Wasser geben, zu

Number of recipes: 83

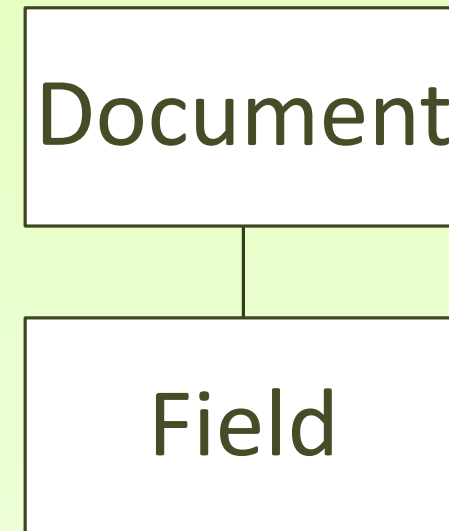
Index erstellen

IndexWriter

Analyzer (e.g. stopwords)

Document

Field



Methode CreateIndex

```
private void CreateIndex(cookml cmlData)
{
    var fsDirectory = FSDirectory.Open(indexDirectory);

    using (var writer = new IndexWriter(fsDirectory, GetAnalyzer(), true, IndexWriter.MaxFieldLength.UNLIMITED))
    {
        foreach (var recipe in cmlData.recipe)
        {
            if (recipe.head == null || recipe.head.Length == 0
                || recipe.preparation == null || recipe.preparation.Length == 0)
                continue;
            var document = new Document();

            //Indexing with stemming
            var titleField = new Field("title", recipe.head[0].title, Field.Store.YES, Field.Index.ANALYZED);
            titleField.Boost = 2;
            document.Add(titleField);
            StringBuilder prep = new StringBuilder();
            foreach (var step in recipe.preparation)
            {
                prep.AppendLine(step.text[0]);
            }
            document.Add(new Field("preparation", prep.ToString(), Field.Store.YES, Field.Index.ANALYZED));

            //Indexing without stemming
            document.Add(new Field("_title", recipe.head[0].title, Field.Store.NO, Field.Index.ANALYZED_NO_NORMS));
            document.Add(new Field("_preparation", prep.ToString(), Field.Store.NO, Field.Index.ANALYZED_NO_NORMS));

            document.Add(new Field("rid", recipe.head[0].rid, Field.Store.YES, Field.Index.NOT_ANALYZED));
            //if (recipe.head[0].picbin != null && recipe.head[0].picbin.Length > 0)
            //{
            //    document.Add(new Field("picbin", Convert.ToBase64String(recipe.head[0].picbin[0].Value), Field.Store.YES, Field.Index.ANALYZED));
            //}
            if (recipe.head[0].cat != null && recipe.head[0].cat.Length > 0)
            {
                foreach (var cat in recipe.head[0].cat)
                {
                    document.Add(new Field("category", cat, Field.Store.NO, Field.Index.NOT_ANALYZED));
                }
            }
            writer.AddDocument(document);
        }
    }
}
```

Methode GetAnalyzer

```
private static Analyzer GetAnalyzer()
{
    var file = new FileInfo("data/stopwords.txt");
    var stopwords = WordlistLoader.GetWordSet(file, ";");
    return new GermanAnalyzer(Version.LUCENE_30, stopwords);
}
```


Suche ausführen

QueryParser

IndexSearcher

ScoreDoc

SimpleFacetedSearch

Methode SearchIndex - Query

```
var parser = new MultiFieldQueryParser(
    Version.LUCENE_30,
    new[] { "title", "preparation" },
    GetAnalyzer()
);
Query customQuery = parser.Parse(term.Trim());

var query = new BooleanQuery();
query.Add(new BooleanClause(customQuery, Occur.MUST));

if (FilterCollection.Count > 0)
{
    foreach (var filter in FilterCollection)
    {
        var categoryTerm = new Term("category", filter.Name);
        var categoryFilter = new TermQuery(categoryTerm);
        query.Add(new BooleanClause(categoryFilter, Occur.MUST));
    }
}
IndexSearcher searcher = GetIndexSearcher();
TopDocs result = searcher.Search(query, numberOfResults);
```

