JM2050 – Natural Language Processing

Introduction

September 5th, 2024

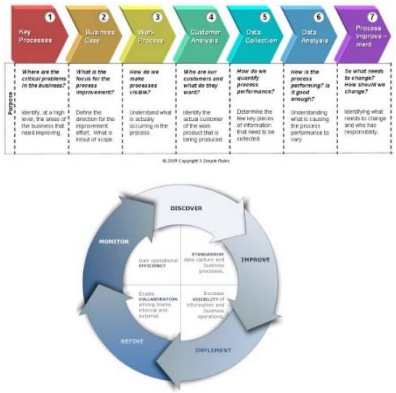
Prof.dr.ir. U. Kaymak

Outline

- Motivation
- Course organization
- Natural language processing – preliminaries
- Machine learning preliminaries

A lot of data out there….

Process improvement



Monitors and analyses events in an organization and proposes business improvement actions.

Smart power grids

Measures, monitors, and manages energy production, transport, and consumption is heterogeneous distributed grids.

Clinical decision support

Provides instant clinical decision support by correlating information from different part of uncorrelated sources.

www.jads.nl

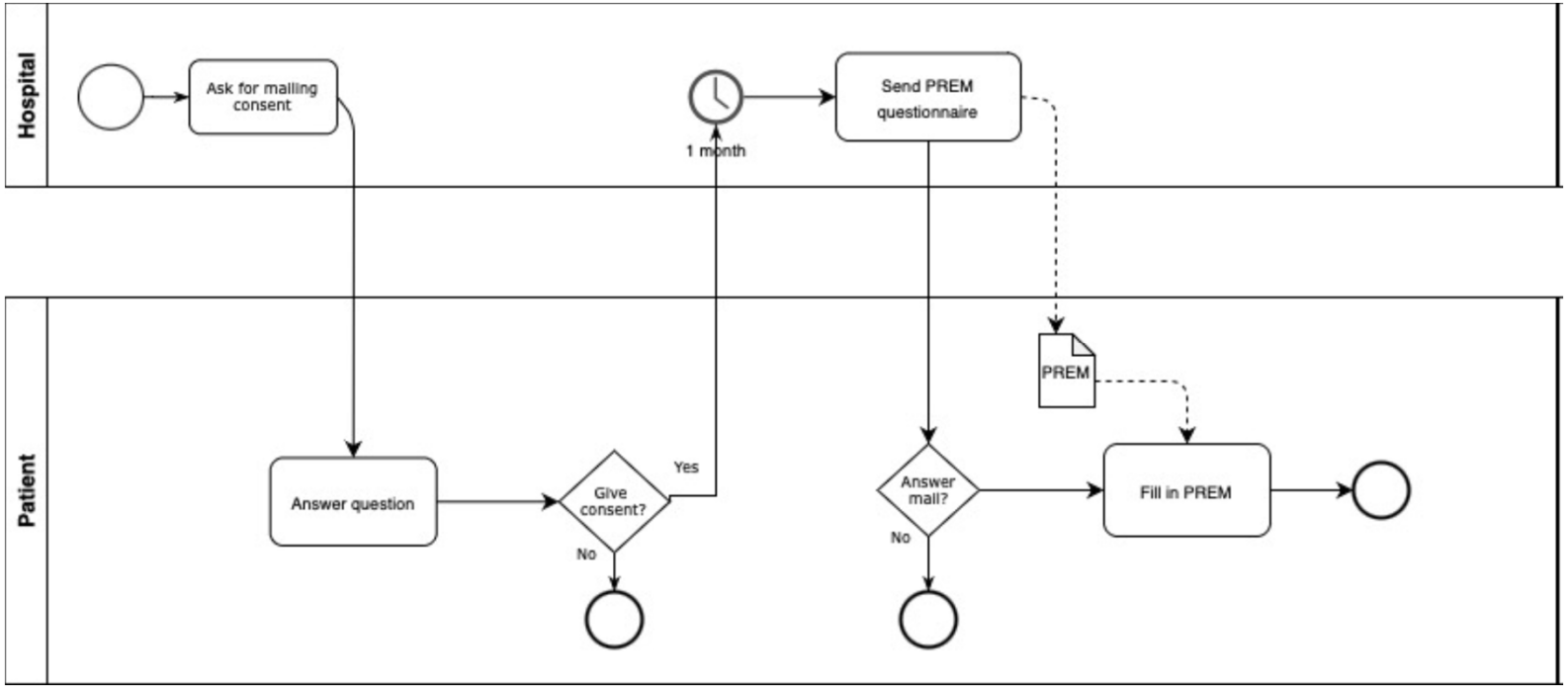
Jheronimus Academy of Data Science

Slide by DSC/e 3

Text as the prevailing medium in human communication

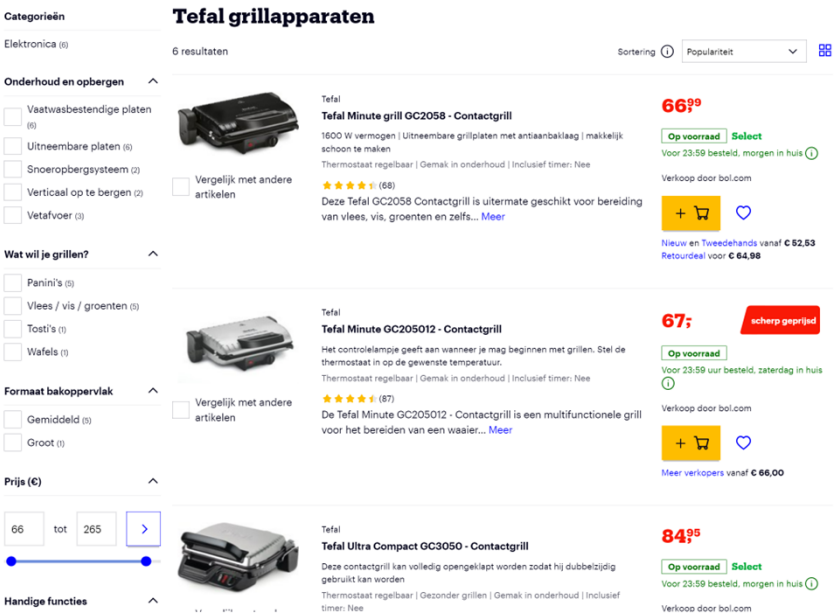
- electronic health records
- scientific publications
- quality reports & questionnaires
- online discussion fora
- social media
- e-mails…

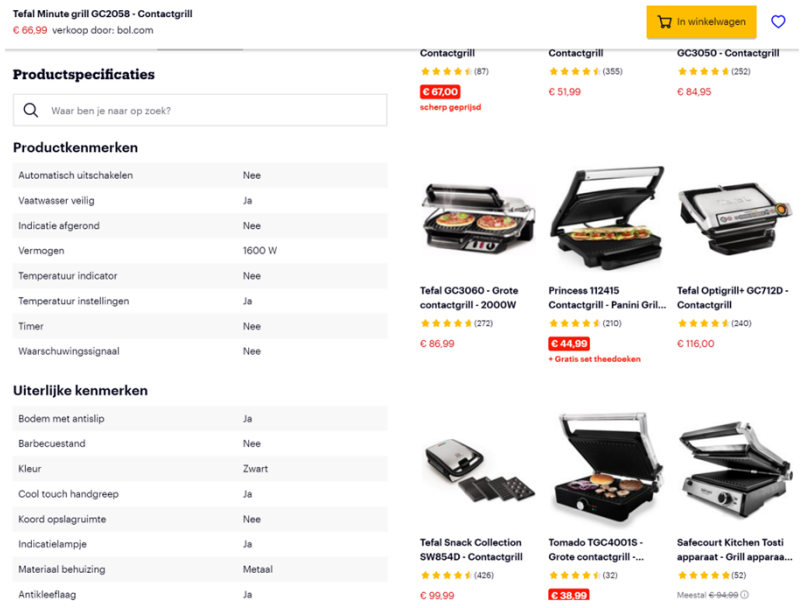
Client satisfaction



[5]

Populating web shop database for faceted search





[6]

www.jads.nl

Jheronimus Academy of Data Science

What can we do by analyzing text?

