

# Introduction

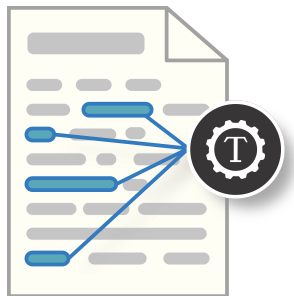
Jelke Bloem

Text Mining  
Amsterdam University College

February 6th, 2024

# What this course is about

**Text Mining is the process of deriving high-quality information from text**



# Text Mining

Marti Hearst on text mining (2003):

*Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.*

<https://people.ischool.berkeley.edu/~hearst/text-mining.html>

# Text Mining

- Subfield of data mining
- Often considered part of *data science*
- Interdisciplinary
  - ▶ Computational Linguistics
  - ▶ Natural Language Processing
  - ▶ Machine Learning
- Input: Unstructured text, text without semantic structure
- Output: Structured information

# Example applications

- Text classification (e.g. spam detection)
- Information extraction and annotation (e.g. Named Entity Recognition)
- Information retrieval and recommendation systems
- Conversational agents and Q&A systems
- Clustering (e.g. topic modelling)

# Example applications in research

- Quantitative linguistics (e.g. use of code switching (mixing languages within an utterance) in web text)
- Computational social science (e.g. mapping interactions between cultural communities using social media data)
- New Media studies (e.g. analysing the impact of tweets with certain types of content)
- History of Ideas research (e.g. studying the history of the concept of “infinity” throughout the scientific literature)
- Data Journalism (e.g. collecting data on causes of injury from varying online sources)
- Dialectometry (e.g. use of dialects on Twitter)
- Bibliometrics (e.g. studying which authors have influenced each other based on their references to each other)

# This course

Text Mining techniques applied to real-world systems and data

# Scientific and Technology goals

- Understanding vs doing
- Sometimes, you don't need to understand something in order to do it...
  - ▶ For example?
- Our primary focus is technological (doing)
- However, language technology can also be a method for understanding things



# Who am I?

I work at the University of Amsterdam as assistant professor of computational humanities

<https://www.uva.nl/profiel/b/l/j.bloem/j.bloem.html>

[j.bloem@uva.nl](mailto:j.bloem@uva.nl)

# Who are you?

- Sciences, Social Sciences or Humanities major?
- Concentration in Information?
- Year?
- Programming experience? Python?
- Machine Learning course?
- Linguistics course?
- What is your interest in this course?

*note to self: attendance*

# Q&A

# Language is complex

Take *your onion* and remove *the outer skin*. Take *your knife* and chop **it** in half.

What is *it*?

- The nearest antecedent?
- Subject of last sentence?
- Object of last sentence?
- Topic of the story?
- How is the computer supposed to know?

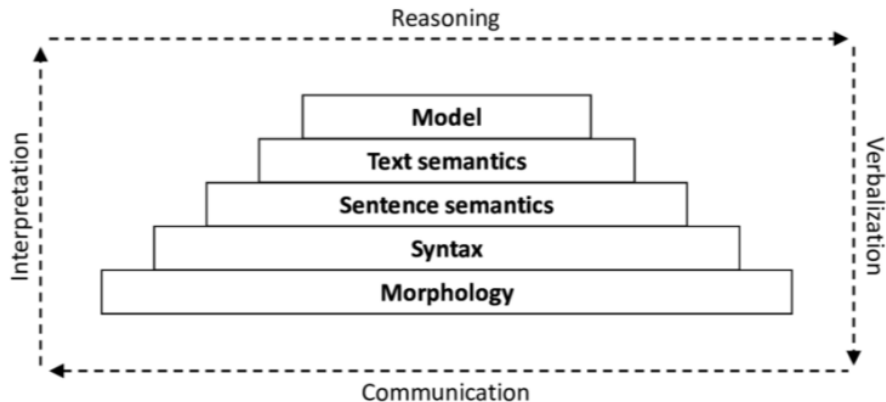
Task of *coreference resolution*

- Is *skin* a noun or a verb?
- Why is this in imperative mood? Is it an order? Advice?
- What is the context?

*Natural language understanding*

# Language is complex

- Language is a complex social object
- Language has multiple layers:



## What even is a word?

*This very course looks very interesting.*

- 01010100 01101000 01101001 01110011 00100000 01110110  
01100101 01110010 01111001
- ['T','h','i','s',' ','v','e','r','y',...]
- ['This','very','course','looks','very','interesting']
- [1,2,1,1,1,0,0,...]