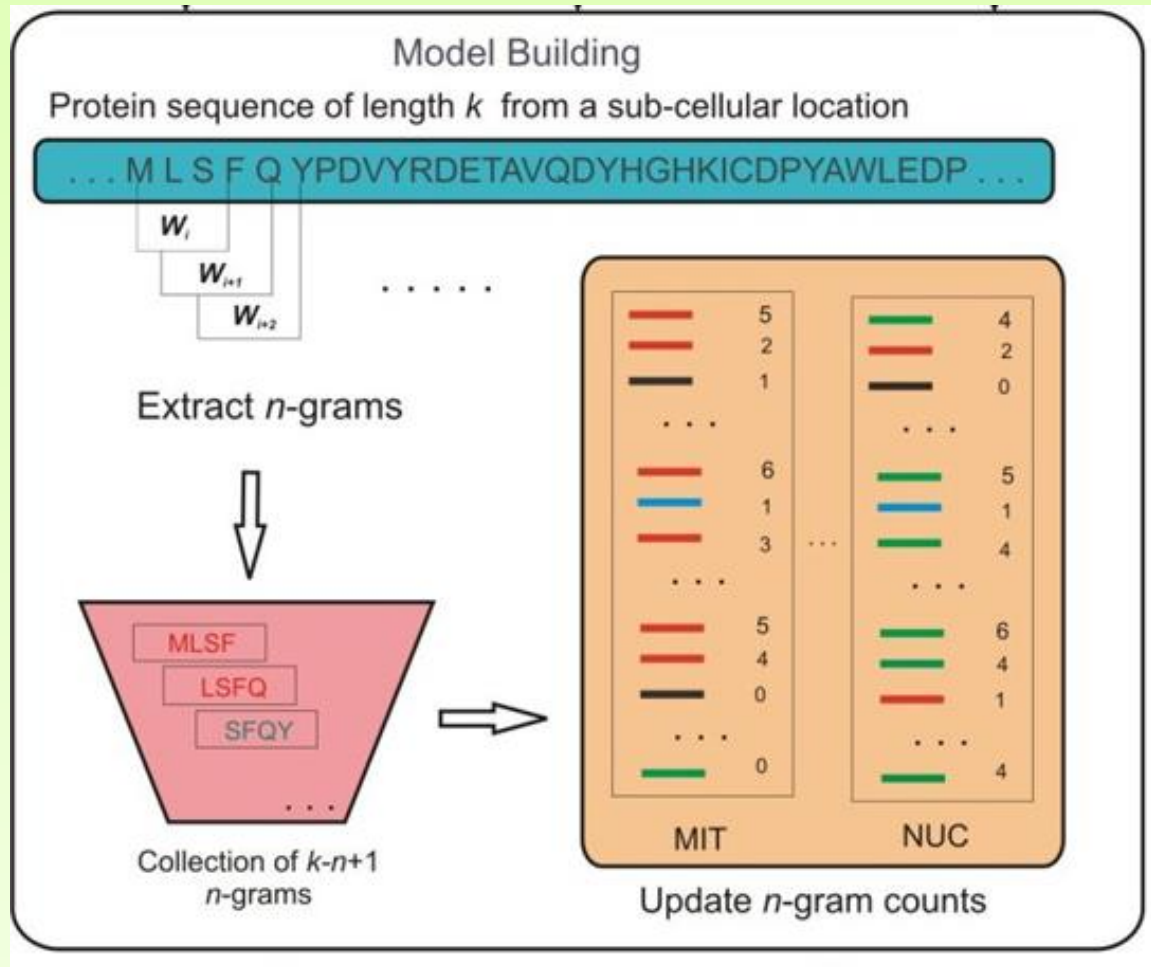
Methode SearchIndex - ScoreDoc

```
OutputFacet(searcher.IndexReader, query, "category", "Category");

ResultCollection.Clear();
NumberOfRecipes = result.TotalHits;
for (var i = 0; i < result.TotalHits; i++)
{
    ScoreDoc scoreDoc = result.ScoreDocs[i];
    var score = scoreDoc.Score;
    var documentIndex = scoreDoc.Doc;
    var document = searcher.Doc(documentIndex);

    ResultCollection.Add(new ResultItem
    {
        Title = document.Get("title"),
        Preparation = document.Get("preparation")
    });
}
```

