Extract insight from texts using SAP Text Analysis

Tomer Steinberg
SAP Israel
Agenda

Why use text analysis functionality?

Background: SAP’s text analysis technology

Search: Full-text search and fuzzy search

Text analysis: Entity and fact extraction

Text mining

Wrap-up
Why use text analytics
Why Text Analytics

Enterprise Challenges

Massive amounts of data locked

Companies are struggling to:

• Search on unstructured text related content
• Extract meaningful, structured information from unstructured text
• Combine unstructured with structured data
• Leverage data in real-time to gauge and guide their business strategy and solve critical problems
Potential use cases

Law enforcement
Intelligence
Social Media Analytics
Precision Marketing
Predictive Maintenance
Investment trade
Credit Scoring
Patents
Mary's interest is leggings

Jane's interest is jackets

Sue's looking for a fleece

REAL TIME INTENT SIGNALS SHOW THAT:

Mary's free shipping offer is focused on leggings

Jane's is on jackets

Sue's is on fleece

Single template with dynamic content

Get yours now, with Free shipping!

Free shipping on %category% and more thru 11/28

Mary's free shipping offer

Jane's is on jackets

Sue's is on fleece
SAP’s Text Analysis Technology
Background: SAP’s text analysis technology

1997
Inxight spun off from PARC, a Xerox Company
Finite-State technology for modeling natural language

2007
Inxight acquired by Business Objects
Integration of text analysis technology into BI applications

2008
Business Objects acquired by SAP
Text analysis technology continues to focus on BI applications

2012
First integration into SAP HANA
Foundation for full-text search, BI and sentiment analysis applications

Today
Text analysis in SAP HANA
Foundation for virtually any type of unstructured textual data processing in the platform
Background: SAP’s text analysis technology

The development team is one of the SAP HANA core teams

SAP Labs in Boston, MA
- Located near Kendall Square in Cambridge, close to MIT

14 engineers

7 computational linguists
SAP HANA Platform – More than just a database

SAP HANA platform converges Database, Data Processing and Application Platform capabilities & provides Libraries for predictive, planning, text, spatial, and business analytics so businesses can operate in real-time.
Why does SAP HANA provide text analysis functionality?

**Capabilities**
- Native full-text and fuzzy search
- In-database text analysis
- Graphical modeling of search models
- Info Access – HTML5 UI toolkit and API for JavaScript

**Benefits**
- Less data duplication and movement – leverage one infrastructure for analytical and search workloads
- Extract salient information from unstructured textual data
- Easy-to-use modeling tools – HANA Studio
- Build search applications quickly – Info Access
What types of text processing capabilities are supported?

Search
In addition to string matching, HANA features full-text search which works on content stored in tables or exposed via views. Just like searching on the Internet, full-text search finds terms irrespective of the sequence of characters and words.

Text analysis
Capabilities range from basic tokenization and stemming to more complex semantic analysis in the form of entity and fact extraction. Text analysis applies within individual documents and is the foundation for both full-text search and text mining.

Text mining
Text mining makes semantic determinations about the overall content of documents relative to other documents. Capabilities include key term identification and document categorization. Text mining is complementary to text analysis.
What types of text processing capabilities are supported?

**Search**

In addition to string matching, HANA features **full-text search** which works on content stored in tables or exposed via views. Just like searching on the Internet, full-text search finds terms irrespective of the sequence of characters and words.

**Text mining** makes semantic determinations about the overall content of documents relative to other documents. Capabilities include key term identification and document categorization. Text mining is complementary to text analysis.

**Text analysis** capabilities range from basic tokenization and stemming to more complex semantic analysis in the form of entity and fact extraction. Text analysis applies within individual documents and is the foundation for both full-text search and text mining.
Search

In addition to string matching, HANA features full-text search which works on content stored in tables or exposed via views. Just like searching on the Internet, full-text search finds terms irrespective of the sequence of characters and words.

Text analysis

Capabilities range from basic tokenization and stemming to more complex semantic analysis in the form of entity and fact extraction. Text analysis applies within individual documents and is the foundation for both full-text search and text mining.

Nicole Kidman, Aaron Eckhart and ‘Rabbit Hole’
By MEKADO MURPHY
Dan Steinberg/Associated Press

Aaron Eckhart and Nicole Kidman at the Toronto International Film Festival

TORONTO — Nicole Kidman returns to Toronto, this time in the role of both actor and producer for her latest project, “Rabbit Hole.” The film, in which she co-stars with Aaron Eckhart, looks at a suburban married couple who experience a tremendous loss. “Rabbit Hole” is based on the play by David Lindsay-Abaire, who also adapted it for the screen. The play received a positive review when it premiered at Manhattan Theater Club to the attention of Ms. Kidman and producing partner Per Saari, who decided to option it.