Putting words into computers is not trivial. We will see more advanced word representations in this course.

# Example 1: Google web search

- Web crawling:
  - ▶ Many “spiders” (programs, bots) move around the Web following links. They start from lists of websites and sitemaps
  - ▶ Website owners have some control over what Google spiders do: e.g., specify what gets crawled, request a recrawl, etc. (robots.txt)
  - ▶ For every webpage, crawlers render it and download relevant data
  - ▶ ((What is the ‘dark web’?))



# Example 1: Google web search

- Indexing:

- ▶ Webpages are processed and content is extracted: text, images, structure (internal and external links), specific information
- ▶ An index is an inverted file. E.g., text is indexed in a full-text engine, while entities (people, persons, concepts) are indexed in a knowledge graph (which is used to populate the infobox)

The screenshot shows a Google search for 'Thanos'. The search bar at the top contains the word 'Thanos'. Below the search bar, there are navigation links for 'All', 'Images', 'News', 'Videos', 'Maps', and 'More'. The search results show 'About 95,600,000 results (0,52 seconds)'. The first result is 'Thanos - Wikipedia' with a link to 'https://en.wikipedia.org/wiki/Thanos'. Below the link, there is a brief description of Thanos as a villain in the Marvel Universe. To the right of the search results, there is an infobox for 'Thanos'. The infobox includes a main image of Thanos, a grid of smaller images, and a 'More images' link. Below the images, the title 'Thanos' is displayed, followed by the subtitle 'Fictional supervillain'. The infobox also contains a paragraph of text about Thanos, his first appearance, his place of origin, and his creators. At the bottom of the infobox, there is a 'Movies' section. On the left side of the infobox, there is a 'People also ask' section with four questions and their corresponding answers.

Google

thanos

Q All Images News Videos Maps More Settings Tools

About 95,600,000 results (0,52 seconds)

**Thanos - Wikipedia**  
<https://en.wikipedia.org/wiki/Thanos>  
Thanos is one of the most powerful villains in the Marvel Universe and has clashed with many heroes including the Avengers, the Guardians of the Galaxy, the Fantastic Four, and the X-Men.  
Team affiliations: **Infinity Watch**; **Black Order**  
Notable aliases: The Mad Titan  
First appearance: The Invincible Iron Man #55  
Place of origin: **Titan**  
The Thanos Quest · Titan · Thanos Rising · Thanos (name)

**People also ask**

Who defeated Thanos? ▾

Can Thanos kill the Hulk? ▾

Is Thanos a God? ▾

How strong is Thanos? ▾

Feedback

**Thanos**  
Fictional supervillain

Thanos is a fictional supervillain appearing in American comic books published by Marvel Comics. The character, created by writer/artist Jim Starlin, first appeared in The Invincible Iron Man #55. [Wikipedia](#)

**Played by:** [Josh Brolin](#)

**Species:** Eternal-Deviant hybrid

**First appearance:** The Invincible Iron Man #55 (February 1973)

**Place of origin:** [Titan](#)

**Creators:** [Jim Starlin](#), [Mike Friedrich](#)

**Movies**



# Example 1: Google web search

Meaning of your query

Relevance of webpages

Quality of content

Usability of webpages

Context and settings

## Replace

How to **change** a light bulb



## Exchange

Does post office **change** foreign currency



## Adjust

How to **change** brightness on a laptop



# Example 1: Google web search

Meaning of your query

Relevance of webpages

Quality of content

Usability of webpages

Context and settings

- How similar a page's content is to the query (in terms of keywords, entities, etc)
- How recently it was updated
- How often it was clicked from the same or similar query by other users ...

# Example 1: Google web search

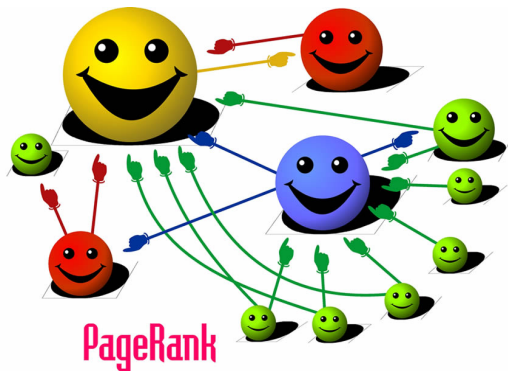
Meaning of your query

Relevance of webpages

Quality of content

Usability of webpages

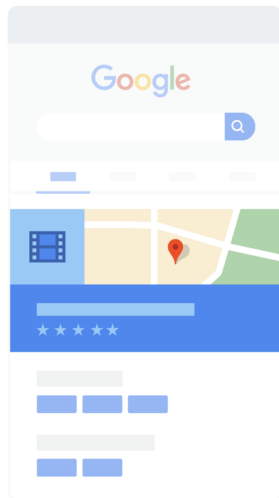
Context and settings



# Example 1: Google web search

- Results:

- ▶ Ranking
- ▶ Providing the right target (a webpage, a link to maps, etc.)
- ▶ User experience / User interface



## Example 1: Google web search

How is text mining used in Google search?