- Customer satisfaction analysis
- Violence prediction in mental healthcare
  - Analyze psychiatrists' and nurses' notes in EPR
- Populate web shop product characteristic database
- Sentiment analysis for a product
- Other popular examples:
  - Machine translation
  - Chatbots
  - Text summarization
  - Email filters

[7]

Overall focus

How can we make sense out of text data

and put it to use to provide solutions to real-world problems?

www.jads.nl

Jheronimus Academy of Data Science

Course organization

Jheronimus Academy of Data Science

9

Learning objectives

- relate NLP methods to other machine learning and deep learning methods;
- pre-process text for various NLP tasks;
- state the differences between classical NLP and deep learning NLP techniques;
- choose and apply an NLP technique that is appropriate for a business problem at hand;
- use insights from NLP models in prediction tasks.

Teaching team

- prof.dr.ir. Uzay Kaymak (coordinator), u.kaymak@jads.nl
- Erik Tromp, MSc. (lecturer), erik@understandling.com
- Niels Scholten, MSc. (teaching assistant), n.c.scholten@tue.nl

- Guest lecturer(s)

Meetings

- 14 sessions – 4 hrs. / week
  Typically: 1 x 2 hrs./week lecture, 1 x 2 hrs./week instruction, 14 weeks long)
  - Thursday 1 – 4, MDB 3.02

  Lectures: introduce and explain main concepts
  Instructions for practical exercises and working on assignments

- Q&A session at the end of the quartile
- Further questions can be asked through Canvas (preferred method) or after lectures

12

Planning – (partial, full program to be announced to Canvas)

Date	Time	Type	Room	Content
Thu Sep 5	09:30 – 12:30	Lecture / Instruction	MDB 0.10	Course introduction NLP overview ML basics (in context of NLP)
Thu Sep 12	09:30 – 12:30	Lecture / Instruction	MDB 0.10	Text mining Pre-processing text Instruction: software install
Thu Sep 19	09:30 – 12:30	Lecture / Instruction	MDB 0.10	Text representation NLP tasks I: sentiment classification, POS tagging Introduction to Assignment 1
Thu Sep 26	09:30 – 12:30	Lecture / Instruction	MDB 0.10	NLP tasks II: NER, dialogue systems, linguistic summarization Tutorial on Word Embeddings

Course material – 1

Mandatory literature

- Lecture slides
- Scientific papers and chapters
- Industry white papers

Software tools

- Python
- Various NLP packages for Python

Course material – 2

Recommended literature

- J. Eisenstein. *Introduction to Natural Language Processing*. MIT Press, 2019.
- D. Jurafsky and J.H. Martin. *Speech and Language Processing*. 3rd. ed. draft, 2024.
- A. Clark, C. Fox and S. Lappin. *The Handbook of Computational Linguistics and Natural Language Processing.* John Wiley & Sons, 2013. 10.1002/9781444324044.
- S. Bird, E. Klein and E. Loper. *Natural Language Processing with Python*. O'Reilly, 2009. (Available as e-book through the library of Eindhoven University of Technology)
- Self-study (video) tutorials.

Assessment

- Components:
  - Individual assignment – 20%
  - Group assignment 1 – 20%
  - Group assignment 2 – 20%
  - Written exam – 40%

- Group assignments will be made in teams of 3 students
  Register through Canvas as soon as possible

It is not possible to re-sit assignments
Assignments are valid only in the current academic year

17

Exam

- Type: written, closed book

- Date: 13 December 2024, 09.00 – 12.00
- Re-sit: 24 January 2025, 09.00 – 12.00

A minimal grade of 5.0 or higher is required for the exam to pass the course!

