Introduction to Text Mining

Advanced Oral Communication Skills for Science and Humanities

TECHNISCHE UNIVERSITÄT DARMSTADT

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

- Imagine the following fact…

> *"The Amount of Digital Information Reached 281 Exabytes"* [2]

- 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 pages manually ?

Information Overload

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- Many information… → Major part of them is represented through text !

Unfortunately ≈ 90% of these information are unstructured… [3]

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

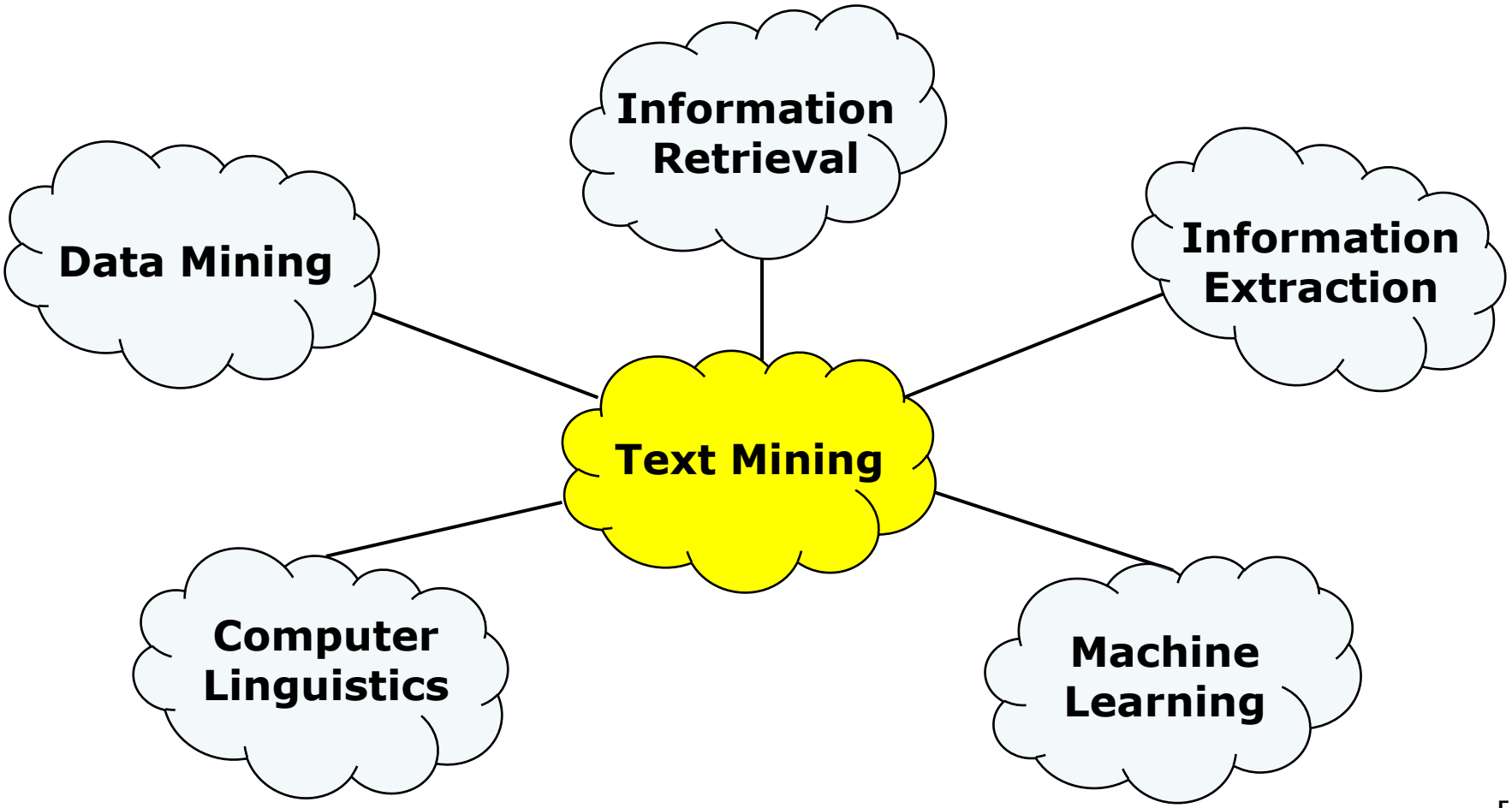
Example: Are you able to interprete something from the following unstructured text?

1190219590179288846 2...
2311220030170439814 9...
3120819910163134754 5...
...

Now better?

ID Number:	Birthdate:	Mobile Number:	...
1	19.Feb.1959	0179 - 2888462	...
2	31.Dec.2003	0170 - 4398149	...
3	12.Aug.1991	0163 - 1347545	...
...	...	...	...

