Natural Language Processing

Prof. Dr. Jannik Strötgen
jannik.stroetgen@h-ka.de

Summer 2024

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Who is Paul?

Paul Löhr

- MA Student Wirtschaftsinformatik
- Participant of the NLP lecture 2023
- Grade: <not shown on the slide>

→ Our TA for the NLP class, in particular the person for various things related to the Tutorial!

Who am I?

Prof. Dr. Jannik Strötgen

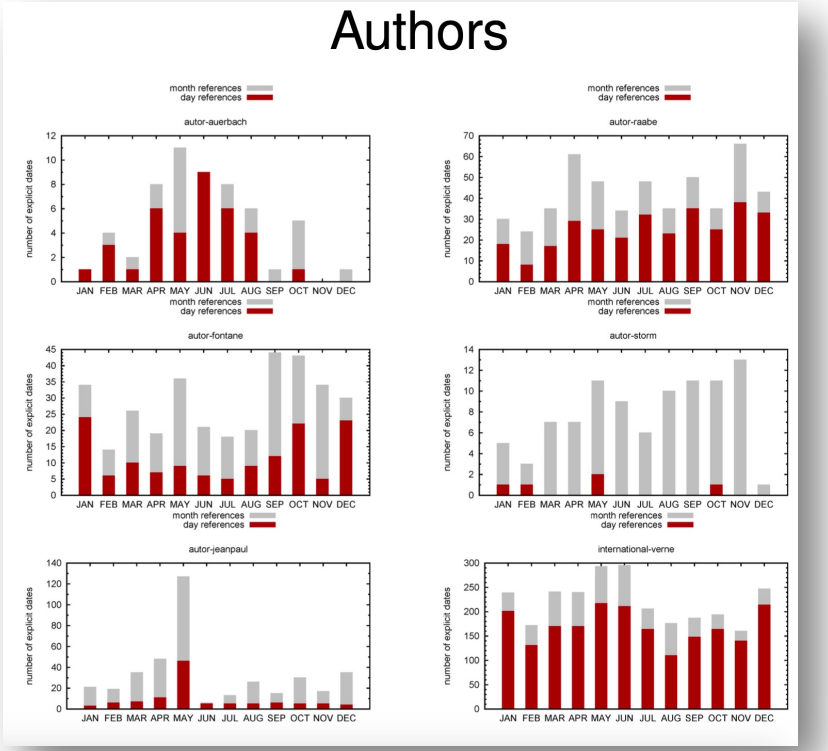
credit: bosch-ai.com

Vita

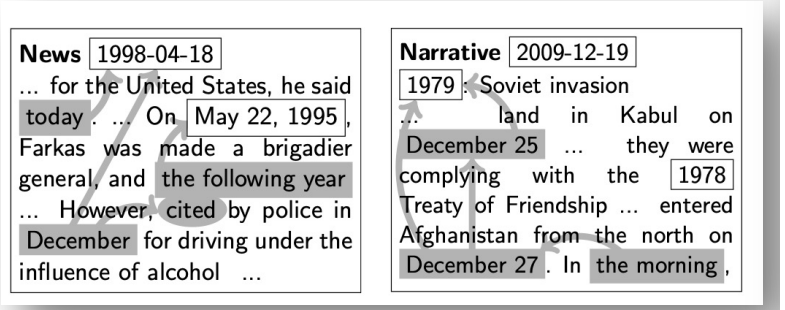
- since 03/2023: professor @ Karlsruhe University of Applied Sciences
- Bosch Center for AI: Research Scientist, Research Group Head (2018 – 2023)
- Max Planck Institute for Informatics: Postdoc, Group Leader (2015 – 2018)
- Heidelberg University: PhD Student, Research Assistant, Postdoc (2009 – 2015)
- PhD (Computer Science) @ Heidelberg University (2015)
- Studies (Computational Linguistics & Economics) @ Heidelberg University (-2009)

Research Interests

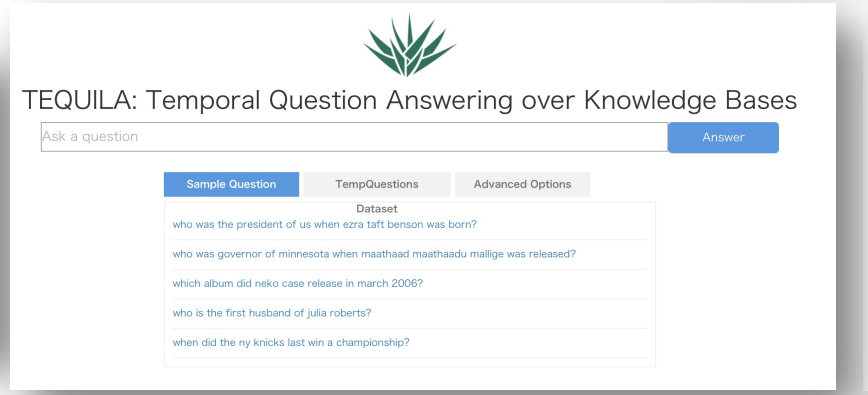
- Natural Language Processing
  - Information Extraction
- Knowledge Graphs
- Information Retrieval
- In general, everything related to “time”



Temporal data analysis



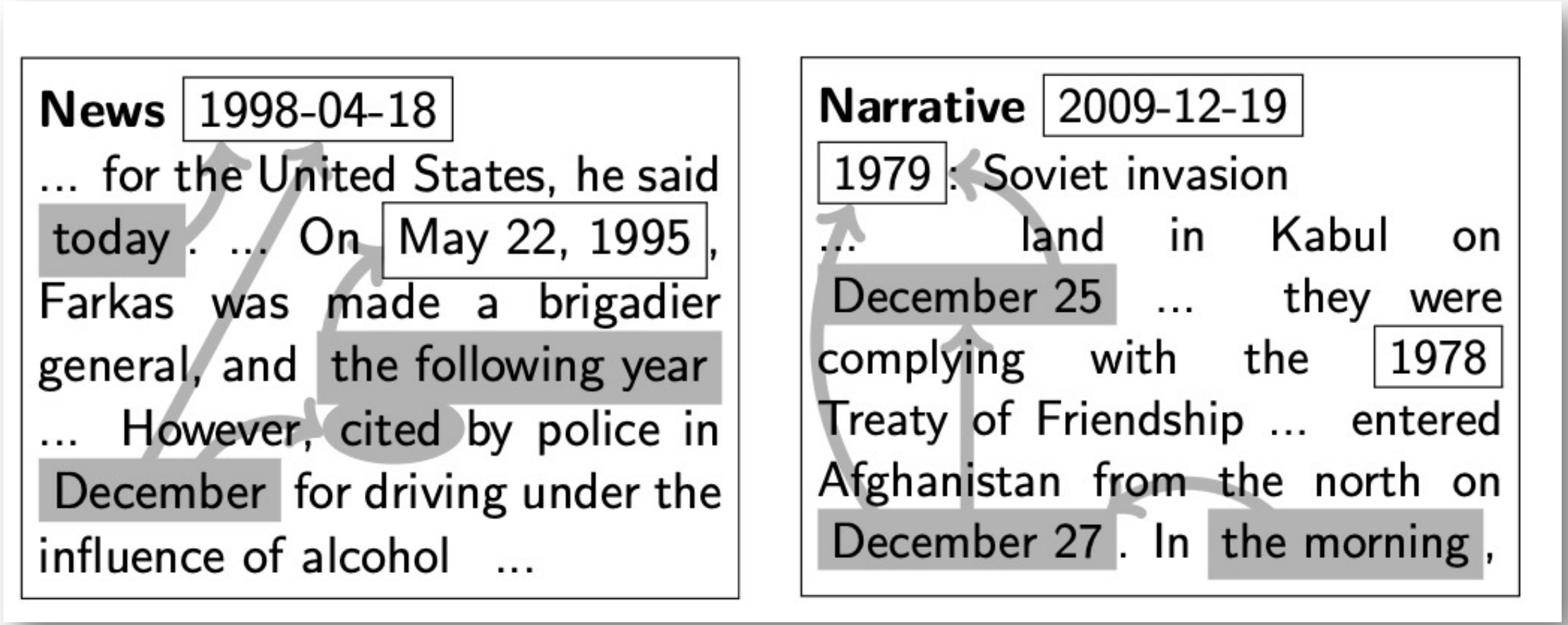
Temporal information extraction (temporal tagging)