18

Policy on AI generated text

- Where authors use generative artificial intelligence (AI) and AI-assisted technologies in the writing process, authors should only use these technologies to improve readability and language. Applying the technology should be done with human oversight and control, and authors should carefully review and edit the result, as AI can generate authoritative-sounding output that can be incorrect, incomplete or biased. AI and AI-assisted technologies should not be listed as an author or co-author or be cited as an author. Authorship implies responsibilities and tasks that can only be attributed to and performed by humans [...].
- Authors must disclose the use of generative AI and AI-assisted technologies in the writing process by adding a statement at the end of their manuscript [...] . The statement should be placed in a new section entitled 'Declaration of Generative AI and AI-assisted technologies in the writing process'.
- *Statement: During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.*

Natural Language Processing – preliminaries

JADS Jheronimus Academy of Data Science

NLP approaches (and evolution)

Rule-based (rationalism)

- Hand-crafted rules
- Symbolic manipulation
- Starts in 1950's

Statistical (empiricism)

- Data-driven (probabilistic or otherwise)
- Shallow machine learning
- Starts in 1990's

Massively parallel processing (deep learning)

- Representation learning
- Human-like performance
- Starts in 2010's

Deng & Liu (2018) [22]

Natural Language Processing (NLP)

How to acquire knowledge that is expressed in natural language?

Symbolic NLP

Given a collection of rules (e.g., a Chinese phrasebook, with questions and matching answers), the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it is confronted with.

Neural NLP

Extension of statistical methods with representation learning and application of deep neural networks, including transformers

Statistical NLP

Application of machine learning techniques to NLP.

23

Wikipedia: natural language processing

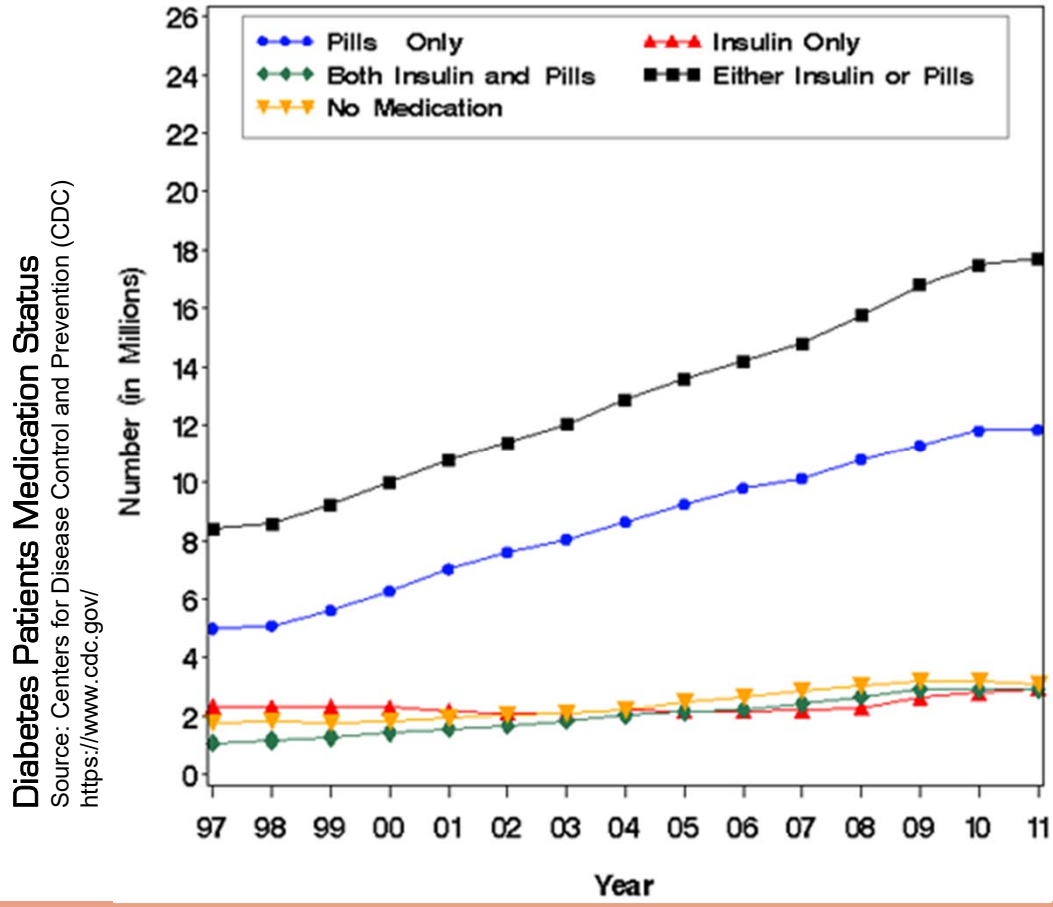
What do we see here?