## Example 2: Research

Word embeddings are a Machine Learning technique to represent the meaning of words in a vector space, according to how they are used in a corpus.

Researchers have trained embeddings on a large dataset of materials science publication abstracts. Results are impressive..

# LETTER

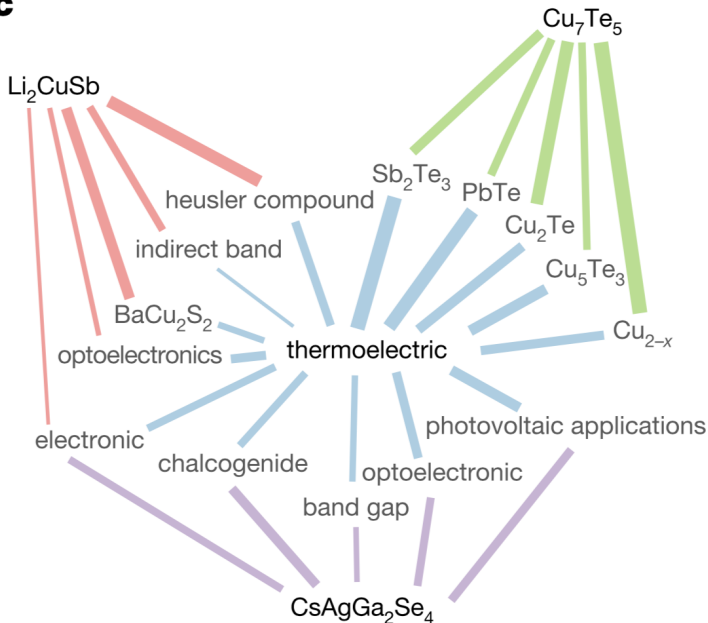
<https://doi.org/10.1038/s41586-019-1335-8>

## Unsupervised word embeddings capture latent knowledge from materials science literature

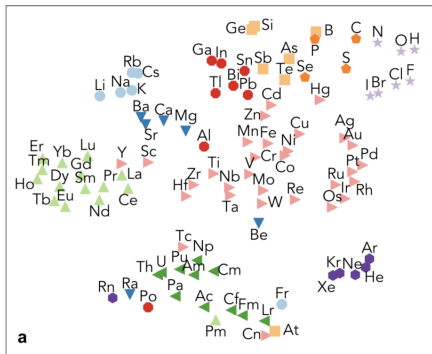
Vahe Tshitoyan<sup>1,3\*</sup>, John Dagdelen<sup>1,2</sup>, Leigh Weston<sup>1</sup>, Alexander Dunn<sup>1,2</sup>, Ziqin Rong<sup>1</sup>, Olga Kononova<sup>2</sup>, Kristin A. Persson<sup>1,2</sup>, Gerbrand Ceder<sup>1,2\*</sup> & Anubhav Jain<sup>1\*</sup>

## Example 2: Material science research

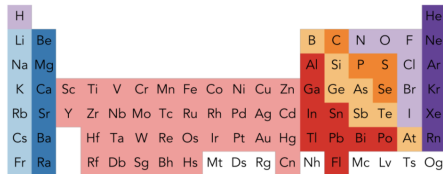
**c**



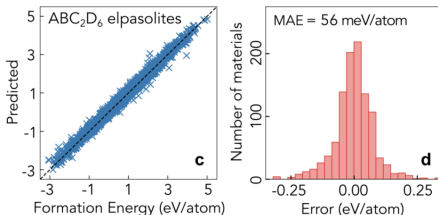
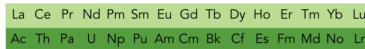
## Example 2: Material science research



- alkali metal
- ▼ alkaline earth metal
- ▲ lanthanide
- ▲ actinide
- ▶ transition metal
- post-transition metal
- metalloid
- ◆ polyatomic nonmetal
- ★ diatomic nonmetal
- noble gas



**b**








## Example 2: Material science research

How can text mining be used in research?