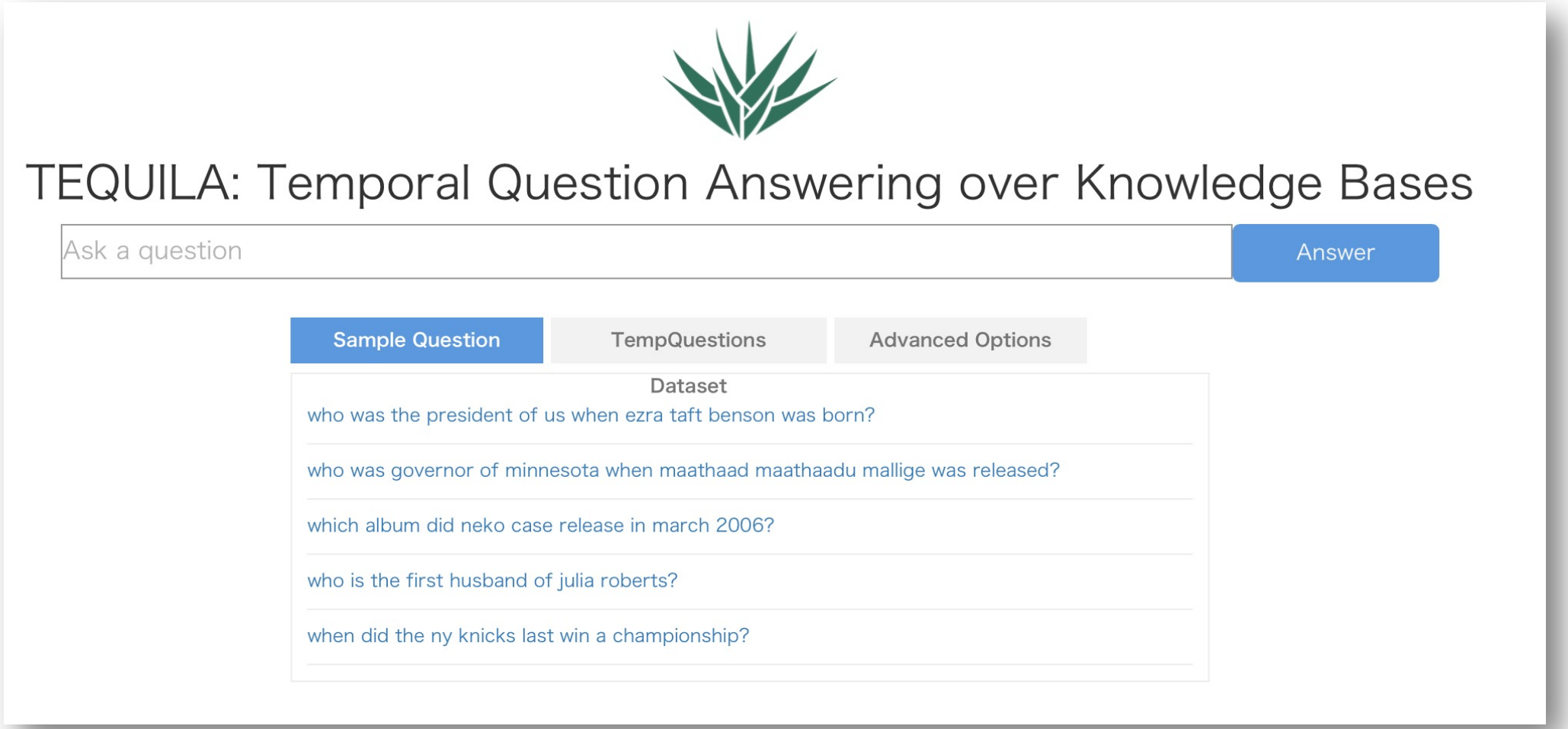
Temporal question answering

Temporal information retrieval

Temporal Information Extraction (Temporal Tagging)

