

APACHE  
LUCENE  
EUROCON



# Multilingual Search and Text Analytics with Solr

Steve Kearns, Basis Technology  
skearns@basistech.com

October 19, 2011

Presented by

**lucid**  
IMAGINATION



# Agenda

- Agenda
- About me
- Language is important
- Approaches for language-aware search
- Configuring Solr

# About Me

- Steve Kearns
- Product Manager at Basis Technology
- MS Information Technology
  
- BBN Technologies
  - **Speech to text, Machine translation**
  - **Distributed architecture**
- EMD Serono (Merck KGaA)
  - **Scientific application development**
  - **Molecular modeling Linux cluster commodity hw**

**Search language** Choose to limit your search based on language.

Search pages written in any language

Limit my searches to pages written in the following languages:

Albanian

German

Polish

Arabic

Greek

Portuguese (Brazil)

Bulgarian

Hebrew

Portuguese (Portugal)

Catalan

Hungarian

Romanian

Chinese Simplified

Icelandic

Russian

Chinese Traditional

Indonesian

Serbian (Cyrillic)

Croatian

Italian

Slovak

Czech

Japanese

Slovenian

Danish

Korean

Spanish

Dutch

Latvian

Swedish

English

Lithuanian

Thai

Estonian

Malay

Turkish

Finnish

Norwegian

Ukrainian

French

Persian

Vietnamese

# Language is Important

- Content is produced and consumed in the language of the user
- Document collections often contain more than one language
- Each language is unique, and presents different challenges to the search engine

# Language is Complex

- Tokenization
  - Some languages do not use spaces
  - Compound words combine two or more words
- Inflection
  - In grammar, inflection or inflexion is the modification of a word to express different grammatical categories such as tense, grammatical mood, grammatical voice, aspect, person, number, gender and case.

# Language is Complex



# Language is Complex

- The Spanish word “pasaportar” has more than 50 inflected forms:

pasaportando  
pasaportes  
pasaportada  
pasaportaba  
pasaportarían  
pasaportarais  
pasaportasen  
pasaportaren  
pasaportado  
pasaportaremos  
pasaportábamos  
pasaportases  
pasaportaríais  
pasaportaran  
pasaportarías  
pasaportaras  
pasaportarás

pasaportareis  
pasaportaron  
pasaportase  
pasaportemos  
pasaportaría  
pasaportara  
pasaportasteis  
pasaportáramos  
pasaportaban  
pasaportásemos  
pasaportamos  
pasaporten  
pasaportaréis  
pasaportabas  
pasaportaríamos  
pasaportáremos  
pasaporto

pasaportarán  
pasaporte  
pasaportan  
pasaporta  
pasaportaste  
pasaportad  
pasaportéis  
pasaportadas  
pasaporté  
pasaportados  
pasaportaré  
pasaportare  
pasaportará  
pasaportó  
pasaportabais  
pasaportaseis  
...

# Language is Complex

- English:
  - Spoke: Noun, part of a wheel
  - Spoke: Verb, Past tense of “speak”
- French:
  - Été: Noun, Summer
  - Été: Verb, Past tense of être (to be)
- German:
  - Samstagmorgen: Compound Noun
    - *Compound of Samstag (Saturday) and Morgen (Morning)*
- Japanese:
  - 首脳会談後、オバマ大統領は記者団の質問に答える予定
    - *Where are the words??*

# Language-Aware Search

- Language Technology
  - Language Identification
  - Tokenization
    - *N-Gram*
    - *Morphological*
  - Token processing
    - *Stemming*
    - *Lemmatization*
  - Higher level analytics
    - *Entity Extraction*
    - *Relationship Extraction*

# Language Identification

- Find a single dominant language in a document
- Find multiple languages in a single document

The New York Philharmonic Orchestra will make a historic trip to North Korea on February, it has announced. Dominique de Villepin a été nommé Premier ministre ce mardi en fin de matinée par Jacques Chirac. The orchestra's president and executive director, Zarin Mehta said it would play in the capital Pyongyang on February 26. In August, the reclusive communist country's Ministry of Culture sent an invitation to the orchestra at Lincoln Center in Manhattan. 朝鲜外务省发言人11月1日在平壤宣布, 朝鲜将重返六方会谈, 但前提条件是朝鲜与美国在六方会谈框架内讨论解除美国对朝鲜金融制裁问题。针对朝鲜方面的动向, 各方均表示欢迎。美联社11月1日报道说: “长期以来一直拒绝与平壤进行直接对话的美国总统布什认为, 各方达成一致、同意恢复六方会谈应归功于中国的斡旋。水泳の世界選手権第11日は27日、豪州・メルボルンで行われ、女子百メートル背泳ぎ決勝で中村礼子(東京SC)が1分0秒40の日本人として大会初の銅メダルを獲得した。優勝は59秒44の世界新をマークしたアメリカ人のナタリー・コーグリッソ。 L'ancien ministre de l'Intérieur, qui n'a jamais participé à une élection, a déjeuné avec les députés UMP et UDF à l'invitation du président de l'Assemblée nationale, Jean-Louis Debré.

