

Apache Lucene and its Ecosystem

Informationsorganisation
& Information Retrieval

Michael Föger, Gunharth Randolph, Helene Wechselberger

Contents

1. First tests with Lucene (simple text crawling, text file crawlings)
2. Lucene Analyzer
3. Indexing the Web
 - 3.1. WebCrawler
 - 3.2. Solr/Nuxt
4. Web Archive
 - 4.1. SolrWayBack
 - 4.2. jwarc
 - 4.3. Indexing Reuters / date range search with luke

Hello Lucene

Let's see what you can do for us!
Getting started with 3 examples.

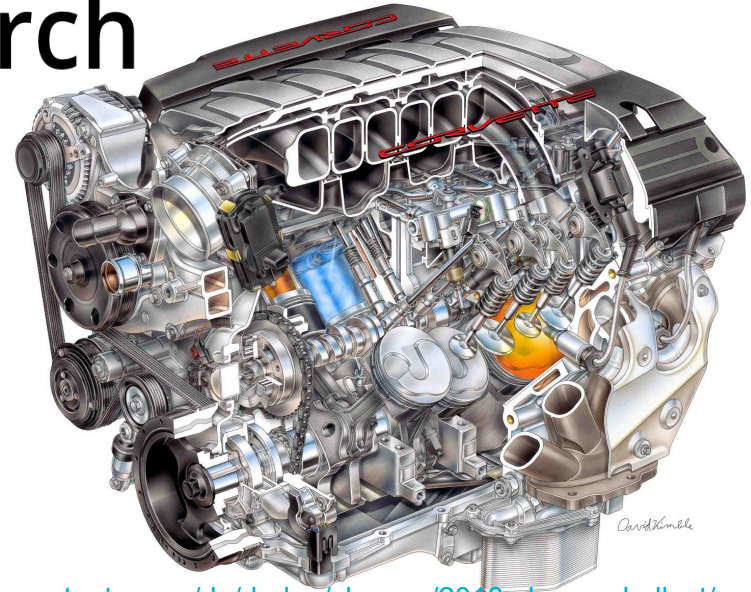
What is Lucene?

Analysis - Index - Search Queries

+ CrateDB  elasticsearch

 DocFetcher Apache
Solr 

 **swifttype**



<https://www.agtauto.com/de/dodge/charger/2018-charger-hellcat/>

getting started exercises

1 - search in code and save to separate folder

```
public static void main(String[] args)
{
    Example_01 hl = new Example_01( path: "example_01_output");
    try {
        hl.index();
        hl.searchAndDisplay( searchText: "information");
        hl.searchAndDisplay( searchText: "lecture");
        hl.searchAndDisplay( searchText: "example");
    } catch (Exception e) {
        e.printStackTrace();
    }
}
```

getting started exercises

2 - search in code and save to temporary buffer memory

```
//Now let's try to search for Hello
```

```
IndexReader reader = DirectoryReader.open(directory);  
IndexSearcher searcher = new IndexSearcher (reader);  
QueryParser parser = new QueryParser ( f: "content", standardAnalyzer);
```

```
Query query = parser.parse( query: "Hello");  
TopDocs results = searcher.search(query, n: 5);  
System.out.println("Hits for Hello -->" + results.totalHits);
```

```
//case insensitive search
```

```
query = parser.parse( query: "hello");  
results = searcher.search(query, n: 5);  
System.out.println("Hits for hello -->" + results.totalHits);
```

