Amsterdam university college

Introduction

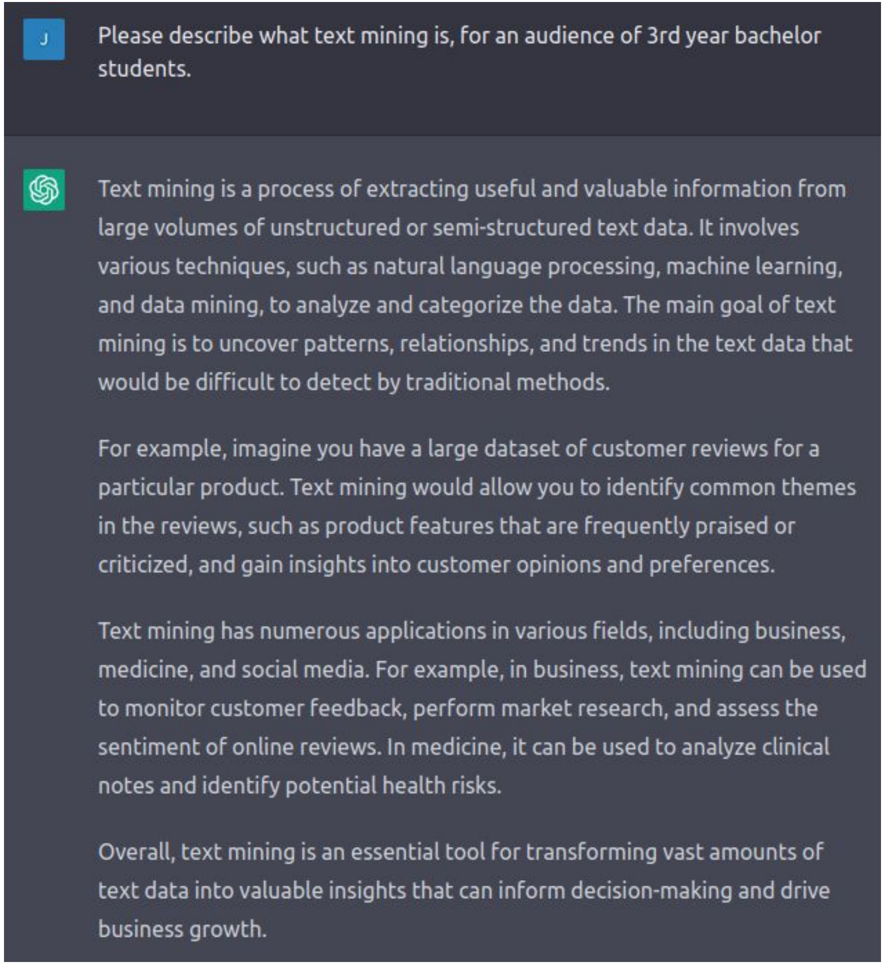
Text Mining

Jelke Bloem

What this course is about

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

What this course is about



J

Please describe what text mining is, for an audience of 3rd year bachelor students.

Text mining is a process of extracting useful and valuable information from large volumes of unstructured or semi-structured text data. It involves various techniques, such as natural language processing, machine learning, and data mining, to analyze and categorize the data. The main goal of text mining is to uncover patterns, relationships, and trends in the text data that would be difficult to detect by traditional methods.

For example, imagine you have a large dataset of customer reviews for a particular product. Text mining would allow you to identify common themes in the reviews, such as product features that are frequently praised or criticized, and gain insights into customer opinions and preferences.

Text mining has numerous applications in various fields, including business, medicine, and social media. For example, in business, text mining can be used to monitor customer feedback, perform market research, and assess the sentiment of online reviews. In medicine, it can be used to analyze clinical notes and identify potential health risks.

Overall, text mining is an essential tool for transforming vast amounts of text data into valuable insights that can inform decision-making and drive business growth.

Text Mining

“[Text Mining] is the discovery [..] of new, previously unknown information, by automatically extracting [it] from different written resources.” Marti A. Hearst

- Subfield of data mining
- Interdisciplinary
- Methods/theory from Computational Linguistics, Natural Language Processing and Machine Learning
- Input: unstructured texts; output: structured information

Example applications

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

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.

