

# Natural Language Processing

Prof. Dr. Jannik Strötgen  
[jannik.stroetgen@h-ka.de](mailto:jannik.stroetgen@h-ka.de)

Summer 2024

**Hochschule Karlsruhe**  
University of  
Applied Sciences

Fakultät für  
**Informatik und**  
**Wirtschaftsinformatik**



# Next Steps

- Please join the ILIAS course:

[https://ilias.h-ka.de/goto.php?target=crs\\_853687&client\\_id=HSKA](https://ilias.h-ka.de/goto.php?target=crs_853687&client_id=HSKA)

**Password:**  
**NLP2024NLP**



Thank you! Questions?

# Motivation and Introduction

# “Unstructured” Data

## Unstructured data

The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

## Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

# Semi-structured Data

## Text (literary theory)

Article [Talk](#)

From Wikipedia, the free encyclopedia

In [literary theory](#), a **text** is any object that can be "read", whether this object is a work of literature, a street sign, an arrangement of buildings on a city block, or styles of clothing. It is a coherent set of [signs](#) that transmits some kind of informative [message](#).

[\[1\]](#) This set of signs is considered in terms of the informative message's *content*, rather than in terms of its physical form or the medium in which it is represented.

Within the field of [literary criticism](#), "text" also refers to the original information content of a particular piece of writing; that is, the "text" of a work is that primal symbolic arrangement of letters as originally composed, apart from later alterations, deterioration, commentary, translations, [paratext](#), etc. Therefore, when literary criticism is concerned with the determination of a "text", it is concerned with the distinguishing of the original information content from whatever has been added to or subtracted from that content as it appears in a given textual document (that is, a physical representation of text).

Since the [history of writing](#) predates the concept of the "text", most texts were not written with this concept in mind. Most written works fall within a narrow range of the types described by [text theory](#). The concept of "text" becomes relevant if and when a "coherent written message is completed and needs to be referred to independently of the circumstances in which it was created."[\[citation needed\]](#)

# Applications of Natural Language Processing

## Information retrieval and extraction

- Search
- Sorting
- Knowledge extraction

## Marketing

- Opinion mining
- Customer feedback analysis
- Product recommendation

## Research

- Political and social science
- Humanities
- Meta-studies in Medicine

## Web applications

- Content moderation
- Writing assistants
- News aggregation

# Topics (preliminary)

# Preliminary Agenda

Date	Topic
19.03.2024	Organisation & motivation
26.03.2024	Introduction to NLP and
02.04.2024	<i>no lecture (Easter)</i>
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	<i>no lecture (Whitsun break)</i>
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

# Preliminary Topics

1. Pre-processing (Tokenization, Stemming and Lemmatization)
2. Part-of-speech Tagging
3. Parsing
4. Named Entity Recognition and Linking
5. Similarity and Search
6. Language Models: Static Word Embeddings
7. Contextual Language Models
8. Text Mining (Classification, Clustering, and Topic Models)
9. Opinion Mining and Sentiment Analysis
10. Information extraction and question answering
11. Applications exploiting NLP

# 1. Pre-processing

Ah distinctly I remember it was in the bleak December and each separate dying ember wrought its ghost upon the floor. Eagerly I wished the Morrow;—vainly I had sought to borrow from my books surcease of sorrow—sorrow for the lost Lenore—for the rare and radiant maiden whom the angels nameless here forevermore ntm adj: distinct  
adv: distinctly

Ah, distinctly I remember it was in the bleak December;  
And each separate dying ember wrought its ghost upon the floor.  
Eagerly I wished the Morrow;—vainly I had sought to borrow  
From my books surcease of sorrow—sorrow for the lost Lenore—  
For the rare and radiant maiden whom the angels nameless here forevermore.

to seek  
1. person sg.  
past perfect



## 2. Part-of-Speech (POS) Tagging

Adding information about word types helps in determining the roles that words play and in understanding the meaning of a text. For example, we can identify **verbs**, **adverbs**, **nouns**, **determiner (articles)**, etc.

Ah, distinctly I remember it was in the bleak December;  
And each separate dying ember wrought its ghost upon the floor.  
Eagerly I wished the morrow;—vainly I had sought to borrow  
From my books surcease of sorrow—sorrow for the lost Lenore—  
For the rare and radiant maiden whom the angels name Lenore—  
Nameless here for evermore.



### 3. Parsing

Parsing enables us to determine the **syntactic structure** of sentences.

The diagram shows a syntax tree for the first stanza of 'The Raven'. The root node is S, which branches into two S nodes. The left S node branches into NP (Ah, distinctly I remember) and VP (it was in the bleak December). The right S node branches into VP (ember wrought its ghost upon the floor). The VP node for the first sentence branches into PR (Bea) and VB (se). The VP node for the second sentence branches into DT (y) and NN (ember). The NP node for the first sentence branches into RB (Eagerly), PR (I), VB (wished), DT (the), and NN (morrow). The NP node for the second sentence branches into DT (its). The entire sentence structure is then combined into a single S node at the top.

