Introduction to Text Mining

Advanced Oral Communication Skills for Science and Humanities

unstructured text → perform some text mining... → structured data
Overview

• Motivation

• Information Overload

• Information – Structured vs. Unstructured

• Delimitation – Text Mining vs. Related Areas

• Text Mining vs. Data Mining

• Text Mining - Methods Overview

• Take-Home-Messages
Motivation

• What is exactly Text Mining?

• One possible definition (in accordance to Sebastian Blümel, [1])

> "Text mining is a collection of techniques and algorithms for the automatic analysis of unstructured data, where the goal is to derive high-quality knowledge from text."

• Hmmm...is this really important innovation?

• Yes, since in Text Mining is handled by a **machine** rather than by humans...
Information Overload

• Why are we interested in the process of deriving information automatically?

• Imagine the following fact...


• 1 Exabyte = 1.000.000.000.000.000.000 Gigabytes!

• 1 Gigabyte ≈ 262.144 DIN-A4 pages

• So would you like to discover relevant information in:
  294.649.856.000.000.000.000.000.000 pages manually?
Information Overload

• More facts...

- "The information consumption (over the lifetime) of a person in the 17th century equals approx. the amount of information in a today's Sunday newspaper"!

- "In the year 2020 the amount of information will be doubled every 73 days"!

- "Researchers waste 25% of their work with searching for relevant information"!

- "Since 500 years, the ability of our brain to process information has improved slightly"!

[2]
Information –
Structured vs. Unstructured

• Many information... → Major part of them is represented through text!

Unfortunately ≈ 90% of these information are unstructured...

• Web pages (Facebook, blogs, ...), e-Mails, manuals, e-Books, etc...

• Is this a problem? Yes, since we need a structured within the information, in order to interprete it...
Information – Structured vs. Unstructured

Example: Are you able to interpret something from the following unstructured text?

11902195901792888462...
23112200301704398149...
31208199101631347545...
...

Now better?

<table>
<thead>
<tr>
<th>ID Number:</th>
<th>Birthdate:</th>
<th>Mobile Number:</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.Feb.1959</td>
<td>0179 - 2888462</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>31.Dec.2003</td>
<td>0170 - 4398149</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>12.Aug.1991</td>
<td>0163 - 1347545</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Delimitation – Text Mining vs. Related Areas

Text Mining

- Information Retrieval
- Information Extraction
- Data Mining
- Computer Linguistics
- Machine Learning
Text Mining vs. Data Mining

The most similar area to Text Mining is the so-called “Data Mining” discipline.

Data Mining: Extract implicit and potentially useful information from structured (e.g. databases...)

Text Mining: Apply Data Mining techniques on unstructured text documents!

<table>
<thead>
<tr>
<th>Search</th>
<th>Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured Data</td>
<td>Data Retrieval</td>
</tr>
<tr>
<td>Unstructured Data</td>
<td>Information Retrieval</td>
</tr>
</tbody>
</table>
Text Mining - Methods Overview

- Many methods are typically involved in a Text Mining process...
  - Automatic Text Summarization
  - Categorization / Classification
  - Clustering
  - Information Extraction
  - Association Rule Learning
  - and many more...
Text Mining - Methods Overview

- Many methods are typically involved in a Text Mining process...

- Automatic Text Summarization

- Categorization / Classification

- Clustering

- Information Extraction

- Association Rule Learning

- and many more...
Text Mining - Methods Overview

- Many methods are typically involved in a Text Mining process...
- Automatic Text Summarization
- Categorization / Classification
- Clustering
- Information Extraction
- Association Rule Learning
- and many more...

1: from sister
2: from cousin
3: from grandpa
4: from uncle Jim
5: from Mr. X

Classifier

Training data
Text Mining - Methods Overview

• Many methods are typically involved in a Text Mining process...

• Automatic Text Summarization

• Categorization / Classification

• Clustering

• Information Extraction

• Association Rule Learning

• and many more...

Programming language...

Island...

Coffee...
Text Mining - Methods Overview

- Many methods are typically involved in a Text Mining process...

- Automatic Text Summarization

- Categorization / Classification

- Clustering

- Information Extraction

- Association Rule Learning

- and many more...

19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb – allegedly detonated by urban guerilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

- bomb
- March 19
- El Salvador: San Salvador (city)
- urban guerilla commandos
- power tower
- destroyed
- no injury or death
- bomb

Terrorist report

<table>
<thead>
<tr>
<th>INCIDENT TYPE</th>
<th>DATE</th>
<th>LOCATION</th>
<th>PERPETRATOR</th>
<th>PHYSICAL TARGET</th>
<th>HUMAN TARGET</th>
<th>EFFECT ON PHYSICAL TARGET</th>
<th>EFFECT ON HUMAN TARGET</th>
<th>INSTRUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>bombing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>El Salvador: San Salvador (city)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>urban guerilla commandos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>power tower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>destroyed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no injury or death</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bomb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Text Mining - Methods Overview

Many methods are typically involved in a Text Mining process...

• Automatic Text Summarization
• Categorization / Classification
• Clustering
• Information Extraction
• Association Rule Learning
• and many more...

| Customer Trans. ID | Corn Flakes | Annanas | Bread | Yoghurt | ...
|--------------------|-------------|---------|-------|---------|-------|
| 1                  | yes         | no      | yes   | yes     | ...
| 2                  | no          | yes     | no    | no      | ...
| 3                  | yes         | no      | no    | yes     | ...
| 4                  | no          | yes     | no    | yes     | ...
| 5                  | yes         | yes     | yes   | yes     | ...
| ...                | ...         | ...     | ...   | ...     | ...

\{Corn Flakes, Bread\} \rightarrow Yoghurt
Text Mining:

• ...is very valuable technique (if you know where and how to use it)

• ...becomes very popular in many fields (beyond computer science)

• ...doesn’t require a deep understanding (semantics) of the text

• ...is maybe not 100% bulletproof, but still offers a great opportunity to gain the most relevant information from a text...
Questions...?
Thanks for your attention...
References

Seminar Slides by: Sebastian Blümel, CN8,
Hochschule Furtwangen University, (unknown year – URL doesn’t exist anymore !)


Aulas Teóricas: Text Mining
http://paginas.fe.up.pt/~ec/edicao04_05.html
http://www8.informatik.uni-erlangen.de/inf8/nocov/TM-slides/information-extraction.pdf


Photo of "Le Penseur", a bronze sculpture made by Auguste Rodin, held in the Musée Rodin in Paris, France.

All links have been accessed successfully on: 29.11.2011