getting started exercises

3 - search in separate *.txt file and save to temporary buffer memory

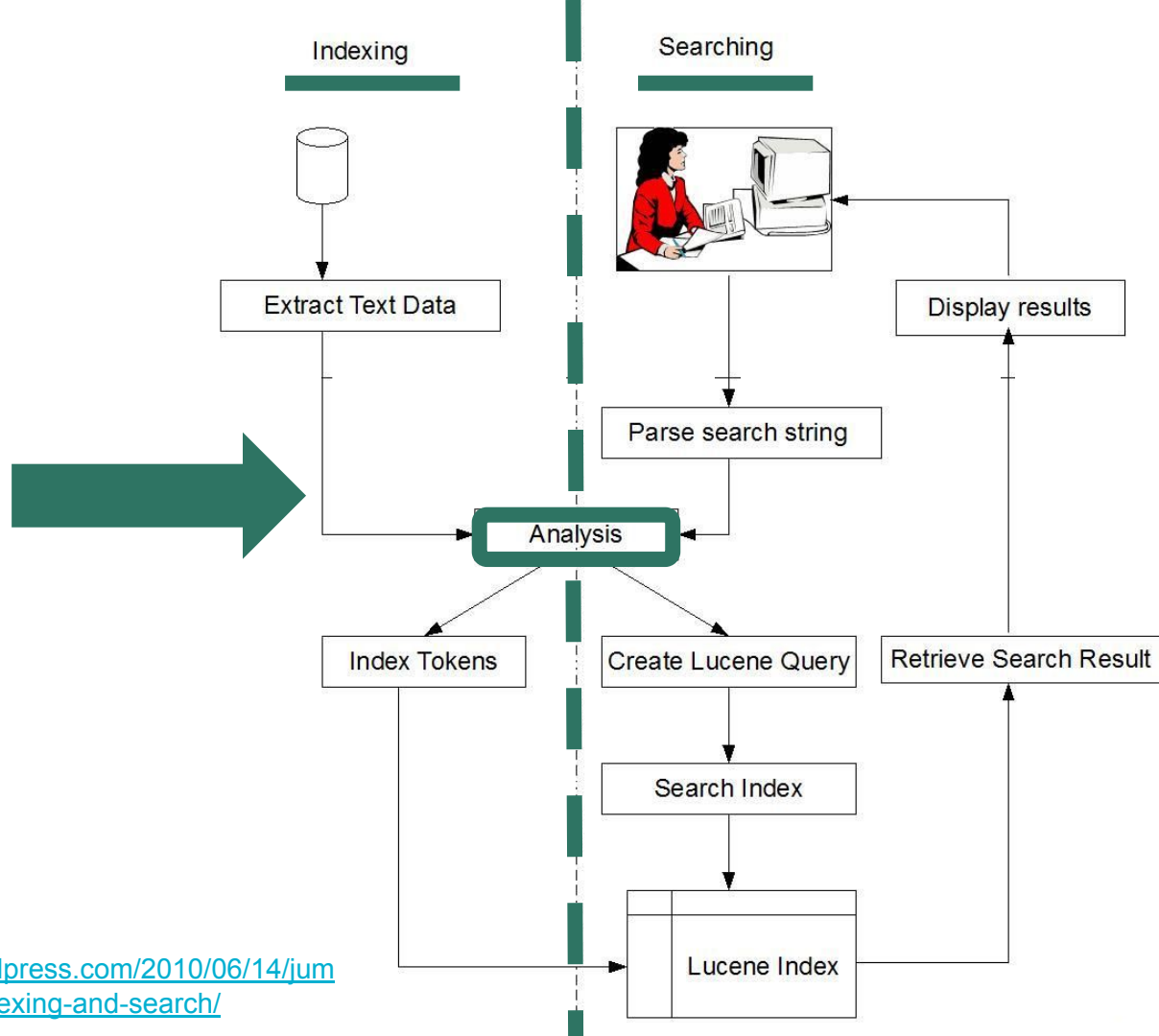
```
public class Example_03
{
    public static void main(String[] args) throws IOException, ParseException {
        // New index
        StandardAnalyzer standardAnalyzer = new StandardAnalyzer();
        String inputFilePath = "src/main/java/com/gunicode/lucene/example_03_input.txt";
        String outputDir = "example_03_output";
        FileReader file = new FileReader(inputFilePath);

        Directory directory = FSDirectory.open(Paths.get(outputDir));
        IndexWriterConfig config = new IndexWriterConfig(standardAnalyzer);
        config.setOpenMode(OpenMode.CREATE);
        // Create a writer
        IndexWriter writer = new IndexWriter(directory, config);

        Document document = new Document();

        try (BufferedReader br = new BufferedReader(file)) {
```

Analysing



Lucene API

Lucene Analyzer

→ used to analyze text while

- indexing and
- searching

documents.

Analyzers mainly consist of **tokenizers** and **filters**.

Required:

Maven dependencies
to pom.xml file

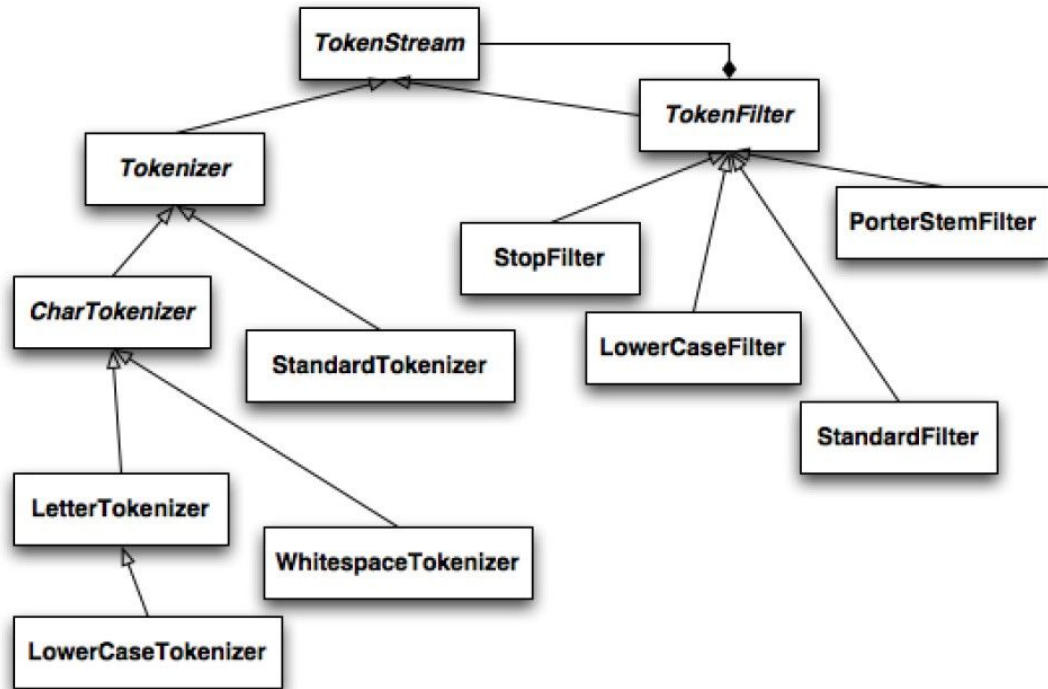
```
1 <dependency>
2   <groupId>org.apache.lucene</groupId>
3   <artifactId>lucene-core</artifactId>
4   <version>7.4.0</version>
5 </dependency>
6 <dependency>
7   <groupId>org.apache.lucene</groupId>
8   <artifactId>lucene-queryparser</artifactId>
9   <version>7.4.0</version>
10 </dependency>
11 <dependency>
12   <groupId>org.apache.lucene</groupId>
13   <artifactId>lucene-analyzers-common</artifactId>
14   <version>7.4.0</version>
15 </dependency>
```