1999	5.6	2.3	1.3	9.2	1.7
2000	6.3	2.3	1.4	10.0	1.8
2001	7.0	2.2	1.6	10.8	1.9
2002	7.6	2.1	1.7	11.4	2.0
2003	8.1	2.1	1.8	12.0	2.1
2004	8.6	2.2	2.0	12.9	2.2
2005	9.2	2.2	2.2	13.6	2.5
2006	9.8	2.2	2.2	14.2	2.6
2007	10.1	2.2	2.4	14.8	2.8
2008	10.8	2.3	2.7	15.7	3.0
2009	11.3	2.6	2.9	16.8	3.2
2010	11.8	2.8	2.9	17.5	3.2
2011	11.8	2.9	2.9	17.7	3.1

Diabetes Patients Medication Status

Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

Year	Pills Only	Insulin Only	Insulin and Pills	Any Medication	No Medication
1999	5.6	2.3	1.3	9.2	1.7
2000	6.3	2.3	1.4	10.0	1.8
2001	7.0	2.2	1.6	10.8	1.9
2002	7.6	2.1	1.7	11.4	2.0
2003	8.1	2.1	1.8	12.0	2.1
2004	8.6	2.2	2.0	12.9	2.2
2005	9.2	2.2	2.2	13.6	2.5
2006	9.8	2.2	2.2	14.2	2.6
2007	10.1	2.2	2.4	14.8	2.8
2008	10.8	2.3	2.7	15.7	3.0
2009	11.3	2.6	2.9	16.8	3.2
2010	11.8	2.8	2.9	17.5	3.2
2011	11.8	2.9	2.9	17.7	3.1


Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/
Pills Only
Insulin Only
Both Insulin and Pills
Either Insulin or Pills
No Medication
Number (in Millions)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
97 98 99 00 01 02 03 04 05 06 07 08 09 10 11
Year

What is special about text?

“ … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …”

Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

Diabetes Patients Medication Status

Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

Can we spot patterns?

Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

Diabetes Patients Medication Status

Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

Diabetes Patients Medication Status

Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

Structured, Unstructured, Semi-structured data

- Machines are good with ***structured data***
  - labeled data in (relational) databases
- We communicate information with language
  - speech & texts
- Our texts are typically ***unstructured data***
  - free-text

