

Integrating Advanced Text Analytics into Solr

Lucene Revolution

Steve Kearns

Product Manager

www.basistech.com



BASIS
TECHNOLOGY

Agenda

- About Basis Technology
- Why Text Analytics and Solr?
- Overview and Uses of Text Analytics
- Integration Strategies

About Basis Technology

- HQ in Cambridge, MA, Offices in:
 - ⇒ *Tokyo, San Francisco, Washington DC*
- Specialists in **multilingual text analytics** for
 - ⇒ *Web/enterprise search*
 - ⇒ *Document/OSINT/media exploitation*
- **Rosette Linguistics Platform** is widely used by commercial enterprises and government agencies

Why Text Analytics and Solr?

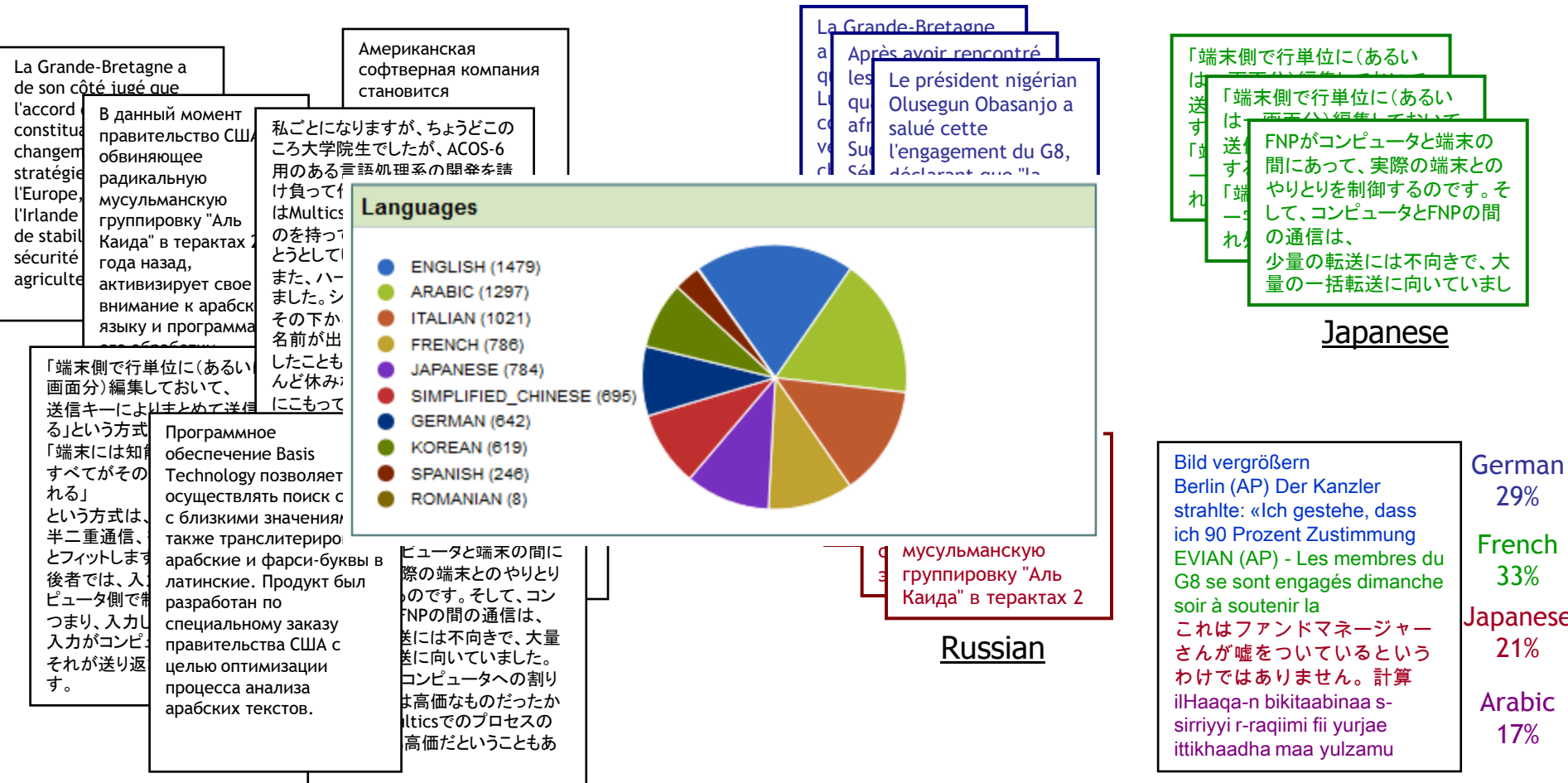
- More than Keyword Search and Result Lists
- More Metadata
 - ⇒ *New ways to visualize, navigate and explore*
 - ⇒ *New knobs to tune relevance*
 - ⇒ *New info to connect disparate data sources*
- Solr can be the consumer, host, or broker