Komplexe Objekte indexieren



FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately

Inbal Budowski-Tala, Yuval Novb, and Rachel Kolodnya

doi: 10.1073/pnas.0914097107

<http://www.pnas.org/content/107/8/3481.abstract>

Luke

Luke - Lucene Index Toolbox, v 3.5.0 (2011-12-28)

File Tools Settings Help

Overview Documents Search Files Plugins

Enter search expression here:
category:Gemüse

Analysis QueryParser Similarity Collector

Analyzer to use for query parsing:
NOTE: use fully-qualified class name here. Default field:
org.apache.lucene.analysis.KeywordAnalyzer _preparation

Optional constructor argument:

Query details: Update Explain structure
category:Gemüse Parsed Rewritten

Last search time: 12347 us Search repeat 1 times.

Results: (Hint: Double-click on results to display all fields) Explain Delete 4359 doc(s) 0-19

#	Score	Doc. Id	_preparation	title	category	preparation	rid	title
0	0,5734	1				1. Grundreze	229616	Stampf mit Rosenkohl
1	0,4778	2				1. Grundreze	229620	Stampf mit Blauschimmelkäse
2	0,4778	3				1. 500 g Knol	229644	Cremeriger Sellerie
3	0,5734	4				1. 500 g Rote	229652	Rote Bete süß-sauer
4	0,3823	5				1. Zwiebel fei	229679	Möhrenlasagne
5	0,5734	13				1. Grundreze	229749	Raclette-Stampf
6	0,4778	14				1. Ingwer sch	231712	Asia-Suppe
7	0,3823	15				1. Zwiebeln h	231720	Rote-Bete-Kartoffel-Eintopf
8	0,4778	18				1. Zwiebeln ir	231736	Saure Zipfel
9	0,4778	19				1. 2 Schalotte	231748	Lauwarmer Gurkensalat
10	0,4778	20				1. 150 g porti	231752	Spinat-Kartoffelstampf
11	0,4778	22				1. 400 g Wirs	231776	Wirsingrahmsuppe
12	0,3823	23				1. 1 kleinen F	231800	Salat mit Ziegenkäse
13	0,4778	24				1. 3 Orangen	231836	Tomaten-Orangen-Salat
14	0,3823	30				1. 300 ml Wa	231879	Kalbs-Sattimbocca
15	0,5734	31				1. Möhren, Se	231883	Linsengemüse
16	0,3823	33				1. Hühnerke	231891	Reistopf
17	0,4778	34				1. Chicorée u	231895	Chicorée-Salat mit Clementinen
18	0,4778	35				1. Weißkohl	2818	Kohl-Letscho
19	0,6689	36				1. Garnelen e	2834	Peperoncini-Garnelen

Index name: C:\NewInfo...LuceneCook\bin\Debug\Index

Suchmaschinen

Elasticsearch

Apache Solr

Open Semantic Search

SEBOL

...

Wer/Wo wird Lucene verwendet

 COMPAREGROUP.EU

 SOUND CLOUD

Instagram

NETFLIX

in



salesforce



WIKIMEDIA

ebay

myspace

XING

AUTO
SCOUT 24

ticketmaster®

HolidayCheck


Audi

BEST
BUY

 WÜRTHPHOENIX

OTTO

theguardian



Referenzen/Links

<http://lucene.apache.org/pylucene/index.html>

<http://lucene.apache.org/solr/>

<https://www.elastic.co/products/elasticsearch>

<https://www.opensearch.org/de/>

<http://db-engines.com/de/ranking/suchmaschine>

<http://www.dotnetpro.de/update/dotnetpro/wider-fluchen-suchen-1126651.html>

Lucene in Action von Erik Hatcher, Otis Gospodnetic, Mike McCandless

Instant Lucene.NET von Michael Heydt

Lucene 4 Cookbook von Edwood Ng, Vineeth Mohan

Fragen ?

Jörg Kessenich
joerg.kessenich@gmail.com