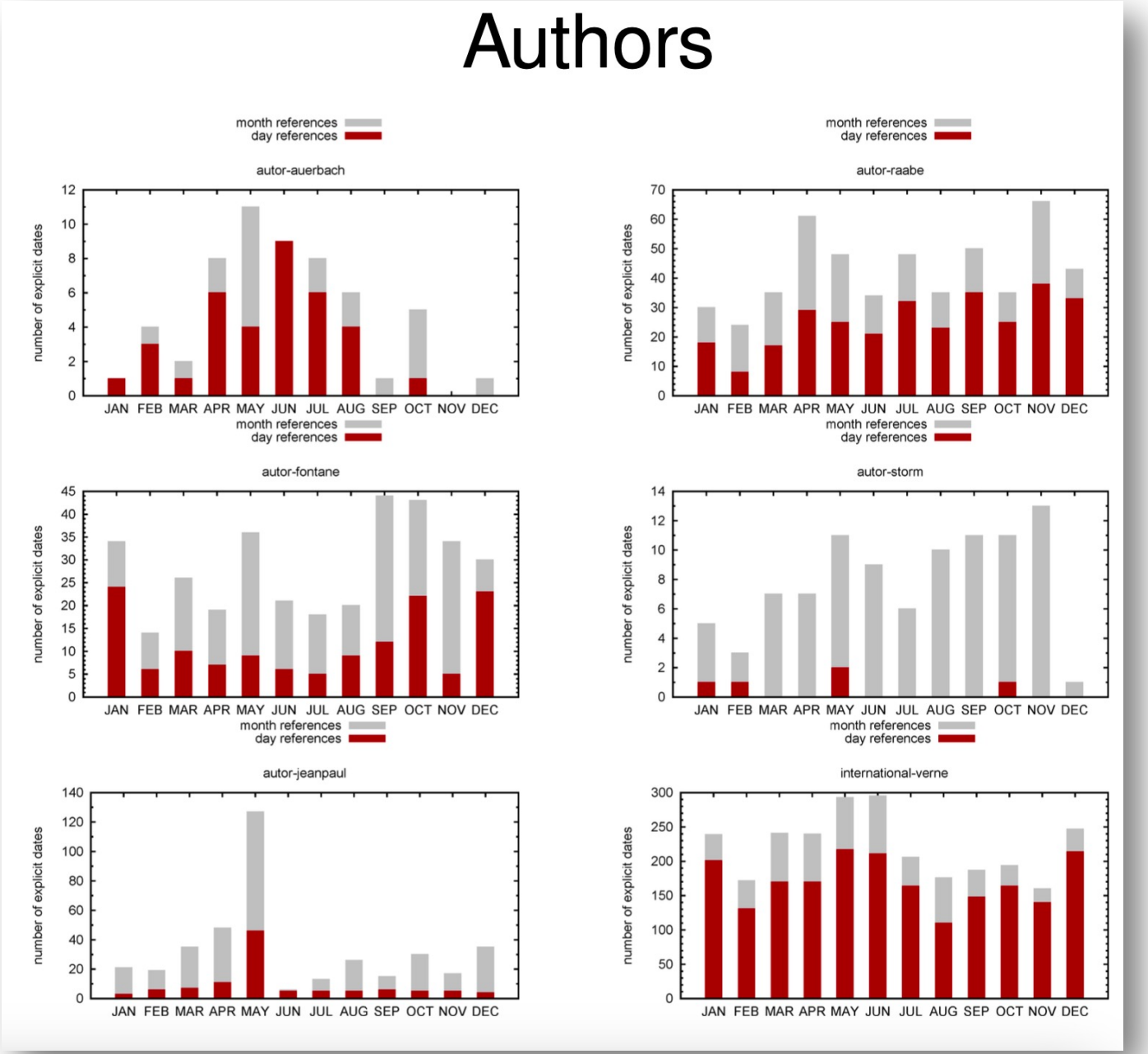


Temporal Question Answering



Temporal Information Retrieval

Temporal Data Analysis

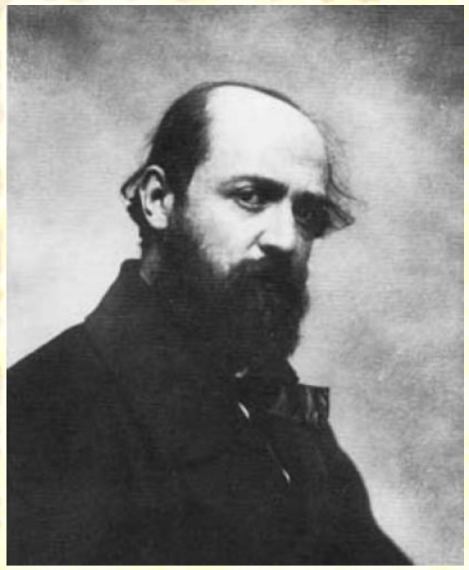


Today in World Literature

« Previous

Quote of the Day (19th of March)

Next »



Henri Murger
Image Attribution: Unknown (Public domain)

Henri Murger: La Vie de Bohème (Source)

It was **the nineteenth of March**, 184—. Should Rodolphe reach the age of Methuselah, he will never forget the date; for it was on that day, at three in the afternoon, that our friend issued from a banker's where he had just received five hundred francs in current and sounding specie.

Continue reading »

Quotes for date: **#1**

Home About Tiwoli Random Quote

https://dh.phil-fak.uni-koeln.de/forschung/tiwoli

Research Interests

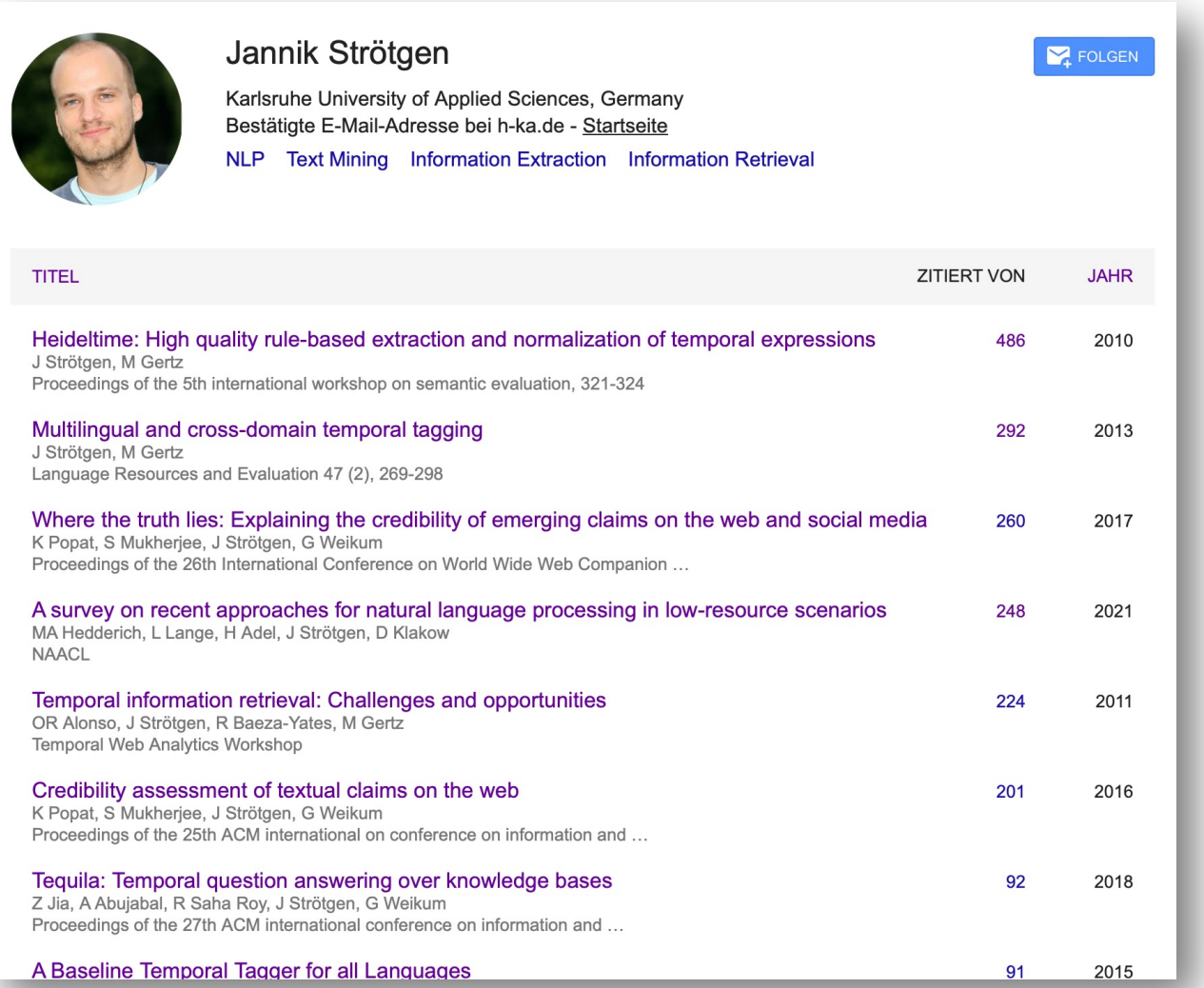
Antrittsvorlesung

- April 17, 2024, 11:30 am
- room Li he

Zeit für Data Science und Natural Language Processing in Zeiten von ChatGPT

(Time for Data Science and Natural Language Processing in Times of ChatGPT)

Research Interests



Jannik Strötgen

FOLGEN

Karlsruhe University of Applied Sciences, Germany
Bestätigte E-Mail-Adresse bei h-ka.de - Startseite

NLP Text Mining Information Extraction Information Retrieval

TITEL	ZITIERT VON	JAHR
Heideltime: High quality rule-based extraction and normalization of temporal expressions J Strötgen, M Gertz Proceedings of the 5th international workshop on semantic evaluation, 321-324	486	2010
Multilingual and cross-domain temporal tagging J Strötgen, M Gertz Language Resources and Evaluation 47 (2), 269-298	292	2013
Where the truth lies: Explaining the credibility of emerging claims on the web and social media K Popat, S Mukherjee, J Strötgen, G Weikum Proceedings of the 26th International Conference on World Wide Web Companion …	260	2017
A survey on recent approaches for natural language processing in low-resource scenarios MA Hedderich, L Lange, H Adel, J Strötgen, D Klakow NAACL	248	2021
Temporal information retrieval: Challenges and opportunities OR Alonso, J Strötgen, R Baeza-Yates, M Gertz Temporal Web Analytics Workshop	224	2011
Credibility assessment of textual claims on the web K Popat, S Mukherjee, J Strötgen, G Weikum Proceedings of the 25th ACM international on conference on information and …	201	2016
Tequila: Temporal question answering over knowledge bases Z Jia, A Abujabal, R Saha Roy, J Strötgen, G Weikum Proceedings of the 27th ACM international conference on information and …	92	2018
A Baseline Temporal Tagger for all Languages	91	2015