Overview of Text Analytics

- Document-Level
 - ⇒ *Language Identification, Categorization*
- Sub-Document Level
 - ⇒ *Entity Extraction, Fact Extraction, Sentiment, Linguistics*
- Cross-Document
 - ⇒ *Cross-Document Entity Resolution, Near Duplicate Detection, Unsupervised Clustering*

Document Level Analysis: Language Identification

- Sub-document Lang ID is possible



Document Level Analysis: Categorization

- Group Documents into Pre-defined categories

- World
- San Francisco Bay Area
- U.S.
- Business
- Sci/Tech
- Sports
- Entertainment
- Health

Explore the BBC

News » Africa Asia-Pacific Europe Latin America Middle East South Asia UK US & Canada Business	Sport » Football Cricket Tennis Motorsport Business » Market Data Economy Companies	Entertainment » Entertainment News Comedy Drama Music » Genres Reviews Artists News	Science » Humans Space Nature » Animals Gardening » Plant finder Advice
--	---	---	---

Sub-Document Analysis: Linguistics

- Segmentation of Asian language
- Lemmatization

Stepping

org.apache.solr.analysis.EnglishPorterFilterFactory

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						

er {}

6	7	8
上映	映時	時間
5,7	6,8	7,9

Morphological
Segmentation

com.basistech.rlp.solr.RLPTokenizerFactory






term position	1	2	3	5	6
term text	i	have	spoken	several	conferences
term type	word	word	word	word	word
			lemma		lemma

Factory

Sub-Document Analysis: Sentiment

- Sentence, paragraph, entity, aspect, emotion

What people are saying

<input type="checkbox"/> rooms		"Clean, friendly; great value for the money and great location." - priceline.com
<input type="checkbox"/> service		"Service-oriented elegance just minutes from San Francisco Airport" - orbitz.com
<input type="checkbox"/> location		"Great location, gorgeous lobby and restaurant." - priceline.com
		"Service-oriented elegance just minutes from San Francisco Airport" - orbitz.com
		"Clean, friendly; great value for the money and great location." - priceline.com
		"So we book at this location as we did want to worry about uncleanliness." - priceline.com
		"Good air quality, well-appointed lobby, convenient shuttle to SFO" - priceline.com
		"Good location, private, good staff." - priceline.com
		"Nice hotel, comfortable room, quiet location." - priceline.com
<input type="checkbox"/> value		"It has a restaurant but it's expensive." - priceline.com
<input type="checkbox"/> dining		"Rooms are very nice. Parking \$18 a night, no complementary continental breakfast." - travelpost.com



BASIS
TECHNOLOGY

Sub-Document Analysis: Entity Extraction

- Identify Named Concepts in Unstructured Text
 - ⇒ *Statistical, rules, lists*

The **Haiti** earthquake occurred at a fault that runs right through **Haiti** and is situated

17:50:15 **Buzzing right now** 

ada afraid american attend automatically avi award blah **bosh**
bulls catch depressed doe doors **dunk** duty
dwts earned espn expression finale forth game gaming
handle hang heat idol jackson joint jokes **judges** khloe
laughing lauren lebron legal less lose loser meloni miami **movie**
noah notification pain paper picked pregnant preparing
randy rose sadly sang **scotty**
sends shoot strength studying swift tag tattoos taylor timeline tonite
tooo unknown **wade** weight yuu

n and North
er like a giant
0.8 inches a
h American
very small, a
the **U.S. Ge**
IEIC).