Logistics

Go to Canvas and GitHub

Discuss:

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

Who am I?

I work at UvA as assist. prof. of computational humanities

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

j.bloem@uva.nl

Who are you?

- Which major? Sciences, Social Sciences, Humanities?
- Concentration in Information?
- Which year?
- Previous programming experience? In Python?
- Have taken the ML course?
- Why this course?

Note to self: take attendance.

Q&A

Language is hard

- Measurement?

“This very course looks very interesting.” --> 0.77

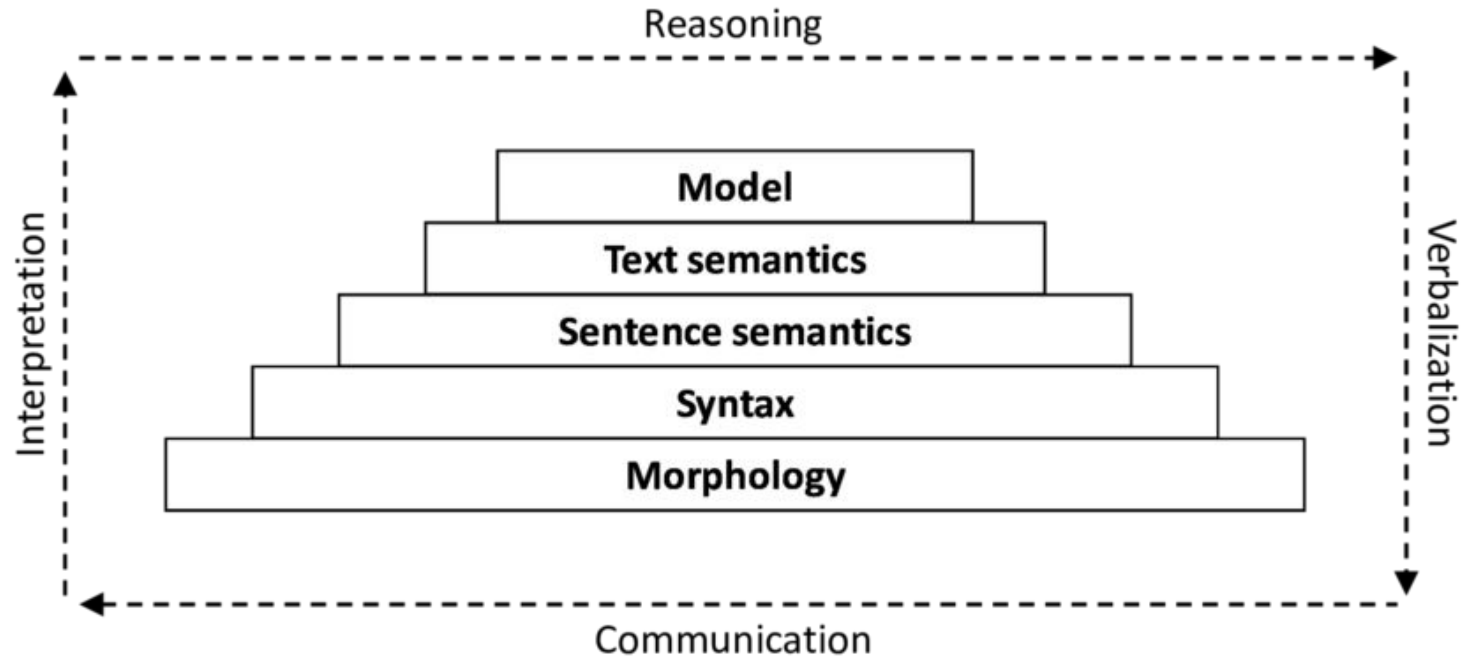
- Representation?

“This very course looks very interesting.” --> [1,2,1,1,1,0,0,...]

[this,very,course,looks,interesting,the,and,...] (bag of words representation)

Language is hard

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



One morning I shot an elephant in my pajamas.
How he got in my pajamas, I don't know

Ambiguity

“One morning I shot an elephant in my pajamas.”