Research Interests

Jannik Strötgen

FOLGEN

Karlsruhe University of Applied Sciences, Germany
Bestätigte E-Mail-Adresse bei h-ka.de - Startseite

NLP Text Mining Information Extraction Information Retrieval

TITEL	ZITIERT VON	JAHR
Device and method for processing temporal expressions from unstructured texts for filling a knowledge database L Lange, J Stroetgen, H Adel-Vu US Patent App. 18/305,896		2023
GradSim: Gradient-Based Language Grouping for Effective Multilingual Training M Wang, H Adel, L Lange, J Strötgen, H Schütze arXiv preprint arXiv:2310.15269		2023
Method for predicting a persistence over time of entries of a knowledge base S Razniewski, I Dikeoulias, J Stroetgen US Patent 11,783,202		2023
Device and method for training a model for linking a mention to an entity across knowledge bases H Soliman, D Milchevski, H Adel-Vu, M Gad-Elrab, J Stroetgen US Patent App. 18/178,373		2023
Device and computer implemented method for adding a quantity fact to a knowledge base D Stepanova, D Milchevski, G Weikum, J Stroetgen, VT Ho US Patent App. 18/168,666		2023
Computer-implemented method and device for processing data H Adel-Vu, J Stroetgen, L Lange US Patent 11,687,725		2023
Tada: Efficient task-agnostic domain adaptation for transformers CC Hung, L Lange, J Strötgen arXiv preprint arXiv:2305.12717	2	2023

Recently accepted:

- ***Rehearsal-Free Modular and Compositional Continual Learning for Language Models***.
  Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, Hinrich Schütze. NAACL'24.
- ***Discourse-Aware In-Context-Learning for Temporal Expression Normalization***.
  Akash Kumar Gautam, Lukas Lange, Jannik Strötgen. NAACL'24.

Teaching

In the past:

- Heidelberg University (mainly seminars)
- Saarland University (seminars, lectures, e.g., *Information Retrieval and Data Mining*)
- Karlsruhe University of Applied Sciences (lectures, e.g., *Data Engineering*)

Teaching

Since summer 2023 (Karlsruhe University of Applied Sciences):

- *Wissenschaftliches Arbeiten (MA Wirtschaftsinformatik)*
- *Natural Language Processing (MA Wirtschaftsinformatik, BA Data Science)*
- *Artificial Intelligence (MA Informatik)*
- *Explainable AI (MA Informatik)*
- *Planung von Informationssystemen (BA Wirtschaftsinformatik)*
- *Anwendungsprojekte (BA Wirtschaftsinformatik, Data Science)*
- *Datenbanken und Datenkunde 2 (BA Data Science)*
- *Data Engineering (BA Data Science)*

Teaching

This semester (Karlsruhe University of Applied Sciences):

- *Wissenschaftliches Arbeiten (MA Wirtschaftsinformatik)*
- ***Natural Language Processing (MA Wirtschaftsinformatik, BA Data Science)***
- *Artificial Intelligence (MA Informatik)*
- *Explainable AI (MA Informatik)*
- *Planung von Informationssystemen (BA Wirtschaftsinformatik)*
- *Anwendungsprojekte (BA Wirtschaftsinformatik, Data Science)*
- *Datenbanken und Datenkunde 2 (BA Data Science)*

My Approach to Teaching

Not me!

Not you!

- I will gladly serve you good food for thought
- Always feel free to ask about the ingredients!
- Always feel free to ask for more!
- … but you need to chew yourself.

Communication 101

I am a big fan of **open, honest, and direct communication**. I am not a fan of excuses and BS. In general, please

- be civil
- be respectful
- be open-minded
- be constructive
- and remember that it is ok to be wrong sometimes.

It is ok to have a problem or to need help from time to time.

- Please let me know asap.
- My door is always open.
- I don‘t hold grudges.

Disclaimer: Don‘t blindly trust me (professors)

Important note:

- I am still fairly new at HKA. I do not necessarily do everything in the way to which you are used (if there even is such a thing). If any organizational detail is a problem for you, please let me know!

Thus:

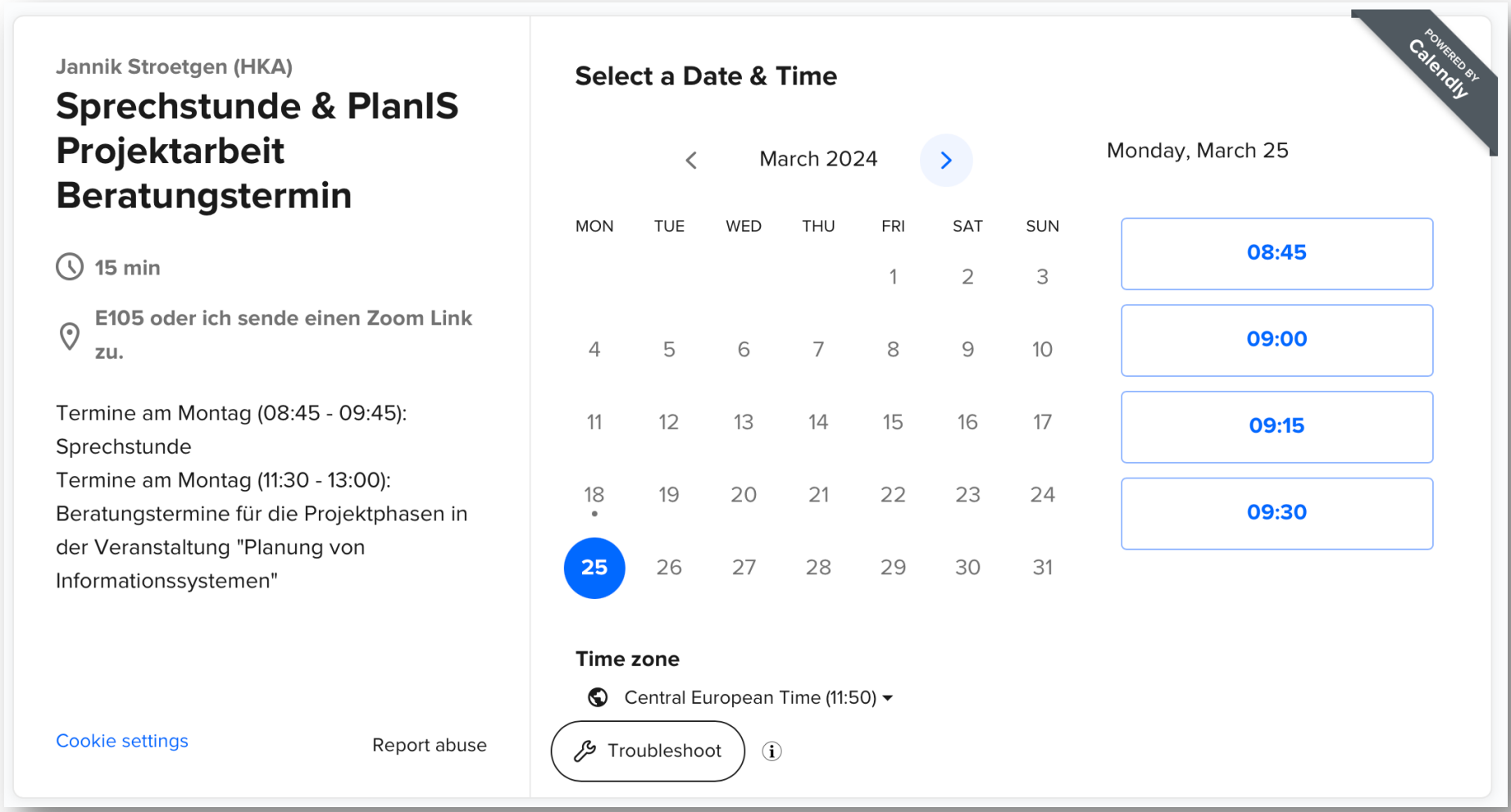
- If something that I say about the schedule or deadlines conflicts with something you know, do not simply assume that I know what I am talking about! If in doubt, please ask and/or correct me!