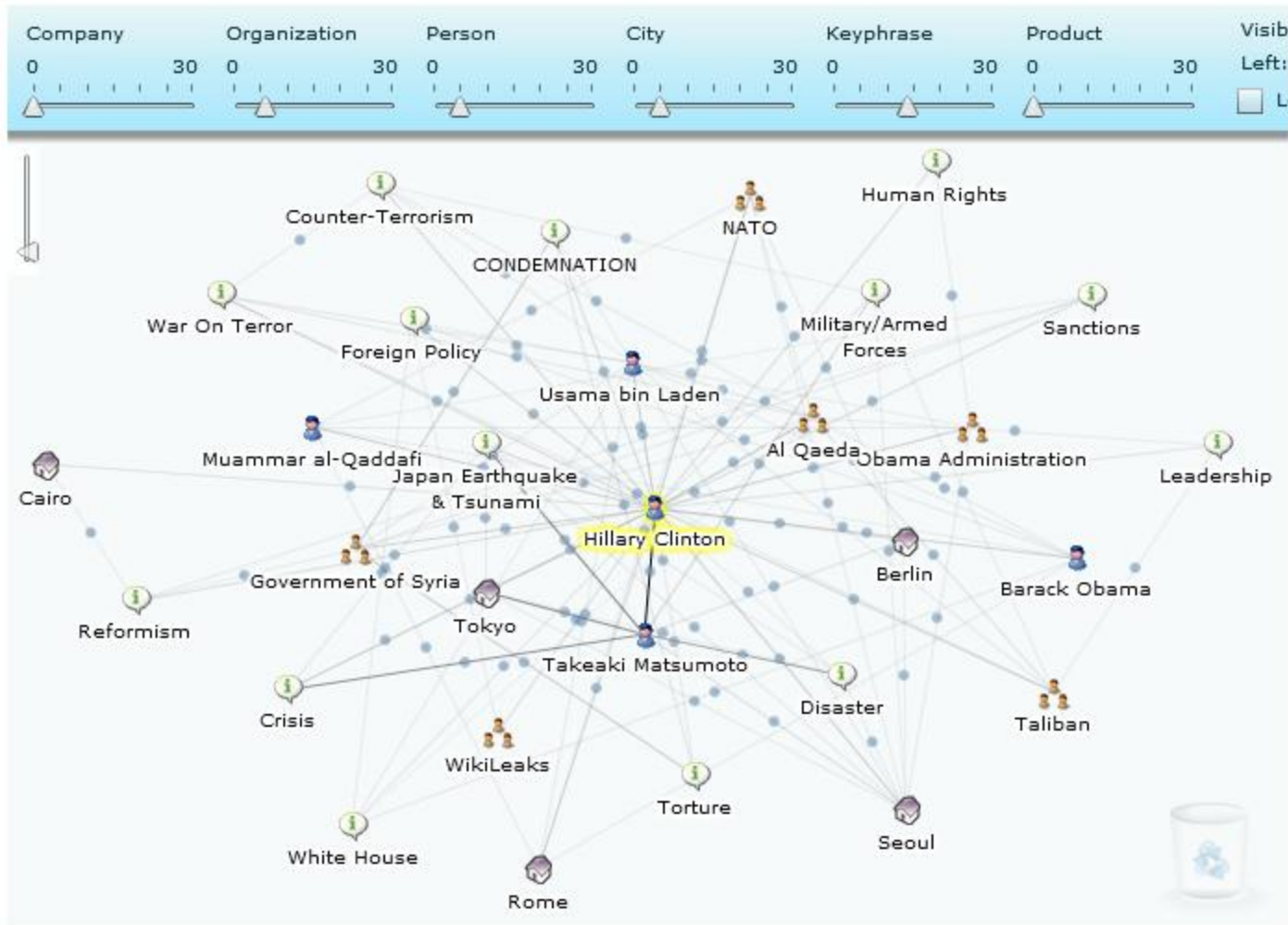
are rocky
two plates
an plate
ple felt,"
and the
med Entity (# instances)

Entity Type	# instances
ENTIFIER:DISTANCE	2
ENTIFIER:NUMBER	16
ENTIFIER:URL	1
CATION	15
ATIONALITY	5
GANIZATION	5
RSON	3
TEMPORAL:DATE	4
TEMPORAL:TIME	5

Filter Results By	
Languages	
ENGLISH	184 🔍
People	
Muammar el-Qaddafi	121 🔍
Bryan Denton	38 🔍
Moussa Ibrahim	34 🔍
C.J. Chivers	26 🔍
John F. Burns	22 🔍
Obama	21 🔍
Locations	
Misurata	184 🔍
Libya	167 🔍
Tripoli	100 🔍
United States	62 🔍
France	56 🔍
Italy	53 🔍