Delimitation – Text Mining vs. Related Areas



[1]

Text Mining vs. Data Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

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 !

	Search	**Discovery**
Structured Data	Data Retrieval	Data Mining
Unstructured Data	Information Retrieval	Text Mining

[3]

Text Mining - Methods Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- Many methods are typically involve
- Automatic Text Summarization
- Categorization / Classification
- Clustering
- Information Extraction
- Association Rule Learning
- and many more...

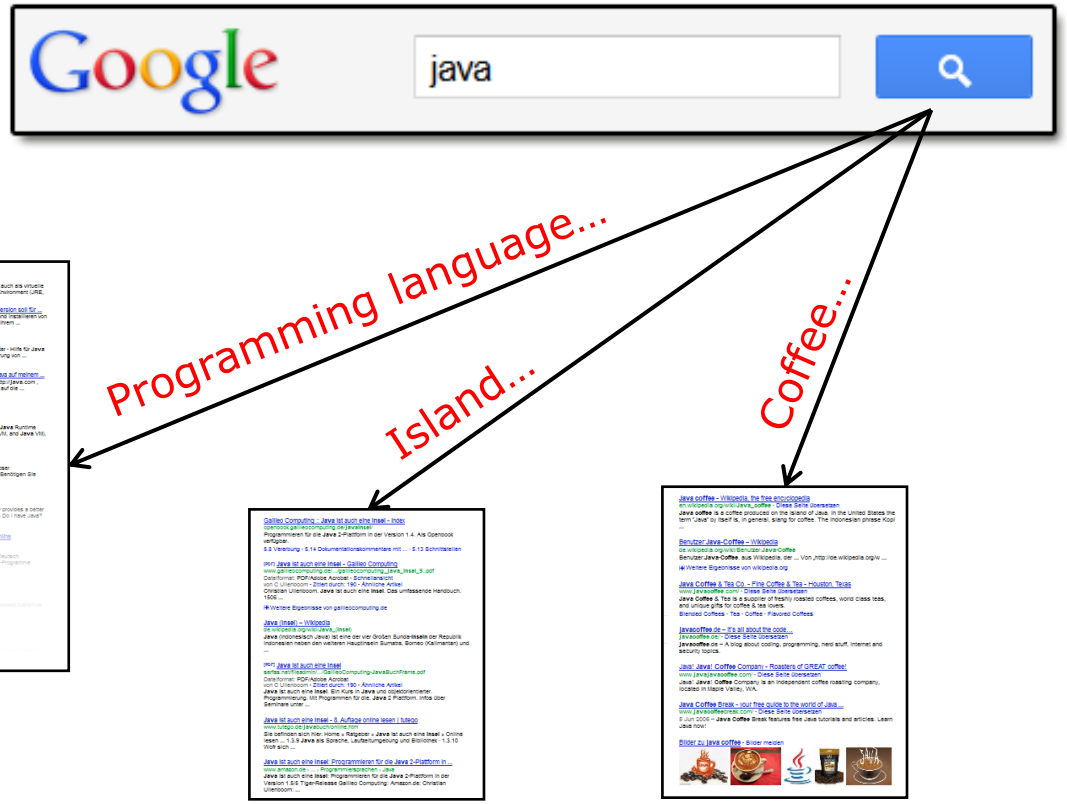
1: from sister
2: from cousin
3: from grandpa
4: from uncle Jim
5: from Mr. X

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…



Text Mining - Methods Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

[4]

Terrorist report	
INCIDENT TYPE	<>
DATE	<>
LOCATION	<>
PERPETRATOR	<>
PHYSICAL TARGET	<>
HUMAN TARGET	<>
EFFECT ON PHYSICAL TARGET	<>
EFFECT ON HUMAN TARGET	<>
INSTRUMENT	<>

- 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).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador: San Salvador (city)
PERPETRATOR	urban guerilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	–
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