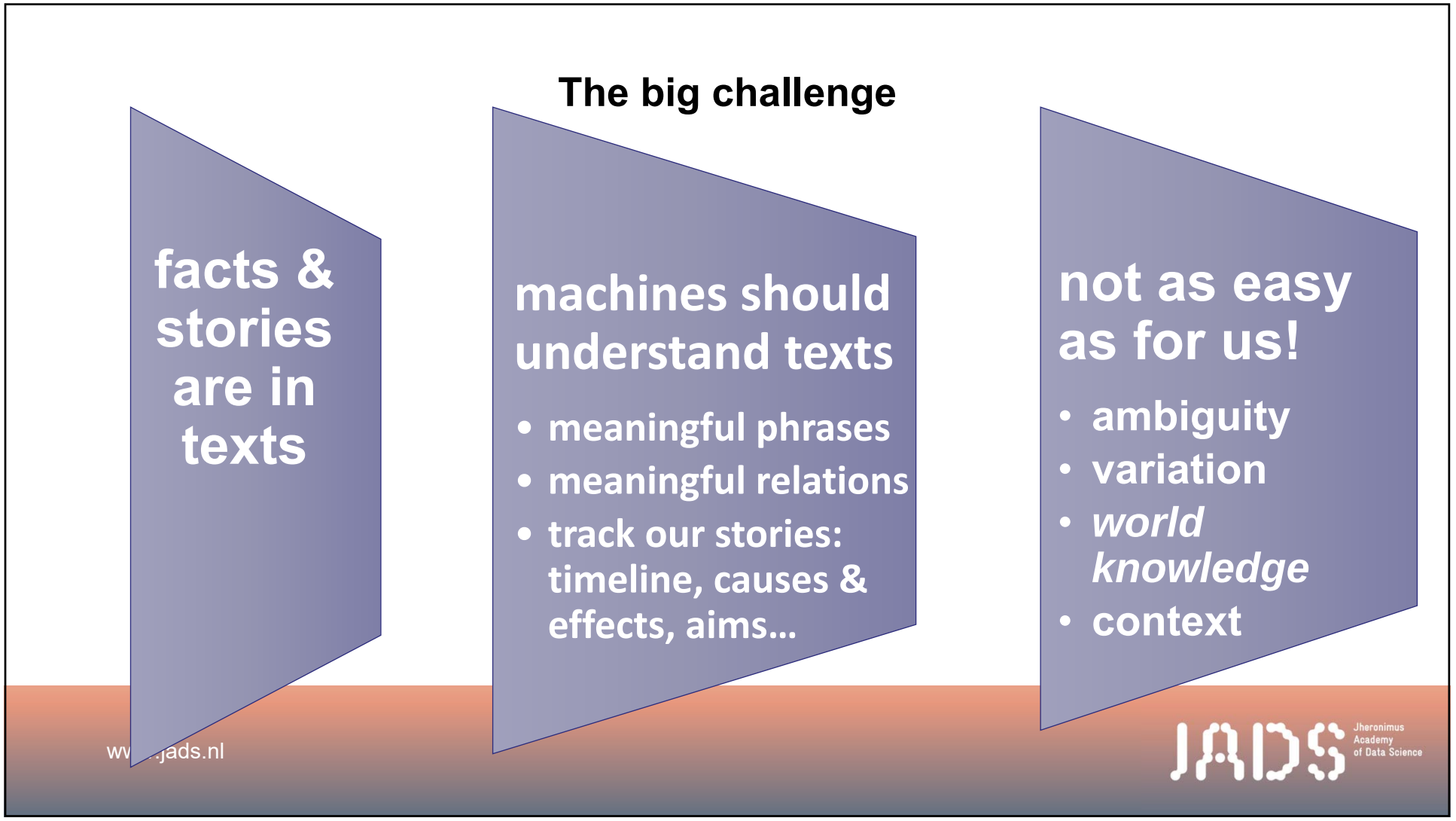
Year	Pills Only	Insulin Only	Insulin and Pills	Any Medication	No Medication
1999	5.6	2.3	1.3	9.2	1.7
2000	6.3	2.3	1.4	10.0	1.8
2001	7.0	2.2	1.6	10.8	1.9
2002	7.6	2.1	1.7	11.4	2.0
2003	8.1	2.1	1.8	12.0	2.1
2004	8.6	2.2	2.0	12.9	2.2
2005	9.2	2.2	2.2	13.6	2.5
2006	9.8	2.2	2.2	14.2	2.6
2007	10.1	2.2	2.4	14.8	2.8

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

Semi-structured data

- Mixture of structured and unstructured data
- E.g. in forms & databases: free-text notes
  - (i.e., ***semi-structured data***)

Year	Pills	Notes
1999	5.6	Adults report following a good diet
2000	6.3	Data collected in small regions. Other regions: check and combine information from table created later than 2010.
2001	7.0	Sample seems to contain more women. Adult age range: 18 – 50, no data for adults older than 50.