Ah, distinctly I remember it was in the bleak December;  
A **R**Bea **P**re **S**e **V**Brat **D**T **y** **NN**; ember wrought its ghost upon the floor.  
Eagerly I wished the morrow;—vainly I had sought to borrow  
From my books surcease of sorrow—sorrow for the lost Lenore—  
For the rare and radiant maiden whom the angels name Lenore—  
Nameless here for evermore.



## 4. Named Entity Extraction & Disambiguation

Some words are more important than others. **Named entities** are **uniquely identified** by their name, for example specific persons, locations, or times that are important in a story.

Ah, distinctly I remember it was in the bleak December;  
And each separate dying ember wrought its ghost upon the floor.  
Eagerly I wished the morrow;—vainly I had sought to borrow  
From my books surcease of sorrow—sorrow for the lost Lenore—  
For the rare and radiant maiden whom the angels name Lenore—  
Nameless here for evermore.



# 5. Search and Similarity

The diagram illustrates a search process. On the left, a screenshot of the Google search homepage shows a search bar with the query "Help, I'm stuck in a geeky lecture! Who even is Edgar Allan Poe?!" A blue L-shaped arrow points from the search bar to the right side of the image. On the right, a screenshot of a Wikipedia article titled "Edgar Allan Poe" is shown. The page includes a grid of five portrait images of Edgar Allan Poe, a "More images" button, the title "Edgar Allan Poe", the subtitle "American writer", and a biography.

Help, I'm stuck in a geeky lecture! Who even is Edgar Allan Poe?!

X

Google Search I'm Feeling Lucky

Edgar Allan Poe

American writer

Edgar Allan Poe was an American writer, poet, editor, and literary critic. Poe is best known for his poetry and short stories, particularly his tales of mystery and the macabre. He is widely regarded as a central figure of Romanticism in the United States, and of American literature. [Wikipedia](#)

**Born:** January 19, 1809, Boston, Massachusetts, United States

**Died:** October 7, 1849, Church Home & Hospital, Baltimore, Maryland, United States

# 6./7. Language Models



**Does not compute.**

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

**Computes.**

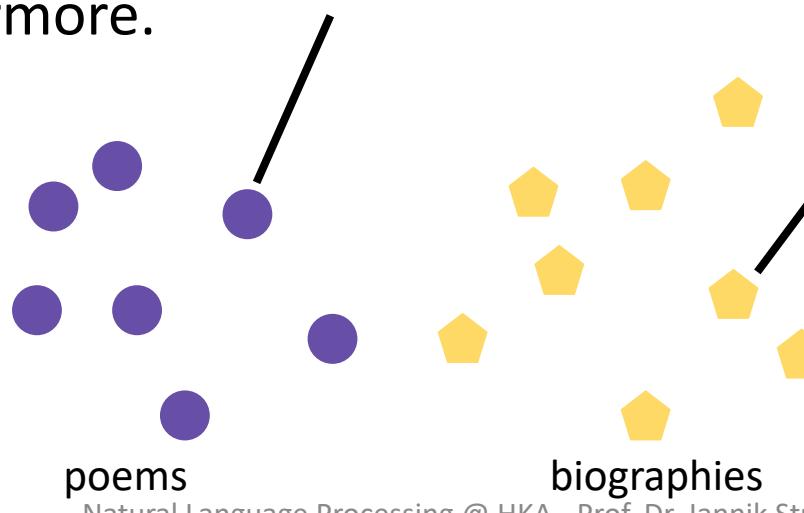
# 8. Text Mining

Ah, distinctly I remember it was in the bleak December;  
And each separate dying ember wrought its ghost upon the  
Eagerly I wished the morrow;—vainly I had sought to borro  
From my books surcease of sorrow—sorrow for the lost Lei  
For the rare and radiant maiden whom the angels name Le  
Nameless here for evermore.

Edgar Allan Poe

American writer

Edgar Allan Poe was an American writer, poet, editor, and literary critic. Poe is best known for his poetry and short stories, particularly his tales of mystery and the macabre. He is widely regarded as a central figure of Romanticism in the United States, and of American literature. [Wikipedia](#)