Text Mining - Methods Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

[5]

olved in a Text Mining process…

- 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	…
…	…	…	…	…	…

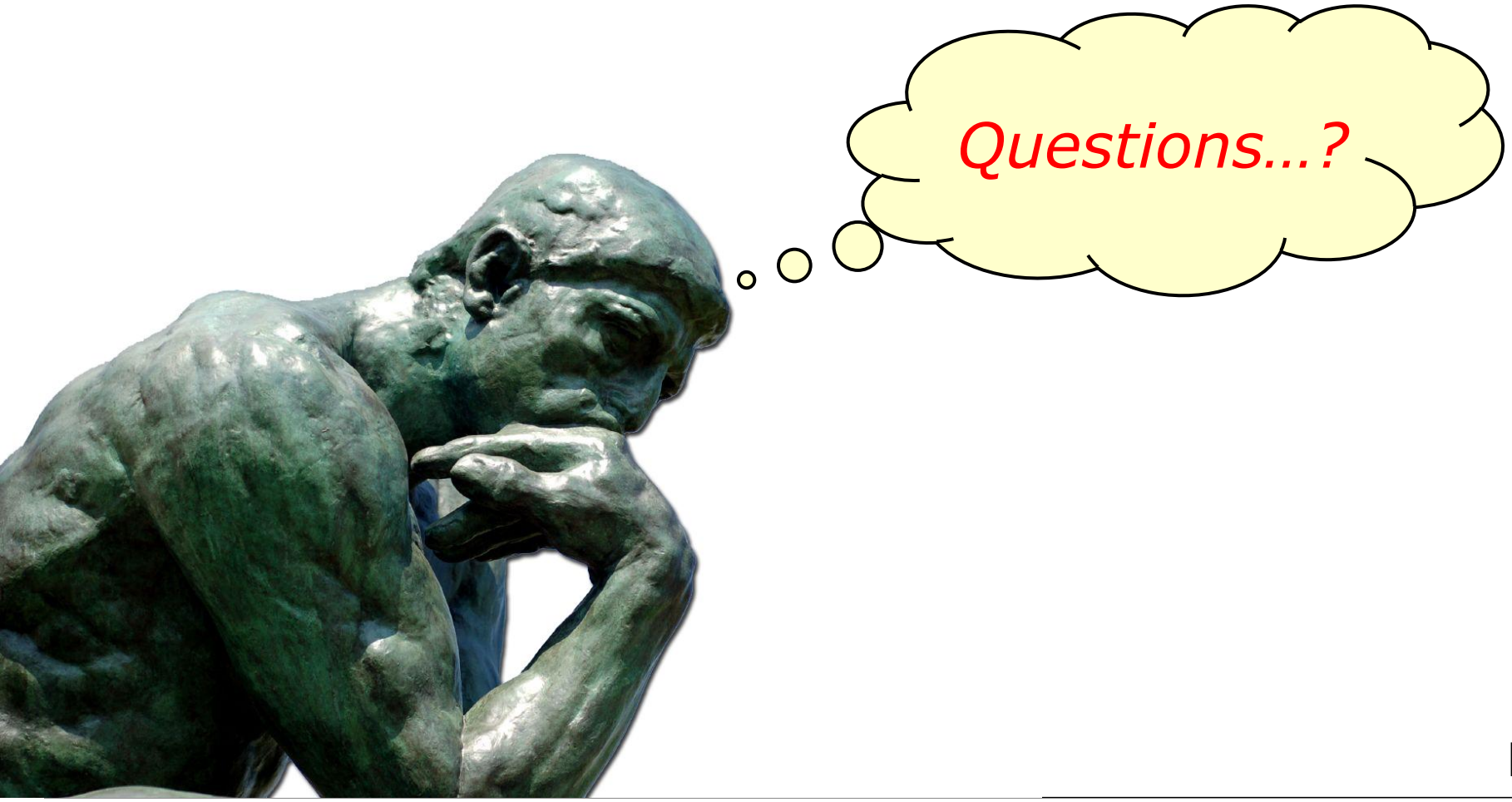
{Corn Flakes, Bread} → Yoghurt

Take-Home-Messages

TECHNISCHE UNIVERSITÄT DARMSTADT

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…



[6]

TECHNISCHE UNIVERSITÄT DARMSTADT

Thanks for your attention…

References

- [1] "**Eine Einführung In Text Mining – Wissensgewinnung aus Texten**",
  Seminar Slides by: Sebastian Blümel, CN8,
  Hochschule Furtwangen University, (unknown year – URL doesn't exist anymore !)

- [2] "**The Amount of Digital Information**",
  www.infoniac.com/hi-tech/amount-digital-information-reached-281-exabytes.html

- [3] "**Extracção de Conhecimento 2004/2005 (LEIC e MEI)**",
  Aulas Teóricas: Text Mining
  http://paginas.fe.up.pt/~ec/edicao04_05.html

References

- [4] "**Textmining – Information Extraction (symbolisch), Dept. Informatik 8 (Künstliche Intelligenz)** "

  http://www8.informatik.uni-erlangen.de/inf8/nocov/TM-slides/information-extraction.pdf

- [5] "**Unumkehrbarer Trend? - Künftig weitere Preissteigerungen**",

  http://www.n24.de/news/newsitem_771291.html

- [6] "**Questions picture**",

  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