- Who is in pajamas?
- Is “shot” a noun or verb?
- What does “shot” mean?
- What is the context?
- …

Technically, this task is usually framed as *coreference resolution*, or more broadly *language understanding* (useful, e.g., in Q&A systems: can you think why?)

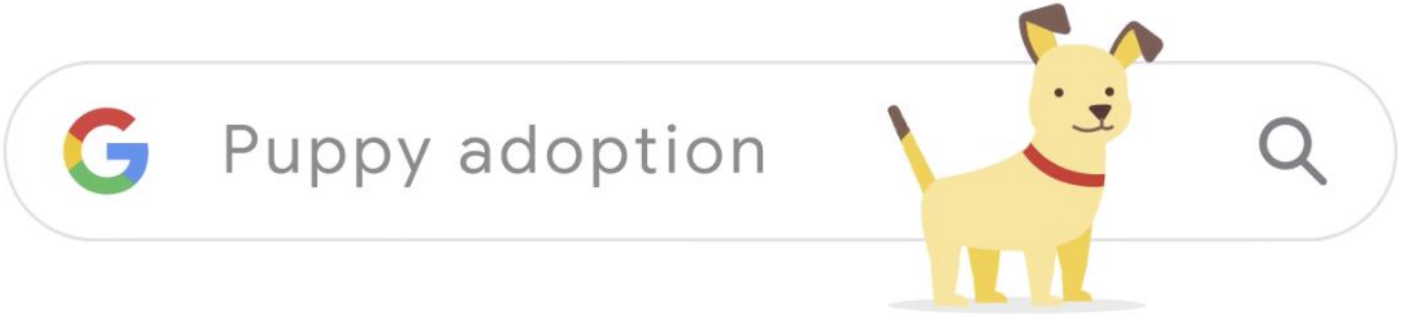
This course

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..
    - Can you think of examples?
- Our primary focus is technological (doing)