The big challenge
facts & stories are in texts
machines should understand texts
• meaningful phrases
• meaningful relations
• track our stories: timeline, causes & effects, aims…
not as easy as for us!
• ambiguity
• variation
• world knowledge
• context
www.jads.nl
JADS
Jheronimus Academy of Data Science

Ambiguity: *table*

Year	Pills Only	Insulin Only	Insulin and Pills	Any Medication	No Medication
1999	5.6	2.3	1.3	9.2	1.7
2000	6.3	2.3	1.4	10.0	1.8
2001	7.0	2.2	1.6	10.8	1.9
2002	7.6	2.1	1.7	11.4	2.0
2003	8.1	2.1	1.8	12.0	2.1
2004	8.6	2.2	2.0	12.9	2.2
2005	9.2	2.2	2.2	13.6	2.5
2006	9.8	2.2	2.2	14.2	2.6
2007	10.1	2.2	2.4	14.8	2.8

Ambiguity…

Los Angeles Times

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH

Big rig carrying fruit crashes on 210 Freeway, creates jam

McDonald's Fries the Holy Grail for Potato Farmers

Wednesday, September 23, 2(
Associated Press

FOX
TOP

DutchNews.nl

TUESDAY 27 MARCH 2012

Home | Opinion | Features | International | In Dutch | Dictionary | What's On |

PERFECT HOUSING
Perfect solutions
Stay with us

Chinese cooking fat heads for Holland

Tuesday 27 March 2012

A first consignment of 20 tonnes of processed waste cooking oil was shipped from the Chinese port of Qingdao en route to the Netherlands on Monday, where it will be used to power KLM planes, Chinese news agency Xinhua reported.

www.jads.nl

JADS Jheronimus Academy of Data Science

Variation

Dialect variation	
	
vacation	holiday
cash point	ATM
lift	elevator
film	movie
queue	line

Spelling variation
kilometers per hour
kilometres per hour
klometer pr hotr
km per hour
km hr
km/h

Synonyms & syntactic variants
Jill's house
Jill's home
Jill's residence
the house of Jill

Diabetes Patients Medication Status

Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

" … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …"

www.jads.nl

Domain, Knowledge & Text context to the rescue

- **Domain:** document context, style, genre, purpose, characteristics
- **Knowledge:** general and domain ***knowledge resources***
- **Text context:** use of ***linguistic information***

www.jads.nl

Jheronimus Academy of Data Science

What does analyzing text mean?

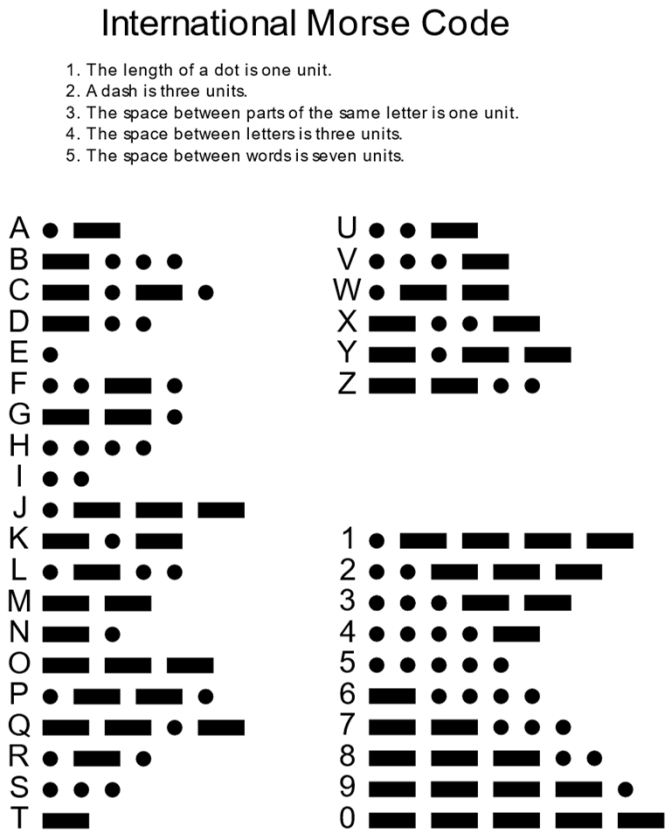
[40]

How shall we analyse this?

Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/

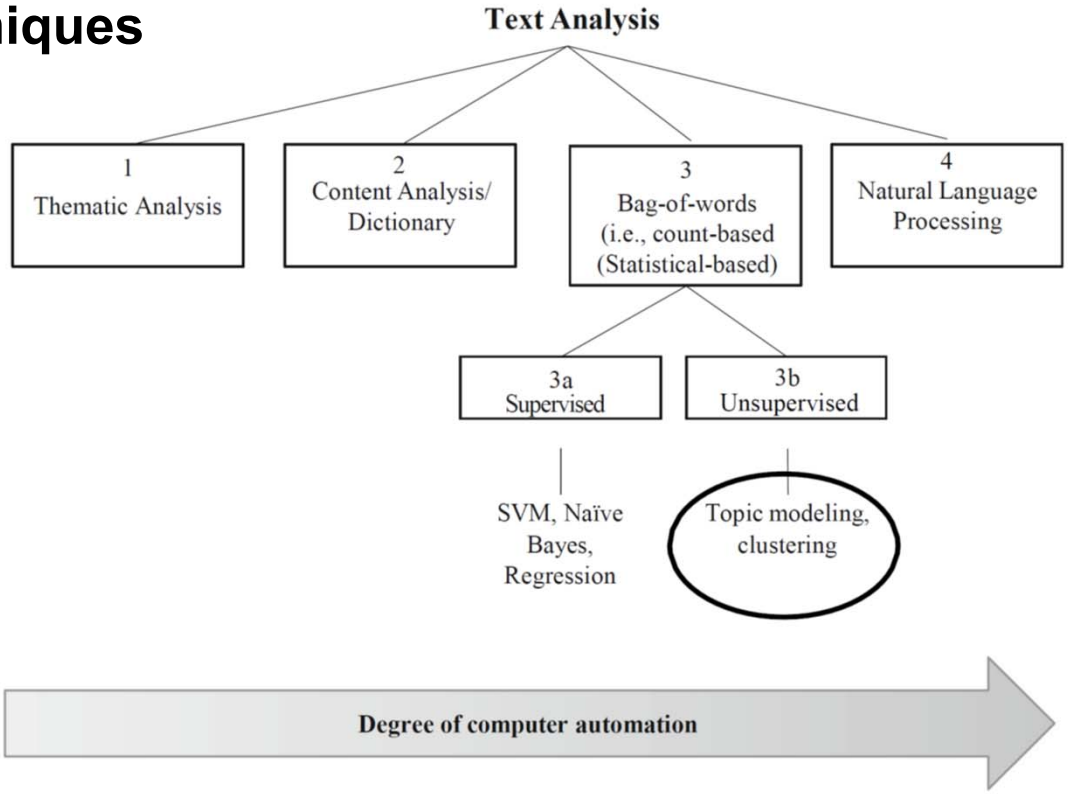
“ … From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until 2007 and increased afterwards. …”

Language analysis



[42]

Text analysis techniques



(Banks et al., 2018)

[43]

Some common natural language processing tasks

Text classification	Information retrieval	Information extraction
spam filtering	recommender systems	Template-filling
topic modeling	search engine	named entity recognition (NER)
sentiment analysis	question answering	relationship extraction
	Summarization	ontology extraction

44

Recap

- Three general approaches to NLP
  - Rule-based (rationalism)
  - Statistical (ML) models (empiricism)
  - Deep learning models (massive parallelism)
- Text is considered unstructured data
- Challenges with text data
  - Ambiguity
  - Variation
  - World knowledge
  - Context
- NLP tasks become easier by using limiting the domain, using knowledge resources and context information

[46]