Contact

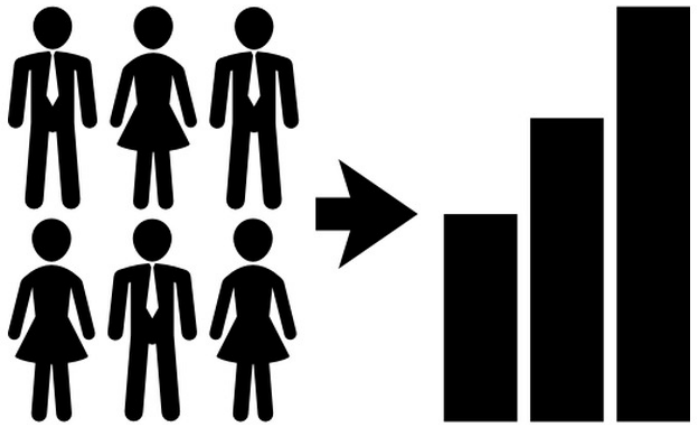
Prof. Dr. Jannik Strötgen

- Room: E105
- jannik.stroetgen@h-ka.de
- Office hours: Monday, 08:45 am – 09:45 am
  - Please make an appointment via email
    or book a time slot via Calendly:
    https://calendly.com/jannik-stroetgen-hka/sprechstunde-planis

Contact


Jannik Stroetgen (HKA)
Sprechstunde & PlanIS Projektarbeit Beratungstermin
15 min
E105 oder ich sende einen Zoom Link zu.
Termine am Montag (08:45 - 09:45): Sprechstunde
Termine am Montag (11:30 - 13:00): Beratungstermine für die Projektphasen in der Veranstaltung "Planung von Informationssystemen"
Cookie settings
Report abuse
Select a Date & Time
POWERED BY Calendly
March 2024
Monday, March 25
MON TUE WED THU FRI SAT SUN
1 2 3
4 5 6 7 8 9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30 31
08:45
09:00
09:15
09:30
Time zone
Central European Time (11:50)
Troubleshoot

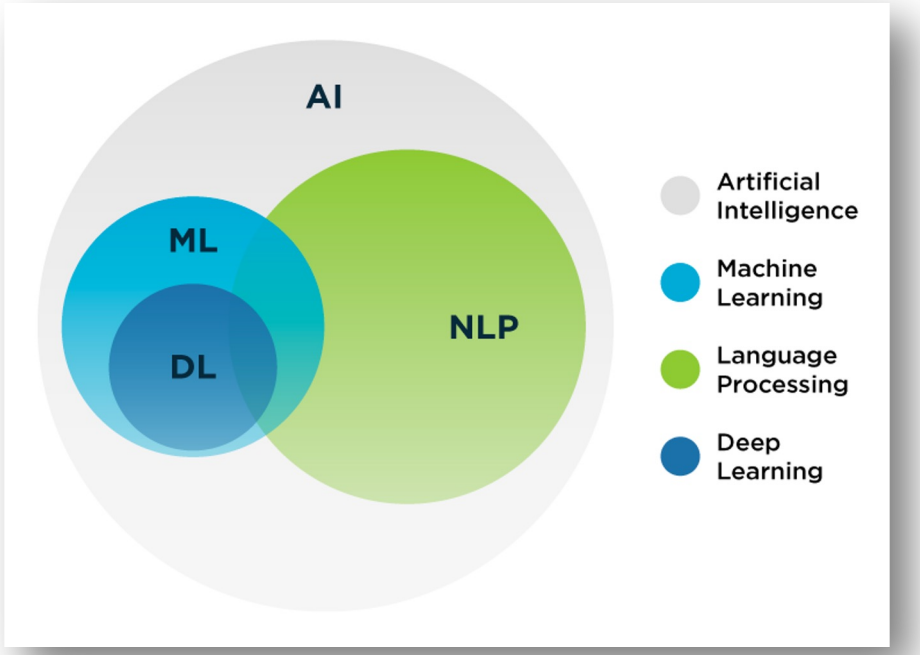
Who are you?



- What do you study?
- In which semester?
- What related lectures did you attend?
- What are your expectations for this lecture?

Motivation: What is NLP? Why is it hard?

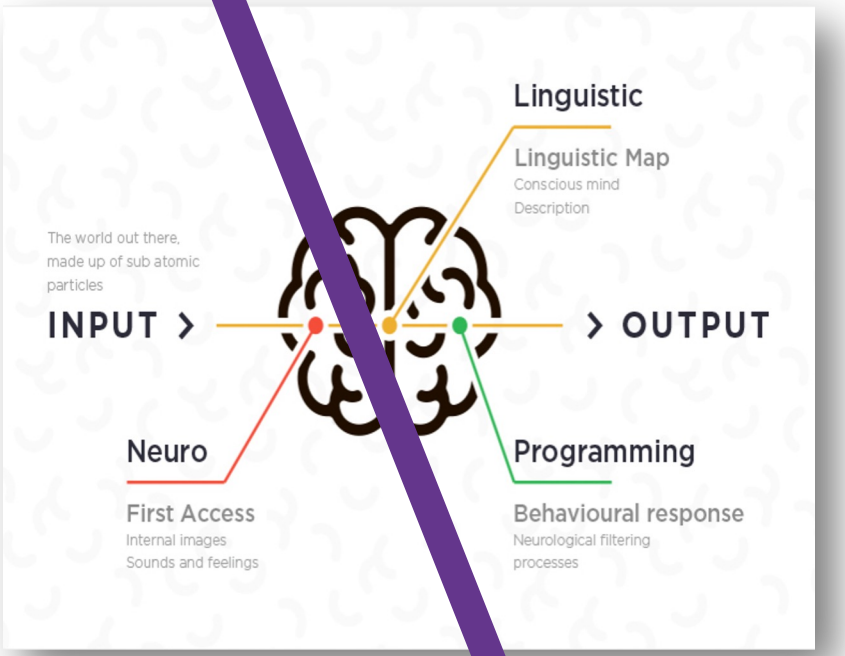
What is NLP?



https://7wdata.be/article-general/natural-language-processing-taking-your-business-to-the-next-level/

Natural Language Processing

Hot Topic in Machine Learning und AI!