Chinese, Simplified	148
English	367
French	283
Japanese	119
<b>Total</b>	<b>917</b>

# Tokenization

- Morphological Analysis vs. N-gram
- Search Term: 東京 ルパン上映時間
- N-gram:

org.apache.lucene.analysis.cjk.CJKAnalyzer {}

term position	1	2	3	4	5	6	7	8
term text	東京	京ル	ルバ	パン	ン上	上映	映時	時間
term type	double	double	double	double	double	double	double	double
source start,end	0,2	1,3	2,4	3,5	4,6	5,7	6,8	7,9
payload								

- Morphological Analysis:

com.basistech.rlp.solr.RLPTokenizerFactory

term position	1	2	3	4
term text	東京	ルパン	上映	時間
source start,end	0,2	3,6	6,8	8,10

# Token Processing

- Stemming vs. Lemmatization
- English: “I have spoken at several conferences”
- Stemming:

org.apache.solr.analysis.EnglishPorterFilterFactory {protected=prot

term position	1	2	3	4	5	6
term text	i	have	spoken	at	sever	conferences.
term type	word	word	word	word	word	word
source start,end	0,1	2,6	7,13	14,16	17,24	25,37
payload						

- Lemmatization:

org.apache.solr.analysis.RemoveDuplicatesTokenFilterFactory {}

term position	1	2	3	5	6
term text	i	have	spoken	several	conferences
term type	word	word	word	word	word
lemma					
source start,end	0,1	2,6	7,13	17,24	25,36
payload					

# Token Processing

German: “Am Samstagmorgen fliege ich zurueck nach Boston.”

- Stemming:

org.apache.solr.analysis.SnowballPorterFilterFactory {lang=de}

term position	1	2	3	4	5	6	7
term text	am	Samstagmorg	flieg	ich	zurueck	nach	boston.
term type	word	word	word	word	word	word	word
source start,end	0,2	3,16	17,23	24,27	28,35	36,40	41,48
payload							

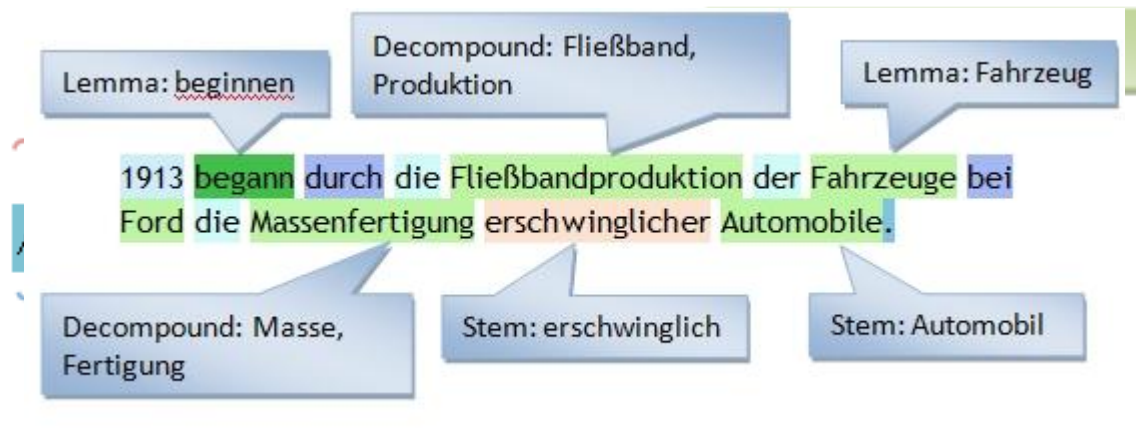
- Lemmaatization (and decompounding!):

com.basistech.rlp.solr.RLPTokenizerFactory

term position	1	2	3	4	5	6	7	8
term text	am	Samstagmorgen	fliege	ich	zurueck	nach	boston.	
	an der	Samstag	fliegen		zurück			
		morgen						
term type	word	word	word	word	word	word	word	word
lemma		comp	lemma		lemma			
		comp						