## Example 3: Language generation

ChatGPT		
 Examples	 Capabilities	 Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

## Example 3: Language generation

S

What should one use chatGPT for?



ChatGPT can be used for a variety of tasks, such as:



- Generating human-like text, such as writing essays, articles, or stories
- Answering questions in a conversational manner
- Generating responses in a chatbot or virtual assistant
- Summarizing or translating text
- Generating programming code
- And many more!

It can also be fine-tuned on specific task or domain to improve performance.

## Example 3: Language generation

J

Describe the incident in which a chatbot was accidentally elected onto the University of Amsterdam student council in 2005. Please answer in 100 words or less and provide a bibliography of relevant historical sources.



In 2005, a chatbot called "Tron" was accidentally elected onto the University of Amsterdam student council. The incident occurred when students mistook the chatbot, which was created as a prank by a group of computer science students, for a real person while voting for representatives online. Tron received more than 100 votes, earning it a spot on the council. The incident sparked debate about the potential for technology to deceive people and the importance of verifying the identities of online candidates.

Bibliography:

- "Tron the Chatbot Elected to Student Council" (2005) The New York Times
- "Chatbot Tron Elected to Student Council" (2005) BBC News
- "The Rise of the Machines: How a Chatbot Was Elected to a Dutch University's Student Council" (2005) Wired

## Example 3: Language generation

What is the connection between text mining and language generation (e.g. ChatGPT)?

# Q&A

# What I want to do

- Foundations of doing computational work with human language
- Python skills for text mining
- One in-depth topic: language models
- First-hand experience with some NLP and Machine Learning tasks:
  - ▶ Classification
  - ▶ Sentiment analysis
  - ▶ Recommendation systems
  - ▶ Clustering
  - ▶ Topic modelling

# How this course will work

**Applied:** we will foreground applications over implementations, and minimize theory to the necessary

**Python:** we will only use this programming language. Motivation: widely adopted, rich ecosystem, gentle onboarding

Python is an interpreted high-level programming language for general-purpose programming. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.



# Learning objectives and assessment

- ① You develop an understanding of the basic concepts and techniques used to mathematically represent and model language, including their underpinnings in linguistics.
- ② You develop programming skills for automatically extracting and processing textual data from different sources (such as textual corpora and social media).
- ③ You will be familiar with the main approaches to distributional semantics, including modern deep learning techniques, and are able to automatically extract distributional information from corpora.
- ④ You will be able to automatically characterise the semantic content of a document.
- ⑤ You will develop programming skills necessary for building his/her own text mining application (e.g., a search engine or a recommender system) by exploiting statistical learning approaches.

# Learning objectives and assessment

- 1 You develop an understanding of the basic concepts and techniques used to mathematically represent and model language, including their underpinnings in linguistics.

→ Reading Assignments (in-class) - 25% (5x 5%)

# Learning objectives and assessment

- 2 You develop programming skills for automatically extracting and processing textual data from different sources (such as textual corpora and social media).
- 3 You will be familiar with the main approaches to distributional semantics, including modern deep learning techniques, and are able to automatically extract distributional information from corpora.
- 4 You will be able to automatically characterise the semantic content of a document.

→ Programming Assignments (take-home) - 30% (3x 10%)

# Learning objectives and assessment

- 5 You will develop programming skills necessary for building his/her own text mining application (e.g., a search engine or a recommender system) by exploiting statistical learning approaches.

→ Group Project - 35%

→ In-class presentation - 10%

## Show Canvas and GitHub

- Timetable
- Schedule and topics
- Assessment
- Materials and references
- Prerequisites

# Q&A

# For the next class

- Set up your Python working environment
  - ▶ We will mostly use Jupyter notebooks <https://jupyter.org>, try to get familiar with them
  - ▶ I recommend Google Colaboratory (online) or Anaconda (offline)
    - ★ <https://colab.research.google.com/>
    - ★ <https://www.anaconda.com/distribution> (install Python 3.5+ and create a virtual environment for the course)
  - ▶ However, you are free to choose how to run your code
- Next lab will be a **Python refresher**: you will have several notebooks with a lot of Python to play with in small groups, to make sure you know the basics.
- Check the prerequisite references
  - ▶ Let me know by Thursday if you are at a loss with any of them
  - ▶ The materials on Python are for reference as well
  - ▶ **Please read the materials on linguistics by next week**
    - ★ Manning and Schütze, ch. 1 and 3