Lucene Analyzer – What it does:

- splits text into **tokens**
- various analyzers available

Different analyzers consist of different combinations of tokenizers and filters.

→ **different function of analyser**



StandardAnalyzer (demo)

- most commonly used analyzer
- recognizes URLs and emails
- removes english stop words
- lowercases the generated tokens

EXAMPLE

Input:

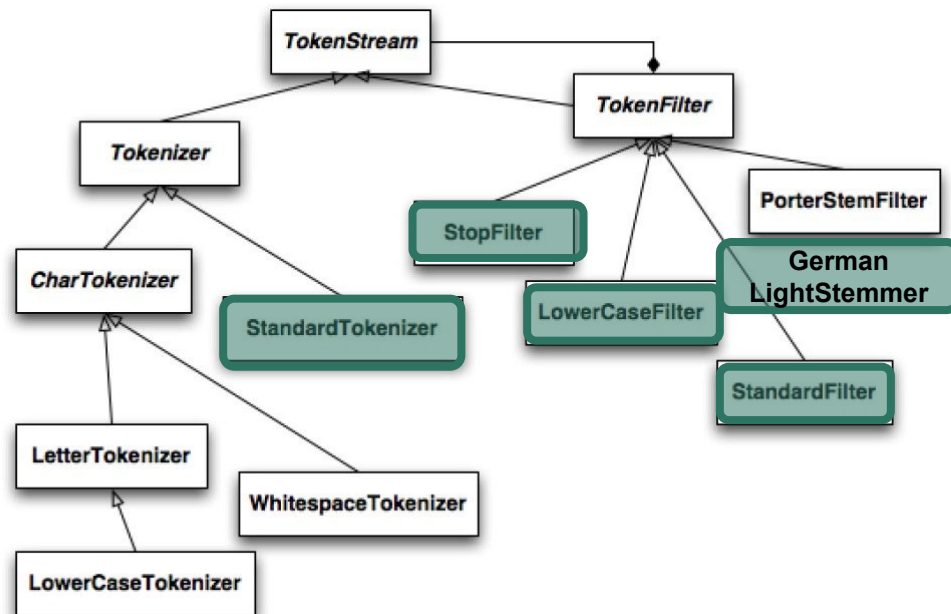
Dieses Wappen auf landesmuseen.at enthält Ritter, Löwe and some english dashed-text

Output:

[dieses, wappen, auf, landesmuseen.at, enthält, ritter, löwe, some, english, dashed, text]

GermanAnalyzer (demo)

- StandardTokenizer
- StandardFilter
- LowercaseFilter
- StopFilter
- GermanLightStemmer



Input:

Dieses Wappen auf landesmuseen.at enthält Ritter, Löwe and some english dashed-text

Output:

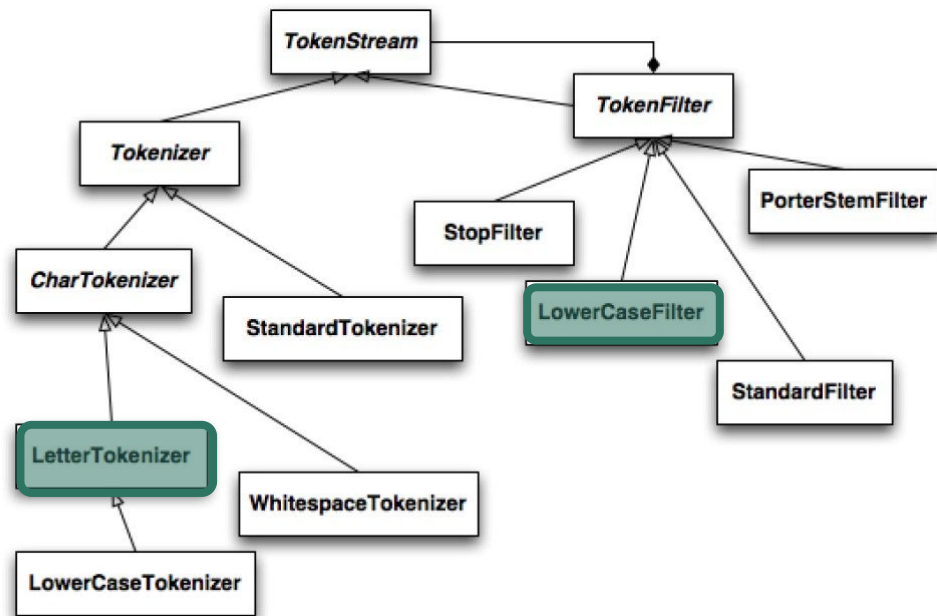
[wapp, landesmuseen.at, enthalt, ritt, low, and, som, english, dashed, text]

SimpleAnalyzer

- LetterTokenizer
- LowercaseFilter

Bear in mind...

- doesn't remove stop words
- doesn't recognize URLs



Input:

Dieses Wappen auf landesmuseen.at enthält Ritter, Löwe and some english dashed-text

Output:

[dieses, wappen, auf, landesmuseen, at, enthält, ritter, löwe, and, some, english, dashed, text]

WhitespaceAnalyzer

Bear in mind...

- splits text by whitespace characters

Input:

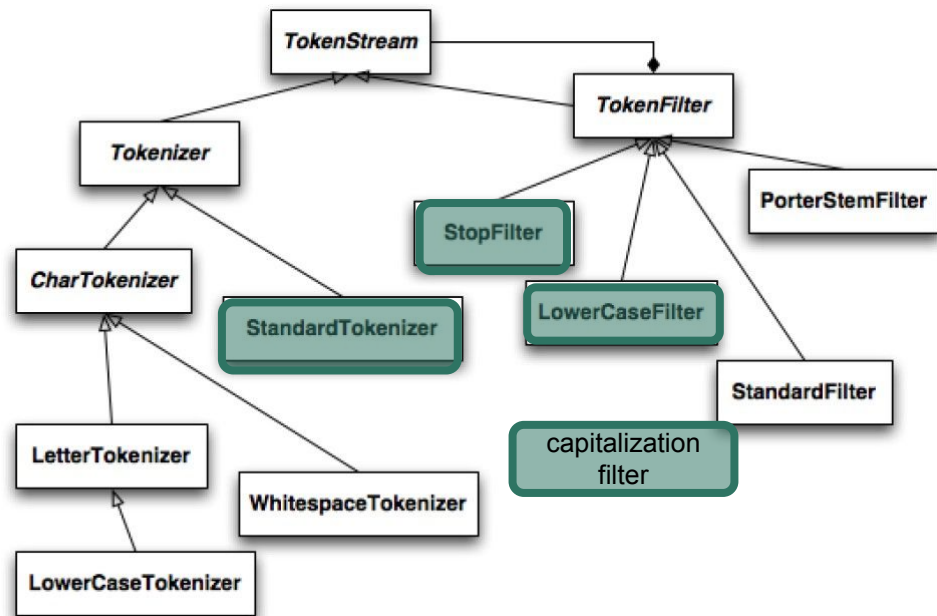
Dieses Wappen auf landesmuseen.at enthält Ritter, Löwe and some english dashed-text

Output:

[Dieses, Wappen, auf, landesmuseen.at, enthält, Ritter,, Löwe, and, some, english, dashed-text]

CustomAnalyzer

- standardTokenizer
- LowerCaseFilter
- StopFilter
- CapitalizationFilter



Input:

Dieses Wappen auf landesmuseen.at enthält Ritter, Löwe and some english dashed-text

Output:

[Dieses, Wappen, Auf, Landesmuseen.at, Enthält, Ritter, Löwe, Some, English, Dashed, Text]

PerFieldAnalyzerWrapper

...allows for assigning **different analyzers to different fields**

e.g. title → StandardAnalyzer
body → CustomAnalyzer

Steps:

- 1 - Create HashMap (AnalyzerMap)
- 2 - Map "title" to StandardAnalyzer, map "body" to customAnalyzerMap
- 3 - Create PerFieldAnalyzerWrapper
 - provide "our" AnalyzerMap
 - provide StandardAnalyzer

Indexing the Web

Objectives

- Build a Web-Crawler
- Use the Java programming language
- Optimise the resulting Lucene Index
- Research different areas of Lucene