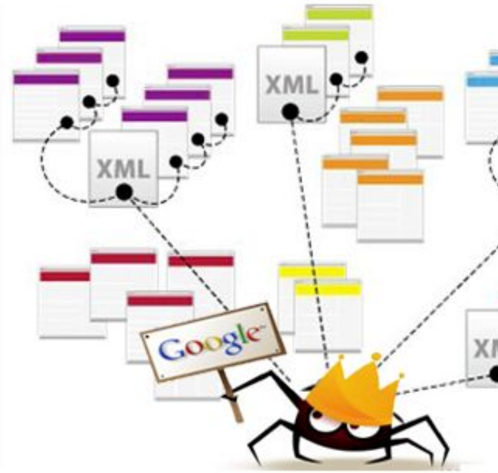
Example 1: Google search



Puppy adoption

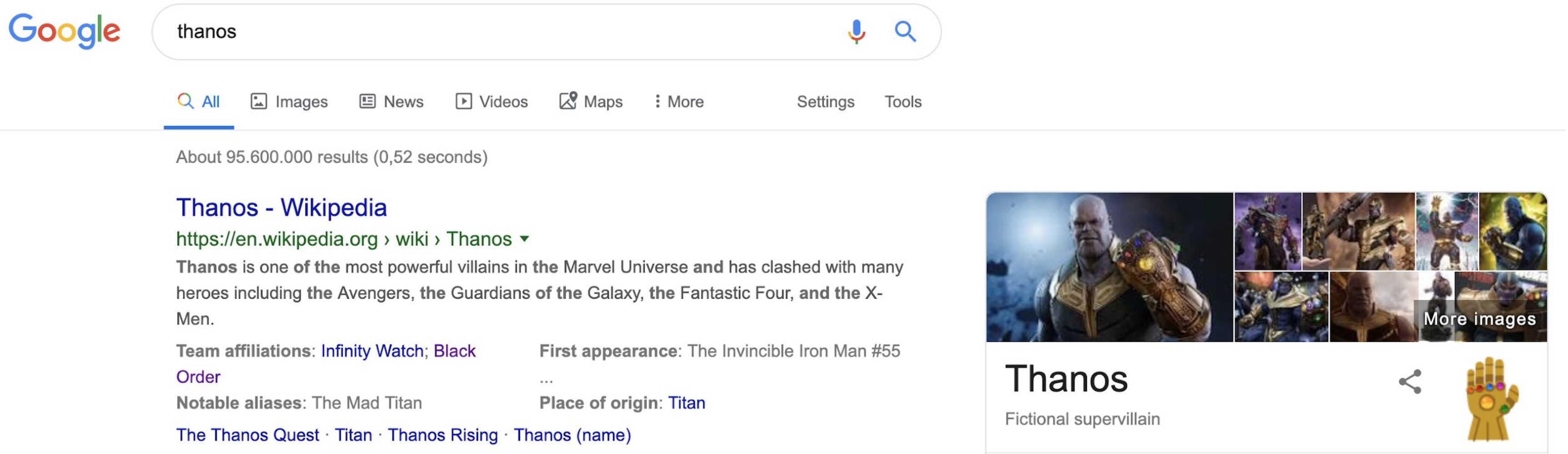
Example 1: Google search

- Web crawling:
    - Many “spiders” (programs) 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.
    - For every webpage, crawlers render it and download relevant data.



Example 1: Google 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 (what is behind the infobox).



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 ...

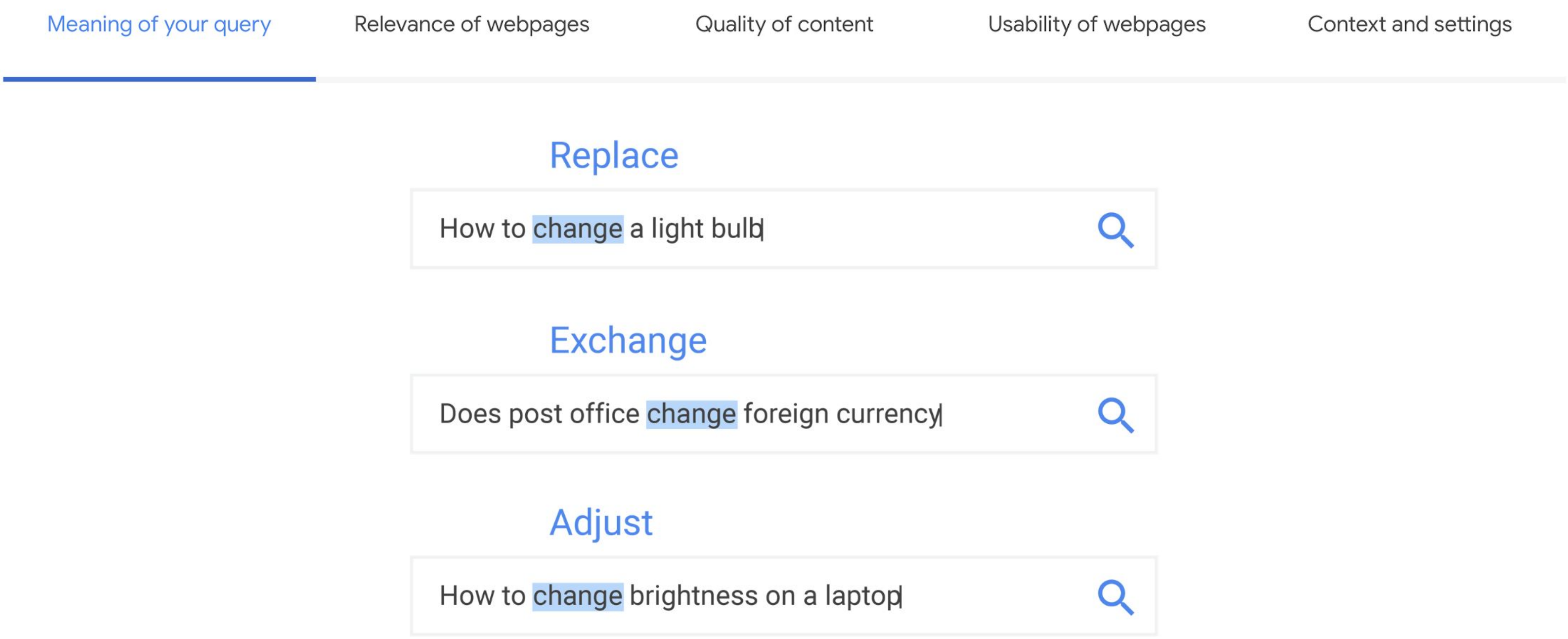
Place of origin: Titan

The Thanos Quest · Titan · Thanos Rising · Thanos (name)