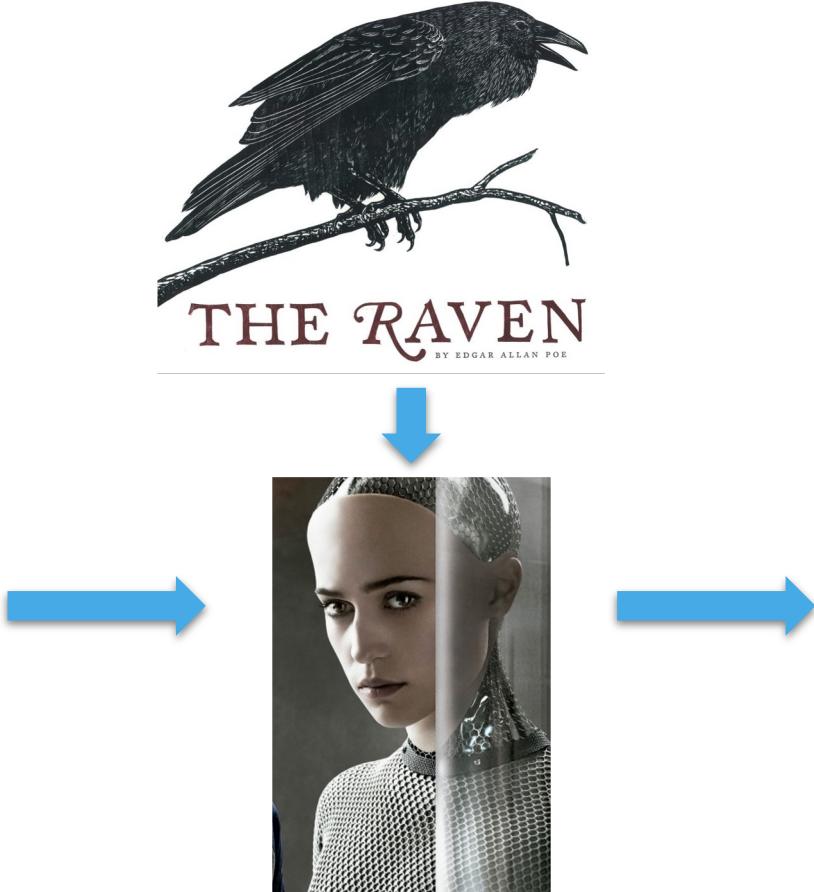
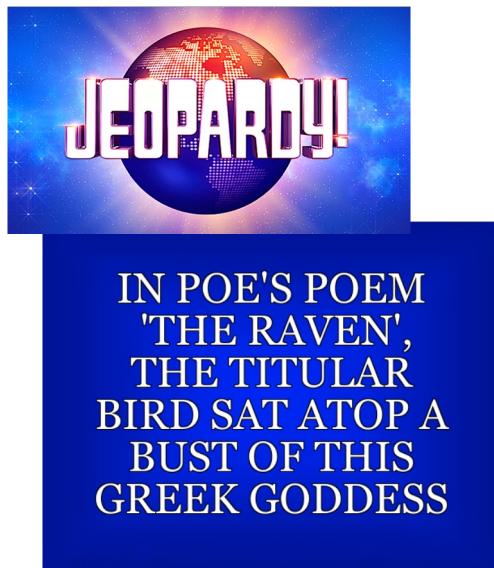


# 9. Opinion Mining and Sentiment Analysis

Ah, distinctly I remember it was in the **bleak December**;  
And each separate **dying ember** wrought its **ghost** upon the floor.  
Eagerly I wished the morrow;—**vainly** I had sought to borrow  
From my **books surcease of sorrow**—**sorrow** for the lost Lenore—  
For **rare and radiant maiden** **from the angels name Lenore**—  
Nameless here for evermore.



# 10. Information Extraction & Question Answering



Pallas Athene

# 11. Applications exploiting NLP



# 12. NLP with LLMs

ChatGPT 3.5 ▾

JA You

I am a lecturer at a university and teach a class about NLP. Is it still useful to teach topics such as linguistic preprocessing?

ChatGPT

Yes, teaching topics such as linguistic preprocessing remains highly useful in a class about Natural Language Processing (NLP), despite the advancements in large language models like GPT (such as ChatGPT). Here's why linguistic preprocessing is still relevant:

1. **Data Cleaning and Standardization:** Linguistic preprocessing techniques such as tokenization, stemming, lemmatization, and stop word removal are essential for cleaning and standardizing raw text data. Even with powerful models like GPT, preprocessing ensures that the input data is in a suitable format for analysis and modeling.
2. **Dimensionality Reduction:** Preprocessing techniques help reduce the dimensionality of the input space, making it more manageable for subsequent analysis and modeling tasks. This can lead to more efficient processing and better performance, especially when dealing with large datasets.
3. **Normalization and Standardization:** Linguistic preprocessing techniques help normalize text data by converting it into a consistent format. This is crucial for ensuring that different variations of words or phrases are treated the same ✓ during analysis, which improves the accuracy and reliability of NLP applications.

Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.

# Next Steps I

- First Tutorial
  - Tomorrow, March 20, 2024
  - Goal: Foundations for exercise sheets → Python

# Next Steps II

- Please join the ILIAS course:

[https://ilias.h-ka.de/goto.php?target=crs\\_853687&client\\_id=HSKA](https://ilias.h-ka.de/goto.php?target=crs_853687&client_id=HSKA)

**Password:**  
**NLP2024NLP**



## Thank you! Questions?