TIROLER WAPPEN

Die Fischtaler Wappenkartei

Suche

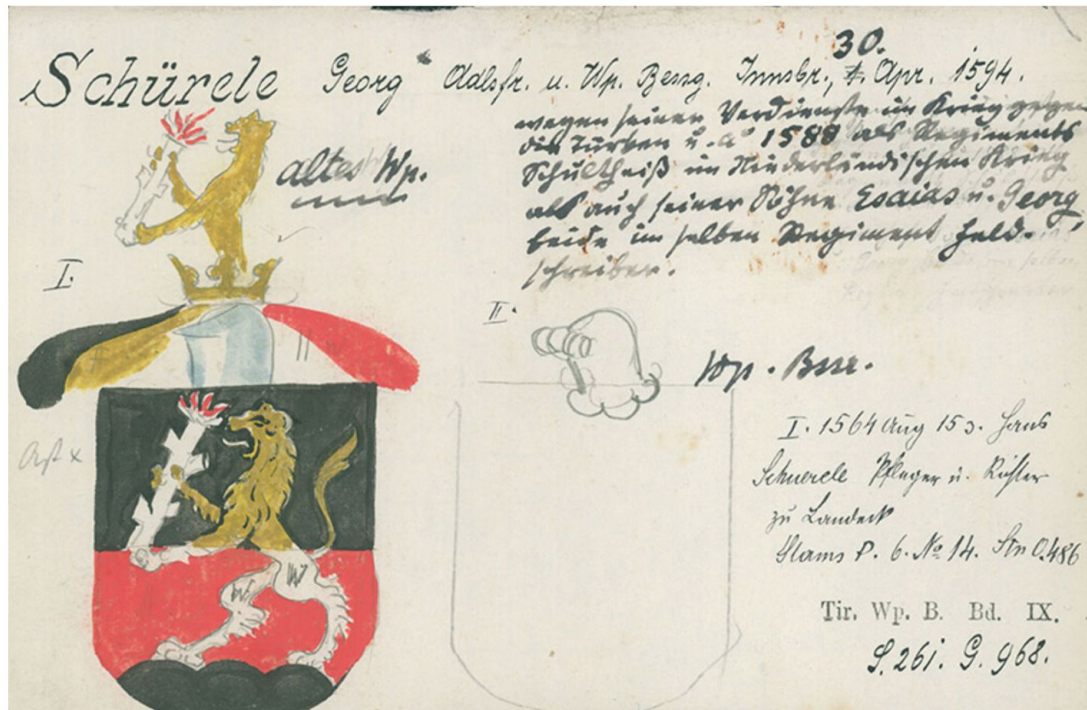
Listen

Entstehungsgeschichte

Hilfe

Suche nach: 'sarley'

[Vorige Karte \(2\)](#) | [\(3 von 9\)](#) | [Nächste Karte \(4\)](#)



Site to index

Seed URL: <http://wappen.tiroler-landesmuseen.at>

- 30.000 Objects
- the demos use the first tab "A-Ban": 1.800 objects
- <http://wappen.tiroler-landesmuseen.at:81/namen.php>
- crawl depth: 1

Repo: <https://gitlab.web.fh-kufstein.ac.at/gunharth/lucene-web-crawler>

- code review

TIROLER WAPPEN Die Fischtaler Wappenkartei

Suche

Listen

Suche nach: 'sarley'

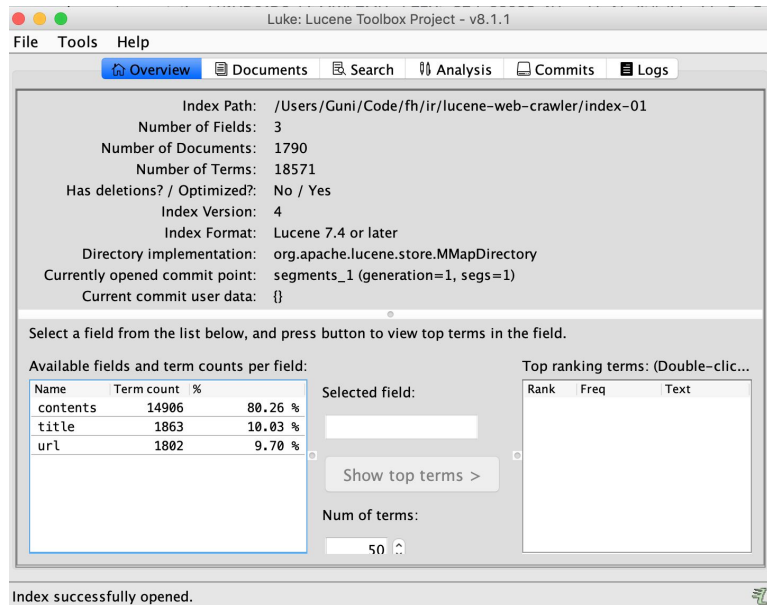
[Vorige Karte \[2\]](#) | [\[3 von 9\]](#) | [Nächste Karte \[4\]](#)



Lucene – Viewing the Index

Lucene Luke

- Lucene Toolbox Project
- GUI tool for introspecting Lucene / Solr / Elasticsearch index
- as of version 8.1 part of Lucene
- in the lucene installation folder run `./luke/luke.sh`



Lucene Web Crawler – Optimising the Index

index-01, StandardAnalyzer

- Index includes words that are on every page, e.g. header
- StandardAnalyzer uses the STOP_WORDS_SET (common english words)