Sub-Document: Fact / Rel. / Event Extraction

- Identify Facts, Link Entities, Events and Times



Cross-Document: Entity Co-reference Resolution

- Map extracted entities to real-world Concepts

People

Osama bin Laden

Barack Obama

Bashar al-Assad

Muammar Gaddafi

Hosni Mubarak

Ali Abdullah Saleh

David Cameron

Osama Bin Laden

David Cameron

Cross-Document Analysis: Clustering

- Near Duplicate Detection
- Unsupervised Clustering

[Stephen Foley: Yandex, the well-oiled search engine](#) ☆

The Independent - 3 hours ago

It is no surprise, then, that interest in the flotation of Yandex, Russia's leading **search engine**, should have spiked after the LinkedIn debut. There were 17 times more orders for Yandex stock than there were shares available, so bankers did what they ...

[Russia's biggest search engine Yandex soars on first day of trading](#) Telegraph.co.uk

[Yandex, Russia's Search Engine, Jumps 43% On Debut](#) Barron's (blog)

[Russian Search Engine Yandex To Debut On Wall Street](#) RTT News

[New Zealand Herald](#)

[all 500 news articles](#) »  LNKD

We'll also celebrate this month's birthdays by having sweet

treats for everyone. celebrate this month's birthdays by having cake/goodies.

Integration Strategies

- Analyzer/Tokenizer/TokenFilter
- UpdateRequestProcessor
 - ⇒ *Run Analysis in Solr*
 - ⇒ *Call External Analysis Service*
- Pre-Processor to Solr

Integration Point: Analyzer/Tokenizer

- Good for:
 - ⇒ *Linguistics*
 - ⇒ *Segmentation of Asian Language*
- Limitations:
 - ⇒ *No access to document object*

Analyzer/Tokenizer Configuration

- Schema.xml
 - ⇒ *FieldType*
 - Analyzer
 - CharFilter
 - Tokenize
 - TokenFilter

Integration Point: UpdateRequestProcessor

- Runs Before Analyzers
- Full Access to Document
- Two options:
 - ⇒ *Run the analysis directly in Solr*
 - ⇒ *Call out to external analysis services*
- Limitations:
 - ⇒ *Think through your indexing strategy*

Integration Point: UpdateRequestProcessor

- Run the analysis directly in Solr
 - ⇒ *Good for light weight analytics*
 - ⇒ *Not good for cross-document analytics*

- Call out to external analysis services
 - ⇒ *Web Services, UIMA, OpenPipeline, GATE, custom code*
 - ⇒ *Note that these external calls are synchronous*
 - ⇒ *Additional complexity / points of failure*

UpdateRequestProcessor Configuration

- SolrConfig.xml
 - ⇒ *RequestHandler*
 - update.processor = UpdateRequestProcessorChain.name
 - ⇒ *UpdateRequestProcessorChain*
 - Processors

```
<requestHandler name="/update" class="solr.XmlUpdateRequestHandler">  
  <lst name="defaults">  
    <str name="update.processor">langidentifier</str>  
  </lst>  
</requestHandler>
```

```
<updateRequestProcessorChain name="langidentifier">  
  <processor class="com.basistech.rlp.solr.LanguageIdentifierUpdateProcessorFactory">  
    <str name="rlpContext">solr/conf/rlp-context-rli.xml</str>  
    <str name="fields">text</str>  
    <bool name="iso2LetterCode">>false</bool>  
    <bool name="preserveField">>false</bool>  
    <bool name="createLangCodeField">>true</bool>  
  </processor>  
  <processor class="solr.LogUpdateProcessorFactory"/>  
  <processor class="solr.RunUpdateProcessorFactory"/>  
</updateRequestProcessorChain>
```



Integration Point: Pre-Processor

- Index in Solr as Last Step of Analysis
- Good For:
 - ⇒ *Finer-grained control*
 - ⇒ *Managing dependencies between components*
 - ⇒ *Scalability*
- Limitations:
 - ⇒ *Complexity / New points of failure*
 - ⇒ *Cannot use Solr's content acquisition features*

Integration Summary

- There are Many Options!
- Document-Level Analysis:
 - ⇒ *Generally, safe to run in UpdateRequestProcessor*
- Sub-Document Analysis:
 - ⇒ *Sometimes run in UpdateRequestProcessor, sometimes external*
- Cross-Document Analysis:
 - ⇒ *Run external*
- Multiple-Analysis Components:
 - ⇒ *Run external document processing pipeline*

Questions?