Thanos

Fictional supervillain

Example 1: Google search



Replace

How to change a light bulb

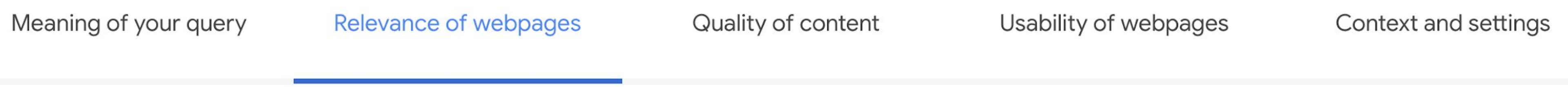
Exchange

Does post office change foreign currency

Adjust

How to change brightness on a laptop

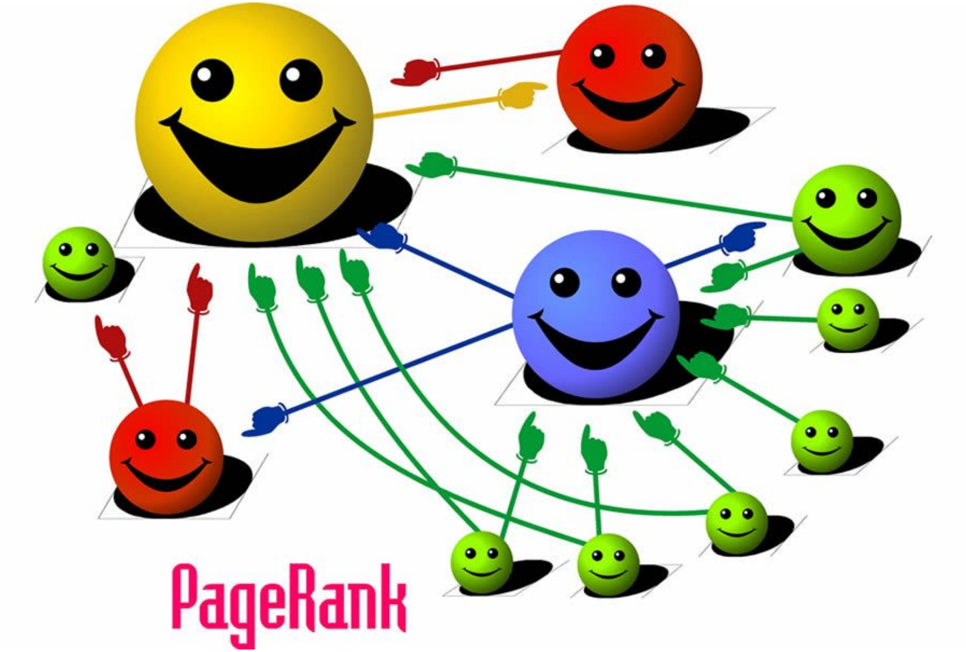
Example 1: Google search



- How similar a page content is to the query (e.g., keywords, entities, etc.).
- How recently was it updated, etc.
- How often was it clicked from the same or similar query by other users?
- ...