Ms. Kidman and Mr. Eckhart shared some thoughts about the new film and the process of working with their director, John Cameron Mitchell.
What types of text processing capabilities are supported?

**Search**

In addition to string matching, **HANA** features full-text search which works on content stored in tables or exposed via views. Just like searching on the Internet, full-text search finds terms irrespective of the sequence of characters and words.

**Text mining**

Text mining makes semantic determinations about the overall content of documents relative to other documents. Capabilities include key term identification and document categorization. Text mining is complementary to text analysis.

**Text analysis**

Capabilities range from basic tokenization and stemming to more complex semantic analysis in the form of entity and fact extraction. Text analysis applies within individual documents and is the foundation for both full-text search and text mining.

---

**At Dresden Semperoper, a New Take on 'Tristan and Isolde'**

By ROSLYN SULCAS February 17, 2015

DRESDEN, Germany — David Dawson’s new “Tristan and Isolde” for the Dresden Semperoper Ballett raises interesting questions about the full-length story ballet, a seldom tackled genre.

The story, a medieval Celtic tale that has long figured in literature, film and in Wagner’s opera of the same name, has been compared to “Romeo and Juliet.”

Once, Europeans were happy to pay for separate cellphone, cable and pay-TV services. Now, they prefer them bundled into a single package that streams content to any device — a smartphone, tablet or Internet-connected television.

Regional rivals like Orange of France and Deutsche Telekom of Germany have moved quickly to offer...

---

**Vodafone Turns Focus to Broadband, Seeking to Catch Up to Rivals**

By MARK SCOTT February 16, 2015

As consumers change the way they use their smartphones, surf the web and watch television, Vodafone is finding itself in need of a face-lift. After years of focusing heavily on its cellphone business, Vodafone, based in Britain and the world's second-largest mobile operator behind China Mobile based on subscribers, is concentrating on high-speed broadband.

Once, Europeans were happy to pay for separate cellphone, cable and pay-TV services. Now, they prefer them bundled into a single package that streams content to any device — a smartphone, tablet or Internet-connected television.

Regional rivals like Orange of France and Deutsche Telekom of Germany have moved quickly to offer...

---

<table>
<thead>
<tr>
<th>Category</th>
<th>Classical_Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key terms</td>
<td>Semperoper, Wagner, ballet, John Cranko, Royal Ballet School</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Telecommunications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key terms</td>
<td>Vodafone, broadband, cellphone business, Orange, Deutsche Telekom, …</td>
</tr>
</tbody>
</table>
A full-text index – required for Google-like searches – is defined on a table column

The table column is ‘aware’ of its index – insert, update, delete is handled automatically

Fast delta indexing

Broad language identification & processing

<table>
<thead>
<tr>
<th>Available Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
</tr>
<tr>
<td>Catalan</td>
</tr>
<tr>
<td>Chinese (Simplified)</td>
</tr>
<tr>
<td>Chinese (Traditional)</td>
</tr>
<tr>
<td>Croatian</td>
</tr>
<tr>
<td>Czech</td>
</tr>
<tr>
<td>Danish</td>
</tr>
<tr>
<td>Dutch</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>Farsi</td>
</tr>
<tr>
<td>French</td>
</tr>
<tr>
<td>German</td>
</tr>
<tr>
<td>Greek</td>
</tr>
<tr>
<td>Hebrew</td>
</tr>
<tr>
<td>Hungarian</td>
</tr>
<tr>
<td>Italian</td>
</tr>
</tbody>
</table>
The following steps are executed on unstructured text:

1. File format filtering
   - Converts any binary document format to text/HTML

2. Language detection
   - Identifies language to apply appropriate tokenization and stemming

3. Tokenization
   - Decomposes word sequences
     E.g. “card-based payment systems” → “card” “based” “payment” “systems”

4. Stemming
   - Normalizes tokens to linguistic base form
     E.g. houses → house; ran → run

5. Full-text index
   - ‘Attaches’ to the table column
Text Analysis
Text analysis
An option to the full-text index

The following steps may be executed on unstructured text to augment full-text indexing:

- **Part-of-Speech**
  - Tags word categories
    - Examples: quick: Adj; houses: Nn-Pl