48	1800	ein
49	1800	die
50	1800	des
51	1800	der

- No stemming

487	1785	ritterordensarchivmandatenbuch
488	1785	rittergeschlechter
489	1785	rittergeschichte
490	1785	ritter

Rank	Freq	Text
1	1790	zusammenarbeit
2	1790	wien
3	1790	wie
4	1790	wappenträger
5	1790	wappenkartei
6	1790	wappen
7	1790	w
8	1790	virgen
9	1790	tyrol
10	1790	tiroler
11	1790	tirol
12	1790	taufers
13	1790	suche
14	1790	stams
15	1790	st
16	1790	sie
17	1790	schwatz
18	1790	schubladen
19	1790	sche
20	1790	s

Lucene Web Crawler – Optimising the Index

index-01 search

- contents:ritter contents:löwe

Doc ID	Score	Field Values
7	4.12	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
1306	3.677	title=Tiroler Wappen: Augsburg; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1372&drawe...
1735	3.677	title=Tiroler Wappen: Panthaleb von Florenz Pernhart; url=http://wappen.tiroler-landesmuseen.at:81/index34a.ph...
214	3	title=Tiroler Wappen: Aicher Hans; Aicher Georg Christ.; Aicher von Aichenegg Hans; Aicher von Aichenegg Georg...
6	2.999	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
56	2.999	title=Tiroler Wappen: Abbondi; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=68&drawer=a...
660	2.999	title=Tiroler Wappen: Am Weg Hans; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=709&dra...
1005	2.999	title=Tiroler Wappen: Ekk; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1065&drawer=a-ban;
1108	2.999	title=Tiroler Wappen: Arnold; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1168&drawer=...
1119	2.999	title=Tiroler Wappen: Arquin Hans; Arquin Christoff; Arquin Simon; Arquin Valentin; Arquin Caspar; url=http://...

Lucene Web Crawler – Optimising the Index

index-02, StandardAnalyzer

- use German stopwords
- add custom German stop words (zusammenarbeit, wappenkartei)

48	1782	ban
49	1782	altneuland
50	1782	a
51	1782	1780

- no stemming

449	1776	ritterordensarchivmandatenbuch
450	1776	rittergeschlechter
451	1776	rittergeschichte
452	1776	ritter

Rank	Freq	Text
1	1782	wien
2	1782	wappenträger
3	1782	wappen
4	1782	w
5	1782	virgen
6	1782	tyrol
7	1782	tiroler
8	1782	tirol
9	1782	taufers
10	1782	suche
11	1782	stams
12	1782	st
13	1782	schwatz
14	1782	schubladen
15	1782	sche
16	1782	s
17	1782	regensburg
18	1782	rattenberg
19	1782	quellen
20	1782	projekt

Lucene Web Crawler – Optimising the Index

index-02 search

- contents:ritter contents:löwe

Doc ID	Score	Field Values
8	4.12	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
1300	3.688	title=Tiroler Wappen: Augsburg; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1372&drawe...
1726	3.688	title=Tiroler Wappen: Panthaleb von Florenz Pernhart; url=http://wappen.tiroler-landesmuseen.at:81/index34a.ph...
216	3.015	title=Tiroler Wappen: Aicher Hans; Aicher Georg Christ.; Aicher von Aichenegg Hans; Aicher von Aichenegg Georg...
7	3.015	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
57	3.015	title=Tiroler Wappen: Abbondi; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=68&drawer=a...
657	3.015	title=Tiroler Wappen: Am Weg Hans; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=709&dra...
1000	3.015	title=Tiroler Wappen: Ekk; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1065&drawer=a-ban;
1102	3.015	title=Tiroler Wappen: Arnold; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1168&drawer=...
1185	3.015	title=Tiroler Wappen: Ascher Franz Ant.; Ascher Franz Anton; url=http://wappen.tiroler-landesmuseen.at:81/inde...

Lucene Web Crawler – Optimising the Index

index-03, StandardAnalyzer

- use PorterStemmer
(Engl. only!)
- optimise jsoup document parsing

```
doc.select("form").remove();
doc.select("div#header").remove();
doc.select("div#navigation").remove();
doc.select("div.tab").remove();
doc.select("div#simple").remove();
doc.select("div#ext").remove();
doc.select("div#info").remove();
doc.select("div.dialogtext4").remove();
doc.select("div#footer").remove();
doc.select("canvas").remove();
doc.select("script").remove();
```

Rank	Freq	Text
1	1794	wappenträger
2	1794	namen
3	1794	list
4	1794	15
5	1786	zweck
6	1786	your
7	1786	webseit
8	1786	urheberrecht
9	1786	un
10	1786	transkript
11	1786	the
12	1786	tag
13	1786	support
14	1786	such
15	1786	sollten
16	1786	senden
17	1786	rückmeldung
18	1786	quellen
19	1786	publikationen
20	1786	projekt