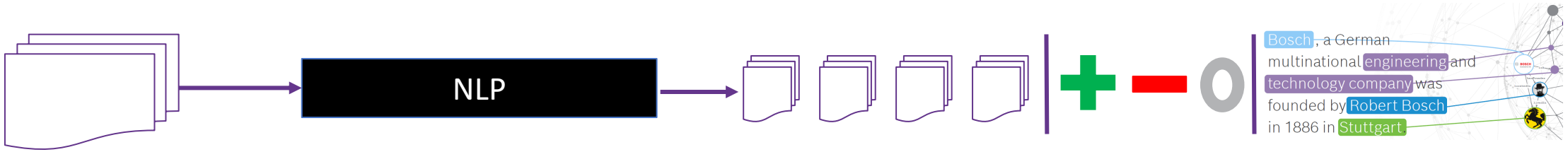
https://www.nlpacademy.co.u[illegible]what_is_nlp/

Neuro-linguistic Programming

Pseudo-research

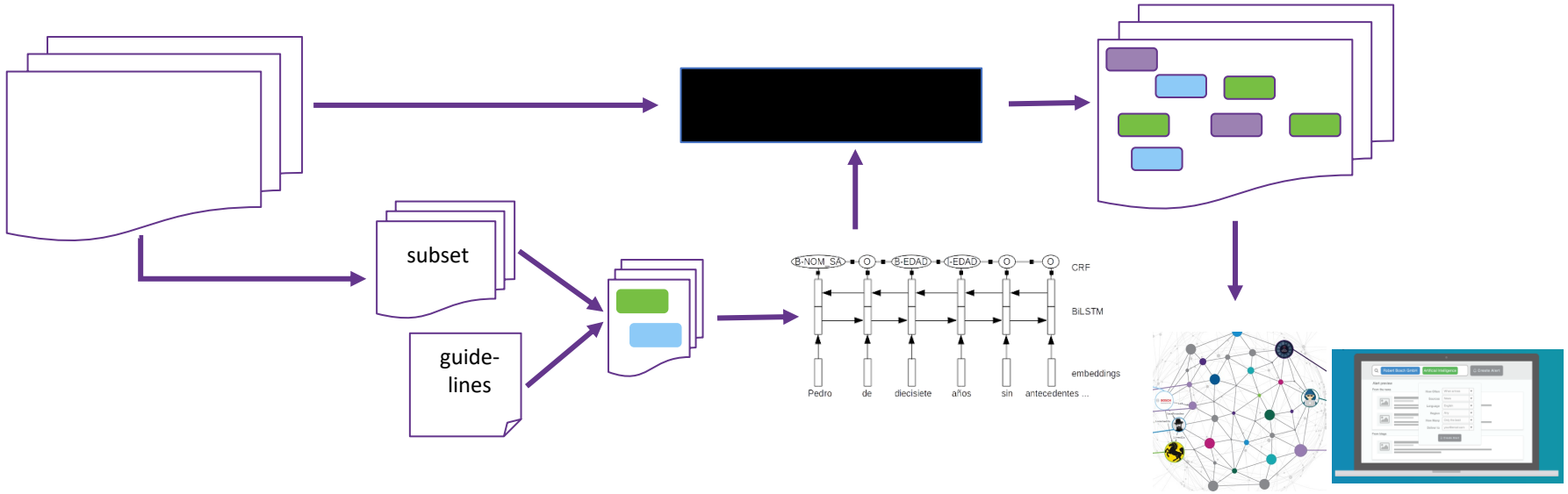
What is NLP?

- Processing of natural language
- Many heterogenous applications: clustering, classification, information extraction, chatbots, search engines, question answering systems, speech recognition, …



What is NLP? Typical Setup

- Typical Setup (nowadays)

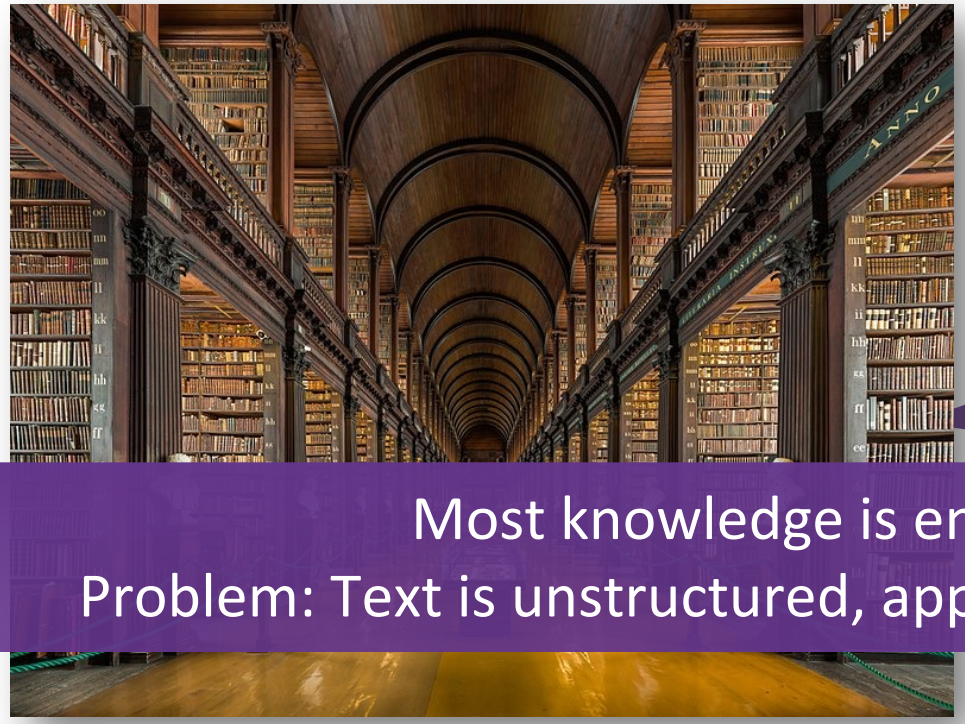


Most Knowledge Encoded in Natural Language

https://de.m.wikipedia.org/wiki/Datei:Long_Room_Interior,_Trinity_College_Dublin,_Ireland_-_Diliff.jpg

Information Extraction

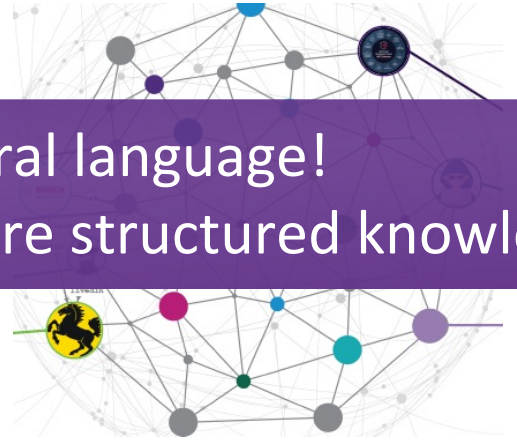
Unstructured knowledge in natural (human) language



https://de.m.wikipedia.org/wiki/Datei:Long_Room_Interior,_Trinity_College_Dublin,_Ireland_-_Diliff.jpg

Summary for humans

Machine-readable summary



Most knowledge is encoded in natural language!
Problem: Text is unstructured, applications require structured knowledge!

Knowledge Graph

Robert Bosch

German industrialist

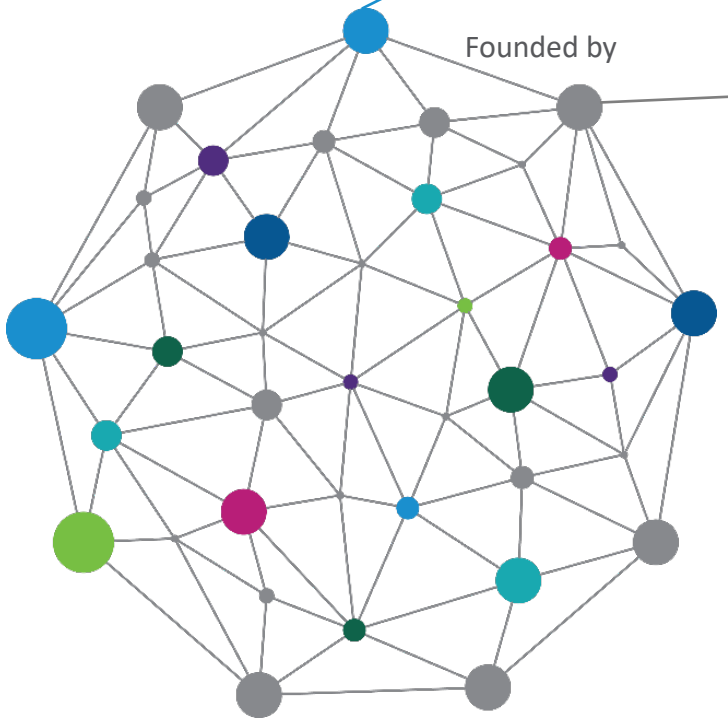
Robert Bosch was a German industrialist, engineer and inventor, founder of Robert Bosch GmbH.