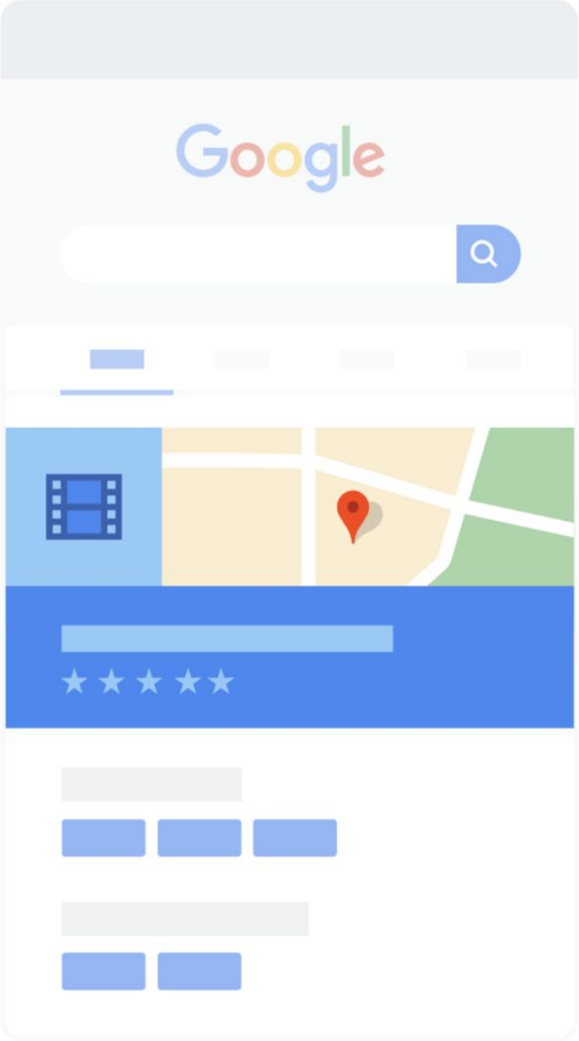
Example 1: Google search

Meaning of your query　Relevance of webpages　Quality of content　Usability of webpages　Context and settings



Example 1: Google search

- Results:
    - Ranking.
    - Providing the right target (a web page, a link to maps, etc.).
    - UX/UI.



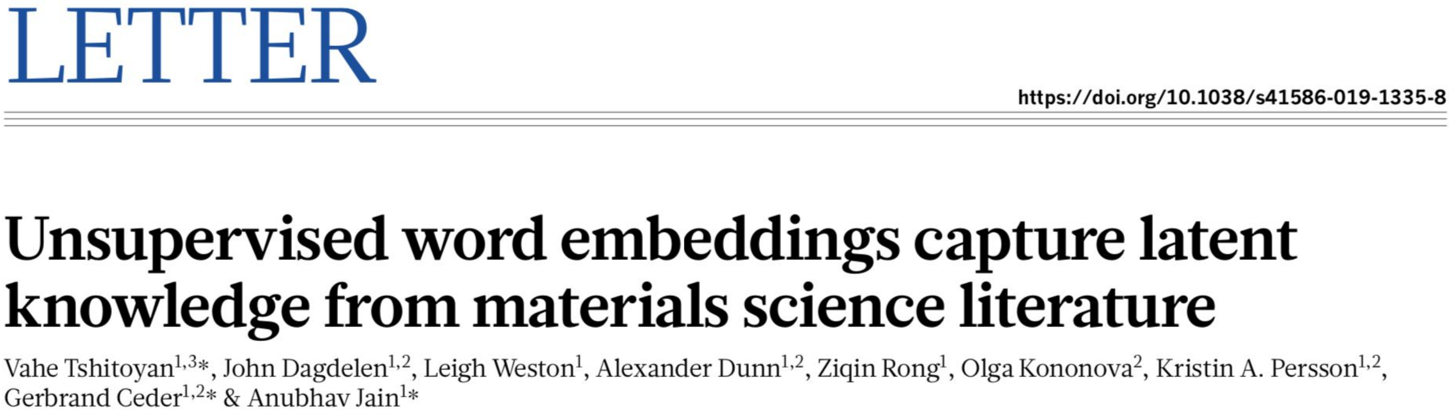
Example 1: Google search

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..