# Additional Linguistic Analysis

- Decomponding
- Part of Speech
- Noun Phrase Extraction



# Higher Level Analysis

- Entity Extraction
- Relationship Extraction

Named Entity (# instances)	
IDENTIFIER:MONEY	10
IDENTIFIER:NUMBER	14
IDENTIFIER:URL	1
LOCATION	3
ORGANIZATION	13
PERSON	2
TEMPORAL:DATE	4
TEMPORAL:TIME	2
	4

By Tomoeh Mu  
Washington Po  
Thursday, Jan

NEW YORK --  
pool of \$16.2

The results re  
financial crisi  
recovered.

The compensa  
- translates to  
bankers will e  
than the pre-c

Add Relationships for:  Add

**Relationships**

**Current Entity Relationships**  
**Ibrahim Hassan al-Asiri**

**hasSibling:**

- Ali Abdullah Saleh

**communicatedWith:**

- al-Qaeda

**Also Mentioned As:**

- Ibrahim
- Hassan al-Asiri
- Ibrahim al-Asiri
- Asiri
- Al-Asiri
- prophet Ibrahim
- Abdullah
- Mr Asiri
- Hassan
- al-Asiri

To display the relationship details, click on an Entity icon

compensation

the  
markets

bonuses -  
ers and  
high lower

# How to Configure Solr

- Goals:
  - Language Identification
  - Language-aware:
    - *Tokenization*
    - *Token Processing*
    - *Entity Extraction*
  
- Challenges
  - Multiple languages in the data set

# How to Configure Solr

- What Solr tools do we have to work with?
  - **UpdateRequestProcessor**
  - **Analyzer/Tokenizer/TokenFilter**
  - **Solr Cores**
  
- Pre-processor to Solr?

# UpdateRequestProcessor

- Runs Before Analyzers
- Full Access to Document
  
- Two options:
  - **Run the analysis directly in Solr**
    - *Good for Lightweight Analysis*
  - **Call out to external analysis services**
    - *Web Services/UIMA. Increases Complexity*
  
- Limitations:
  - **Think through your indexing strategy**

# Analyzer/Tokenizer

- Good for:
  - Segmentation of Asian Language
  - Linguistics
- Limitations:
  - No access to document object
- Schema.xml
  - FieldType
    - *Analyzer*
      - CharFilter
      - Tokenize
      - TokenFilter

# How to Configure Language ID

- UpdateRequestProcessor
  - Runs before field-level analysis takes place
  - Basic Language Identifier URP to be included in Solr
- Outside Solr

What do you do with the language information??

# Multiple Languages: Method 1

- One field for each language
  - **Pro:**
    - *Simple approach and implementation*
    - *Guarantees that queries are processed the same way as index*
  - **Con:**
    - *Increased query-time complexity (Dismax, etc)*
    - *Decreased speed as additional fields are queried*
    - *May require storing multiple copies of text*

Informed by Trey Grainger @ Careerbuilder: <http://www.lucidimagination.com/sites/default/files/Grainger%20Trey%20-%20Extending%20Solr,%20Building%20a%20Cloud-Like%20Knowledge%20Discovery%20Platform%20-%20rev.pdf>

# Multiple Languages: Method 2

- One Solr core per language

Each Core has the same field, with a language-specific Analyzer/Tokenizer

- **Pros:**

- *No query-time performance overhead*
- *Guarantees that queries are processed the same way as index*

- **Cons:**

- *Significant complexity in managing multiple cores*
- *Must implement custom sharding*
- *Does not support multilingual documents*

# Multiple Languages: Method 3

- All Languages in one field
  - **Pros:**
    - *Single field makes queries and indexing easy*
    - *Same schema/core as more languages added*
  - **Cons:**
    - *Requires complex custom Tokenizer/Analyzer*
    - *Must pass in language information for queries and indexing*
    - *Does not guarantee queries are processed the same as the index*
    - *Potential TF/IDF confusion*

# Language is Important

- Use language information at index and query time
- Increase recall, maintain precision
  
- Better search results for your users

# Contact

- Steve Kearns
  - [skearns@basistech.com](mailto:skearns@basistech.com)
  - <http://www.basistech.com>