Founded by

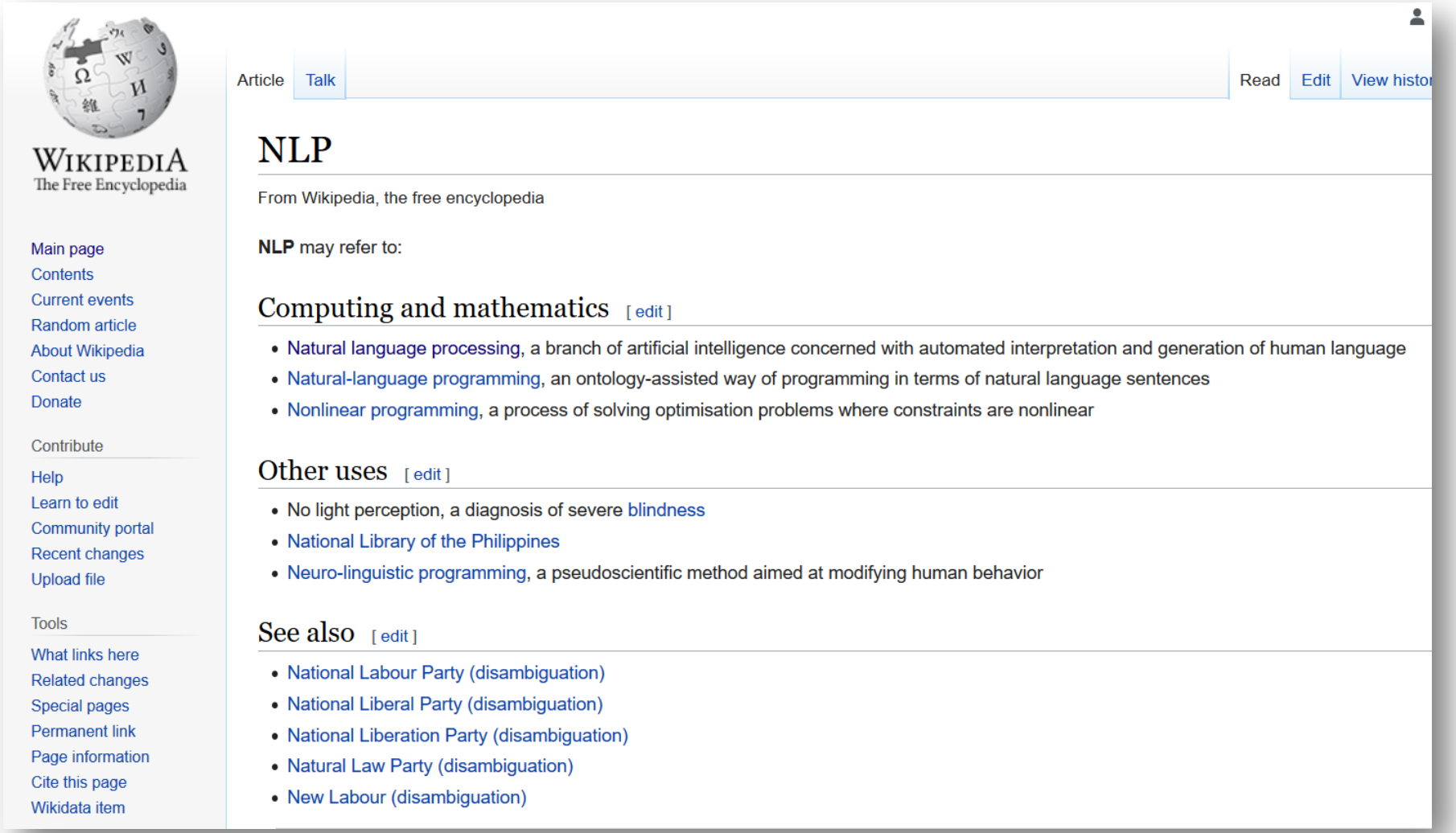
Robert Bosch GmbH

Engineering company

Robert Bosch GmbH, or Bosch, is a German multinational engineering and technology company headquartered in Gerlingen.



Why NLP is hard?



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Article Talk Read Edit View histo

NLP

From Wikipedia, the free encyclopedia

NLP may refer to:

Computing and mathematics [edit]

- Natural language processing, a branch of artificial intelligence concerned with automated interpretation and generation of human language
- Natural-language programming, an ontology-assisted way of programming in terms of natural language sentences
- Nonlinear programming, a process of solving optimisation problems where constraints are nonlinear

Other uses [edit]

- No light perception, a diagnosis of severe blindness
- National Library of the Philippines
- Neuro-linguistic programming, a pseudoscientific method aimed at modifying human behavior

See also [edit]

- National Labour Party (disambiguation)
- National Liberal Party (disambiguation)
- National Liberation Party (disambiguation)
- Natural Law Party (disambiguation)
- New Labour (disambiguation)

Resolving **ambiguities** is almost always the main challenge in NLP

(**context** helps!)

Why NLP is hard?

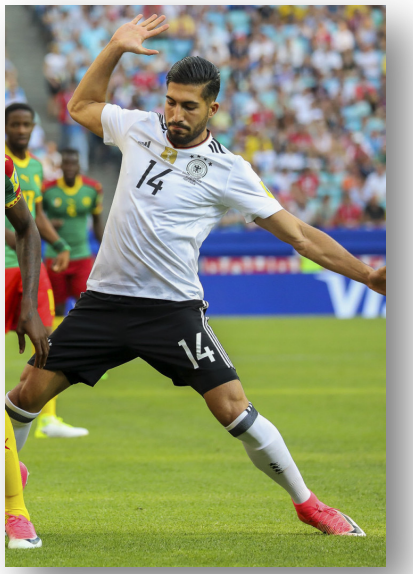
Resolving ambiguities is almost always the main challenge in NLP! (**context** helps!)

- Ambiguities exist on all levels (word, sentence, syntactic, semantic, phonological)
- Example: which part of speech is the English word “can”
- Can we can fish in a can?

→auxiliary, verb, noun

- Can we can fish in a can, Mr. Can?