- **Noun groups**
  - Identifies concepts
    - Examples: text data; global piracy

- **Entity extraction**
  - Classifies pre-defined entity types
    - Examples: Winston Churchill: PERSON; U.K.: COUNTRY;

- **Fact extraction**
  - Relates entities – e.g., classifies sentiments with topics
    - Example: I love SAP HANA:
      [Sentiment] I [StrongPositiveSentiment] love [/StrongPositiveSentiment]
      [Topic] SAP HANA [/Topic].[/Sentiment]
Text analysis gives ‘structure’ to two sorts of elements from unstructured text:

Entities:

John Lennon was one of the Beatles.

\(<\text{PERSON}>\text{John Lennon}\</\text{PERSON}>\text{ was one of the}\n\<\text{ORGANIZATION}@\text{ENTERTAINMENT}>\text{Beatles}\</\text{ORGANIZATION}@\text{ENTERTAINMENT}>\.

Facts:

I love your product.

I \<\text{STRONGPOSITIVESENTIMENT}>\text{love}\</\text{STRONGPOSITIVESENTIMENT}> \<\text{TOPIC}>\text{your product}\</\text{TOPIC}>\.
Text analysis
Supported types for entity extraction

Who: People, job title, and national identification numbers

What: Companies, organizations, financial indexes, and products

When: Dates, days, holidays, months, years, times, and time periods

Where: Addresses, cities, states, countries, facilities, internet addresses, and phone numbers

How much: Currencies and units of measure

Generic concepts: text data, global piracy, and so on

Languages:
Arabic, English, Dutch, Farsi, French, German, Italian, Japanese, Korean, Portuguese, Russian, Simplified Chinese, Spanish, Traditional Chinese
Text analysis
Supported fact extraction (1/2)

Voice of customer

Sentiments: strong positive, weak positive, neutral, weak negative, strong negative, and problems
Requests: general and contact info
Emoticons: strong positive, weak positive, weak negative, strong negative
Profanity: ambiguous and unambiguous

Languages:
English, Dutch*, French, German, Italian, Portuguese, Russian, Simplified Chinese, Spanish, Traditional Chinese

*Emoticons and profanity only
Text analysis
Supported fact extraction (2/2)

Enterprise
Membership information
Management changes
Product releases
Mergers & acquisitions
Organizational information

Public Sector
Action & travel events
Military units
Person-alias, -appearance, -attributes, -relationships
Spatial references
Domain-specific entities

Language: English
Text analysis
How entity extraction works

Built-in entity extraction is **not** keyword search. Text analysis applies full linguistic and statistical techniques (i.e., natural language processing) to make sure the entities which get returned are correct.

**Grammatical Parsing:**

- *Can we bill you?*
- *Bill Smith was the president.*

**Semantic Disambiguation:**

- *I talked to Bill yesterday.*
- *The bill was signed into law*
Significant investment in gold corpus development across languages to achieve objective, repeatable assessments of entity and fact extraction (i.e., blind testing).
## Language Support

<table>
<thead>
<tr>
<th>Language</th>
<th>LINGANALYSIS_BASIC/STEMS</th>
<th>LINGANALYSIS_FULL</th>
<th>EXTRACTION_CORE</th>
<th>EXTRACTION_CORE_VOICEOFCUSTOMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese (Simplified)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese (Traditional)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farsi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norwegian (Bokmal)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norwegian (Nynorsk)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serbian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovak</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thai</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Demo
Step 1

Load Documents

Analyze Results

Create Text Index
Step 2

Load Documents

Create Text Index

Analyze Results
Step 3

1. Load Documents
2. Create Text Index
3. Analyze Results
Text mining
Text mining

Text mining works at the document level

Determinates the overall content of documents relative to other documents.

Used for:
- Identify similar documents
- Identify key terms of a document
- Identify related terms
- Categorize new documents based on a training corpus

Scenarios
- Highlight the key terms when viewing a patent document
- Identify similar incidents for faster problem solving
- Categorize new scientific papers along a hierarchy of topics
Text Mining Demo
Wrap Up

Structure massive amounts of unstructured data

- Search on unstructured text related content
- Extract meaningful, structured information from unstructured text
- Combine unstructured with structured data
- Leverage data in real-time to gauge and guide their business strategy and solve critical problems

Business Benefits

- Understand your Customer/Process
Wrap-Up

Tomer Steinberg
Tomer.Steinberg@sap.com