Lucene Web Crawler – Optimising the Index

index-03 search

- contents:ritter contents:löwe

Doc ID	Score	Field Values
213	4.178	title=Tiroler Wappen: Aicher Hans; Aicher Georg Christ.; Aicher von Aichenegg Hans; Aicher von Aichenegg Georg...
10	3.595	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
1307	3.221	title=Tiroler Wappen: Augsburg; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1372&drawe...
1739	3.124	title=Tiroler Wappen: Panthaleb von Florenz Pernhart; url=http://wappen.tiroler-landesmuseen.at:81/index34a.ph...
1413	2.656	title=Tiroler Wappen: Pach z. Hansenheim Christoph Ulr. v.; Pach z. Hansenheim u. Hohen Eppan Christoph Ulr. v...
185	2.634	title=Tiroler Wappen: Aibling von; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=205&dra...
1285	2.634	title=Tiroler Wappen: Aufenstein v.; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1349&...
1369	2.634	title=Tiroler Wappen: Papa v.; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1435&drawer...
1780	2.634	title=Tiroler Wappen: Panz v.; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1860&drawer...
9	2.599	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...

Lucene Web Crawler – Optimising the Index

index-04, GermanAnalyzer

- uses German stop words
- additional custom stop words
- uses GermanLightStemmer

```
CharCharacterSet stopSet = CharCharacterSet.copy(GermanAnalyzer.getDefaultStopSet());  
stopSet.add("wappenträger");  
stopSet.add("wappenkartei");  
Analyzer analyzer = new GermanAnalyzer(stopSet);
```

Rank	Freq	Text
1	1794	nam
2	1794	list
3	1793	15
4	1788	frag
5	1787	zweck
6	1787	webseit
7	1787	urheberrecht
8	1787	transkript
9	1787	such
10	1787	sollt
11	1787	send
12	1787	ruckmeldung
13	1787	quell
14	1787	publikation
15	1787	projekt
16	1787	preis
17	1787	nachricht
18	1787	mwst
19	1787	kostenpflichtig
20	1787	karteikart

Lucene Web Crawler – Optimising the Index

index-04 search

- contents:ritter contents:löwe

Important:
set Analyzer to
GermanAnalyzer
in Luke!

Doc ID	Score	Field Values
215	3.787	title=Tiroler Wappen: Aicher Hans; Aicher Georg Christ.; Aicher von Aichenegg Hans; Aicher von Aichenegg Georg...
1689	3.294	title=Tiroler Wappen: Baldung von Löwen; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1...
1690	3.294	title=Tiroler Wappen: Baldung von Löwen; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1...
9	3.143	title=Tiroler Wappen: Attlmayr Ferdinand; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=...
25	3.038	title=Tiroler Wappen: Bayern; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=33&drawer=a-...
1308	2.842	title=Tiroler Wappen: Augsburg; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1372&drawe...
1739	2.742	title=Tiroler Wappen: Panthaleb von Florenz Pernhart; url=http://wappen.tiroler-landesmuseen.at:81/index34a.ph...
1371	2.526	title=Tiroler Wappen: Papa v.; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1435&drawer...
1780	2.526	title=Tiroler Wappen: Panz v.; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=1860&drawer...
187	2.501	title=Tiroler Wappen: Aibling von; url=http://wappen.tiroler-landesmuseen.at:81/index34a.php?wappen_id=205&dra...

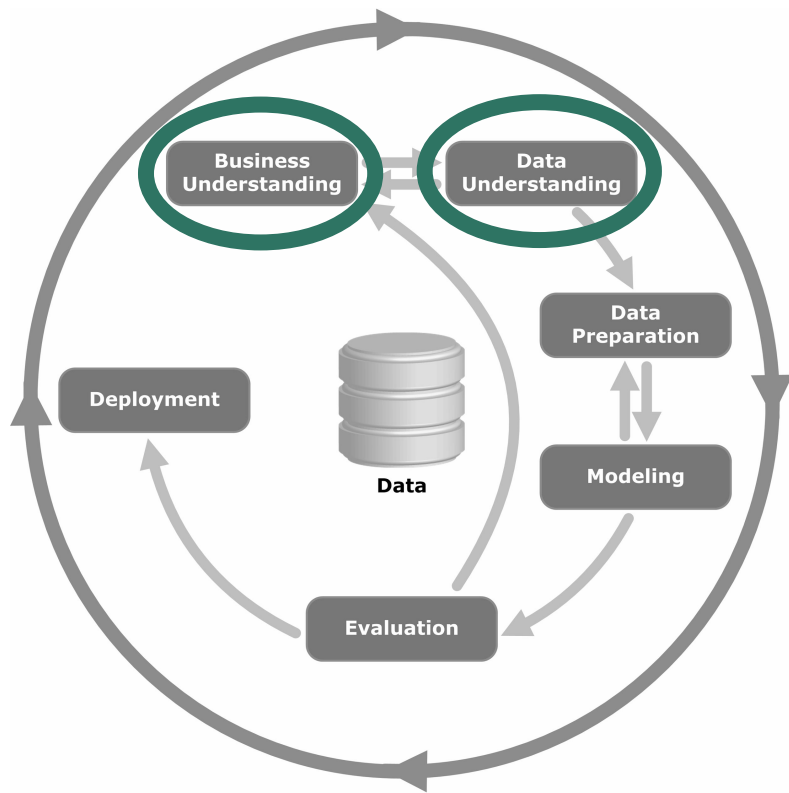
Lucene Web Crawler – Suche

“Test test test... (did we say test?)