→ auxiliary, verb, noun, proper noun

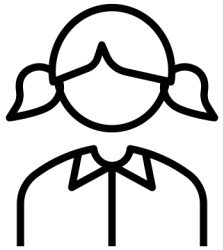


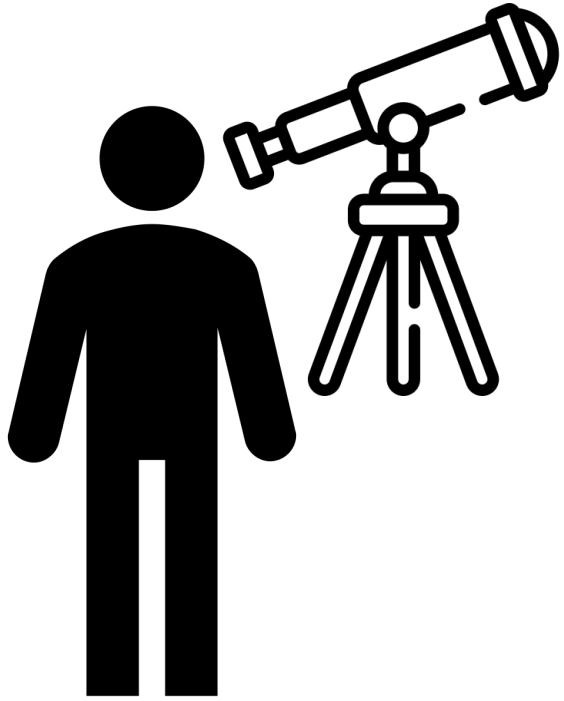
Surname [edit]

- Cem Can (born 1981), Turkish footballer
- Cihan Can (born 1986), Turkish footballer
- Derya Can Göçen, Turkish world record holder female free-diver
- Emre Can (born 1994), German footballer of Turkish origin
- Emre Can (chess player) (born 1990), Turkish Grand Master chess player
- Erkan Can (born 1958), Turkish film and theatre actor
- Eyüp Can (boxer) (born 1964), Turkish boxer
- Eyüp Can (journalist) (born 1973), Turkish journalist
- Melisa Can (born 1984), U.S.-born Turkish female basketball player
- Mustafa Can (born 1969), Swedish author and journalist of Kurdish origin
- Müslüm Can (born 1975), German footballer of Turkish origin
- Osman Can, Turkish jurist
- Sibel Can (born 1970), Turkish folk pop singer
- Şenol Can (born 1983), Turkish footballer
- Yasemin Can (born 1996), Turkish female long-distance runner of Kenyan origin

https://de.wikipedia.org/wiki/Emre_Can#/media/Datei:Germany_VS._Cameroon_(3)_(cropped).jpg

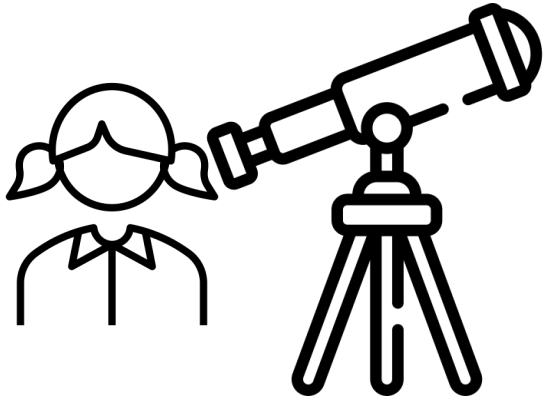
Ambiguity: Syntactic Ambiguity

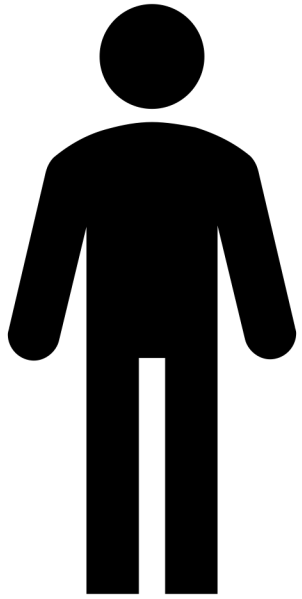




The girl saw the man with the telescope.

Ambiguity: Syntactic Ambiguity





The girl saw the man with the telescope.

Ambiguity: Anaphora Resolution

We gave the monkeys bananas because **they** were hungry.

vs.

We gave the monkeys bananas because **they** were over-ripe.

Ambiguity: Interpretation & World Knowledge

Every American has a president.

vs.

Every American has a mother.

Resolving these Problems Computationally

Consecutive evolution of three major types of approaches

- Rule-based models (until the 1990s)
- Statistical models / corpus linguistics (until the 2010s)
- (Deep) neural network models (ongoing)

What we will cover in the lecture

- Introduction to NLP and Text Analytics
- Preprocessing (e.g., Tokenization, Stemming)
- Linguistic concepts (e.g., syntax, semantics),
- Linguistic preprocessing (tagging, parsing),
- Semantic processing (named entity recognition and linking),
- Language models (e.g., word embeddings)
- Text analytics applications (e.g., classification, clustering, information extraction)

Organization

Natural Language Processing @ HKA SoSe’24

Wirtschaftsinformatik Master

- „Kernfach“ → 4 SWS; 5 ECTS

Data Science Bachelor (6. Semester)

- „Wahlpflichtfach“ → 4 SWS, 5 ECTS

Dual Degree Wirtschaftsinformatik Master

- „Kernfach“ → 4 SWS; 5 ECTS

Natural Language Processing @ HKA SoSe’24

When and Where?

- Tuesday, Block 4 (14:00 – 15:30), E102 (lecture)
- Wednesday, Block 2 (09:50 – 11:20), E002 (tutorial)

Credits

- 4 SWS / 5 ECTS

Lecture slides, exercise sheets, further material and announcements

- ILIAS
- mattermost (maybe)

Prüfungsleistung: Exam + X

Exam

- written exam (90 minutes)

X (pre-requisite to participate in the exam and to gain 5% or 10% bonus points!)

- Presentation and uploaded solutions for one of the exercise sheets
- Depending on number of participants: other types of X

Date

- Exam: some date between July 8 and July 26, 2024.

Agenda

Exam:
some date
between
July 8
and
July 26, 2024.

	Date	Lecture (slides)		Date	Tutorial (exercises)
lecture 1	March 19, 2024	Organization & Introduction to NLP	tutorial 1	March 20, 2024	
lecture 2	March 26, 2024		tutorial 2	March 27, 2024	
	April 2, 2024	*no lecture (Easter)*	tutorial 3	April 3, 2024	
lecture 3	April 9, 2024		tutorial 4	April 10, 2024	
lecture 4	April 16, 2024		tutorial 5	April 17, 2024	
lecture 5	April 23, 2024		tutorial 6	April 24, 2024	
lecture 6	April 30, 2024			May 1, 2024	*no tutorial (holiday)*
lecture 7	May 7, 2024		tutorial 7	May 8, 2024	
lecture 8	May 14, 2024		tutorial 8	May 15, 2024	
	May 21, 2024	*no lecture (Whitsun break)*		May 22, 2024	*no lecture (Whitsun break)*
lecture 9	May 28, 2024		tuorial 9	May 29, 2024	
lecture 10	June 4, 2024		tutorial 10	June 5, 2024	
lecture 11	June 11, 2024		tutorial 11	June 12, 2024	
lecture 12	June 18, 2024		tutorial 12	June 19, 2024	
lecture 13	June 25, 2024		tutorial 13	June 26, 2024	
lecture 14	July 2, 2024		tutorial 14	July 3, 2024	

Preliminary Agenda

Date	Topic
19.03.2024	Organisation & motivation
26.03.2024	Introduction to NLP and
02.04.2024	*no lecture (Easter)*
09.04.2024	Pre-Processing and Part-of-Speech Tagging
16.04.2024	Parsing
23.04.2024	Named Entity Recognition and Disambiguation
30.04.2024	Similarity and Search
07.05.2024	Language Models: Static Word Embeddings

Date	Topic
14.05.2024	Contextual Embeddings
21.05.2024	*no lecture (Whitsun break)*
28.05.2024	Text Mining and Sentiment Analysis
04.06.2024	Information Extraction & QA
11.06.2024	Applications exploiting NLP
18.06.2024	NLP with LLMs
25.06.2024	My Research Topics
02.07.2024	Recap, exam preparation

Sources

Main source

- Material of my former colleague J-Prof. Dr. Andreas Spitz (Uni Konstanz)

In addition:

- Chris Manning & Richard Socher's Natural Language Processing with DL.
- Dan Jurafsky & James H. Martin: Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/

Next Steps

- Please join the ILIAS course:
  https://ilias.h-ka.de/goto.php?target=crs_853687&client_id=HSKA

Password:
NLP2024NLP

Thank you! Questions?