Beware of too much analysis – it might hurt indexing performance.

Start with the same analyzer for indexing and search, otherwise searches would not find what they are supposed to...”

https://lucene.apache.org/core/8_1_1/core/org/apache/lucene/analysis/package-summary.html#package.description



Indexing the Web

with Apache Solr and Apache Nutch

Apache Solr: open source enterprise search platform built on Apache Lucene

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

Apache Nutch: Highly extensible, highly scalable Web crawler

<https://nutch.apache.org/>

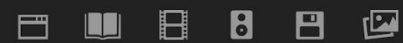
Repo:


<https://github.com/gunharth/solr-nutch-docker>



Web Archives

-



 SIGN IN

 UPLOAD

 Search

ABOUT

CONTACT

BLOG

PROJECTS

HELP

DONATE

JOBS

VOLUNTEER

PEOPLE

DONATE

INTERNET ARCHIVE
WayBackMachine

Explore more than 371 billion [web pages](#) saved over time

Enter a URL or words related to a site's home page



INTERNET ARCHIVE
WayBackMachine

<https://archive.org/web/>



Feedback

Web Archives – SolrWayBack

SolrWayBack

A search interface and wayback machine

<https://github.com/netarchivesuite/solrwayback>

Java8 required!

- Image search
- Search by uploading a file. See if the resource has been harvested and from where.
- Link graph showing links (ingoing/outgoing) for domains using the D3 javascript framework.
- Raw download of any harvested resource from the binary Arc/Warc file.
- Export a search resultset to a Warc-file. Streaming download, no limit of size of resultset.
- Build in SOCKS proxy to view historical webpages without browser leaking resources from the live web.

Solr Admin: <http://localhost:8983/solr/#/netarchivebuilder> (Demo)

SolrWayBack: <http://localhost:8080/solrwayback/> (Demo)

Web Archives – WARC

warc file from command line:

Create a warc file of an entire website:

```
wget --mirror --warc-cdx --page-requisites --warc-file=wappen -i
```

<http://wappen.tiroler-landesmuseen.at:81/namen.php>

--mirror:	shortcut for -N -r -l inf --no-remove-listing
-N:	don't re-retrieve files unless newer than local
-r:	recursive
-l:	max recursive depth
--no-remove-listing	don't remove '.listing' files
--warc-cdx	write CDX index files
--page-requisites:	get all images, etc. needed to display HTML page
--warc-file=FILENAME	save request/response data to a .warc.gz file
-i, --input-file=FILE	download URLs found in local or external FILE

Web Archives: Reuters-21578

<http://www.daviddlewis.com/resources/testcollections/reuters21578/>

21578 news articles in 1987

.sgm Dateien: Standard Generalized Markup Language

```
<DATE>26-FEB-1987 15:01:01.79</DATE>
```

```
<TITLE>BAHIA COCOA REVIEW</TITLE>
```

```
<BODY>...</BODY>
```

Indizierung mit <https://github.com/tdebatty/java-datasets>

Java8!

Web Archives

Reuters-21578 and the Lucene range query

Date range query in Lucene:

Convert date

26-FEB-1987 15:01:01.79

to

19870226

and add to Index

Lucene search:

query: 1987*

query: 198702*

query: [19870301 TO 19870430]

parsed: date:1987*

parsed: date:198702*

parsed: date:19870215 date:to date:19870315

result: all docs

result: Feb only

result: Mar - Apr (incl)

THE END

Questions?

Find sources on the upcoming slides! ;)

Sources

1. Apache Lucene: <https://lucene.apache.org/>
2. Apache Solr: <https://lucene.apache.org/solr/>
3. Apache Nutch: <https://nutch.apache.org/>
4. Apache Maven: <https://maven.apache.org/>
5. Jsoup: <https://jsoup.org/>
6. <https://www.baeldung.com/lucene-analyzers>
7. <http://www.evelix.ch/unternehmen/Blog/evelix/2013/11/11/inner-workings-of-the-german-analyzer-in-lucene>
8. <https://github.com/netarchivesuite/solrwayback>
9. <https://github.com/iipc/jwarc>
10. Elasticsearch: <https://www.elastic.co>
11. <http://www.lucenetutorial.com>
12. Lucene Query Syntax: <http://www.lucenetutorial.com/lucene-query-syntax.html>
13. <https://www.ionos.de/digitalguide/server/konfiguration/apache-lucene>
14. <https://www.javacodegeeks.com/2015/09/lucene-analysis-process-guide.html>
15. <https://github.com/manishkanadje/reuters-21578>