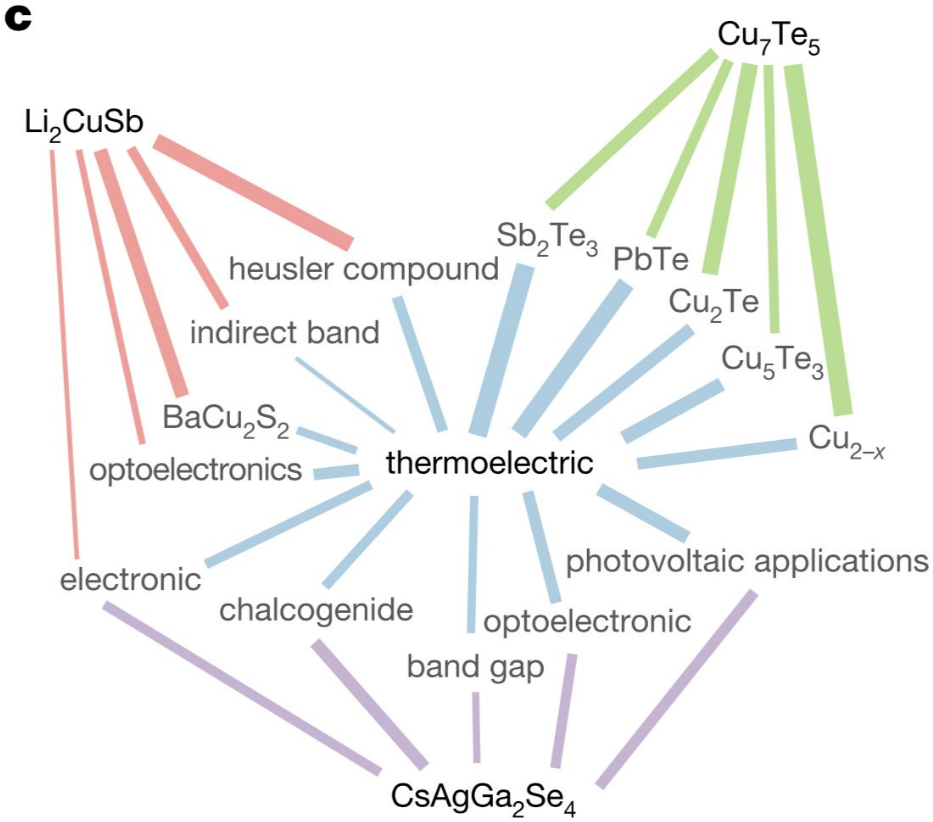
LETTER

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

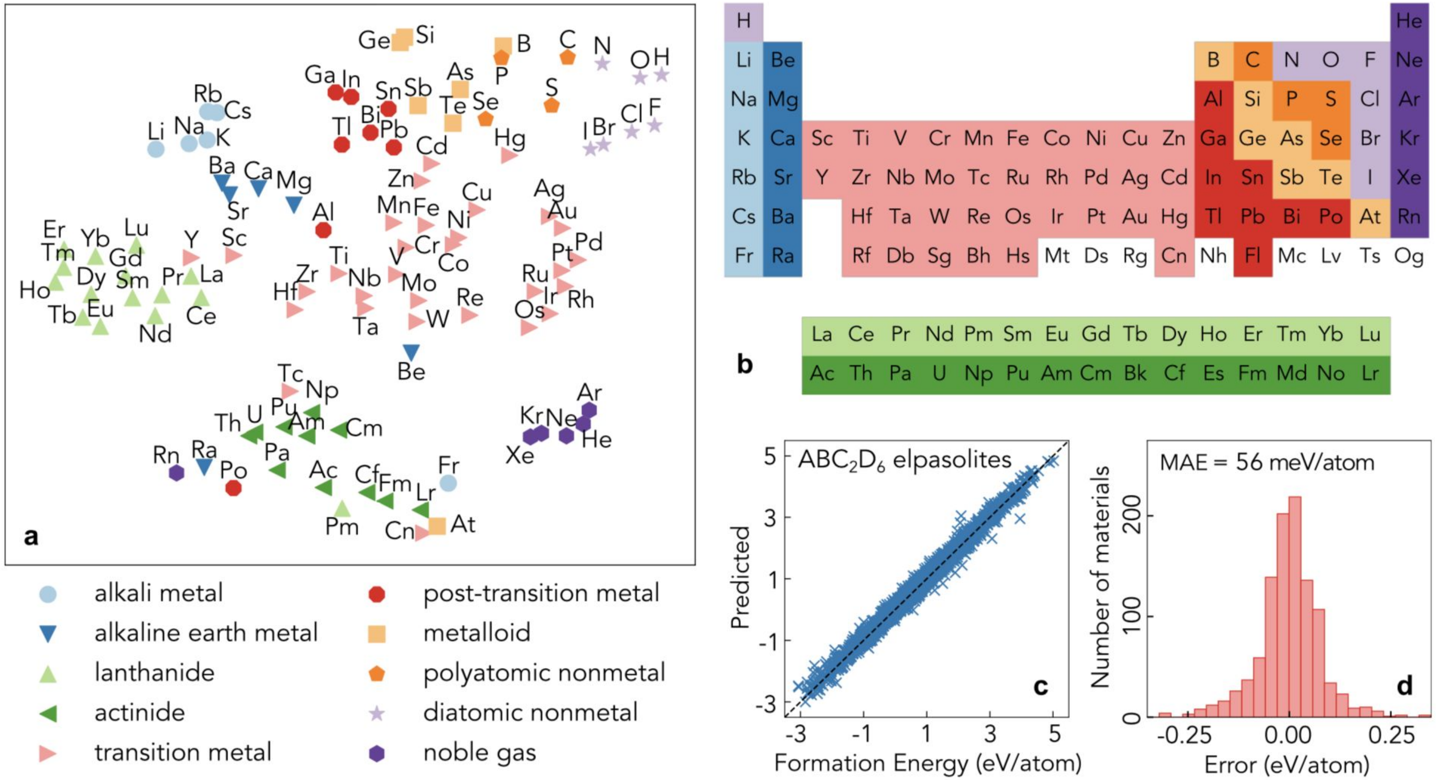
Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

Example 2: Research



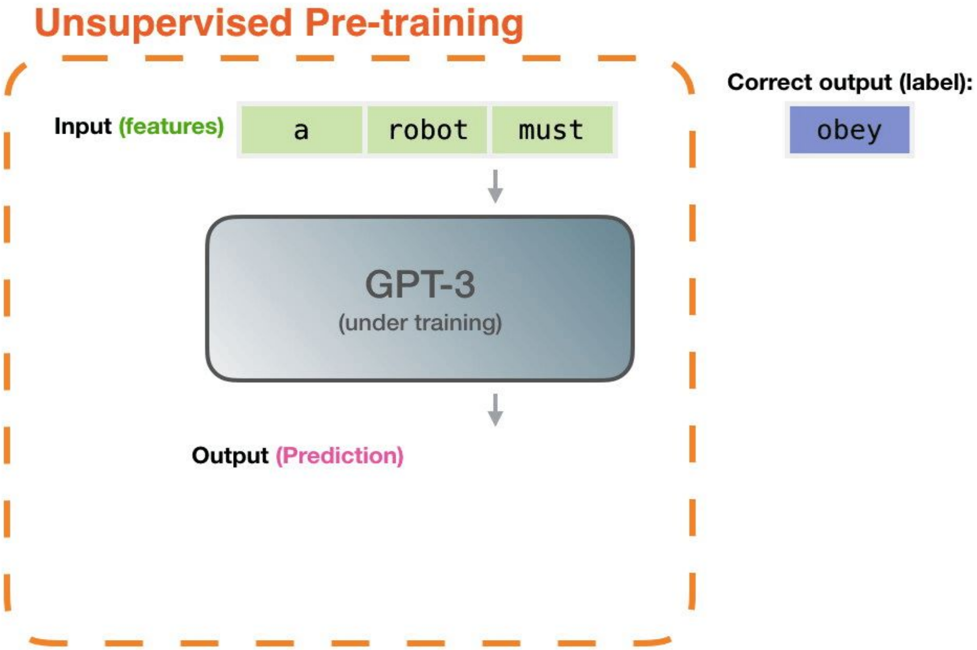
Example 2: Research



Example 2: Research

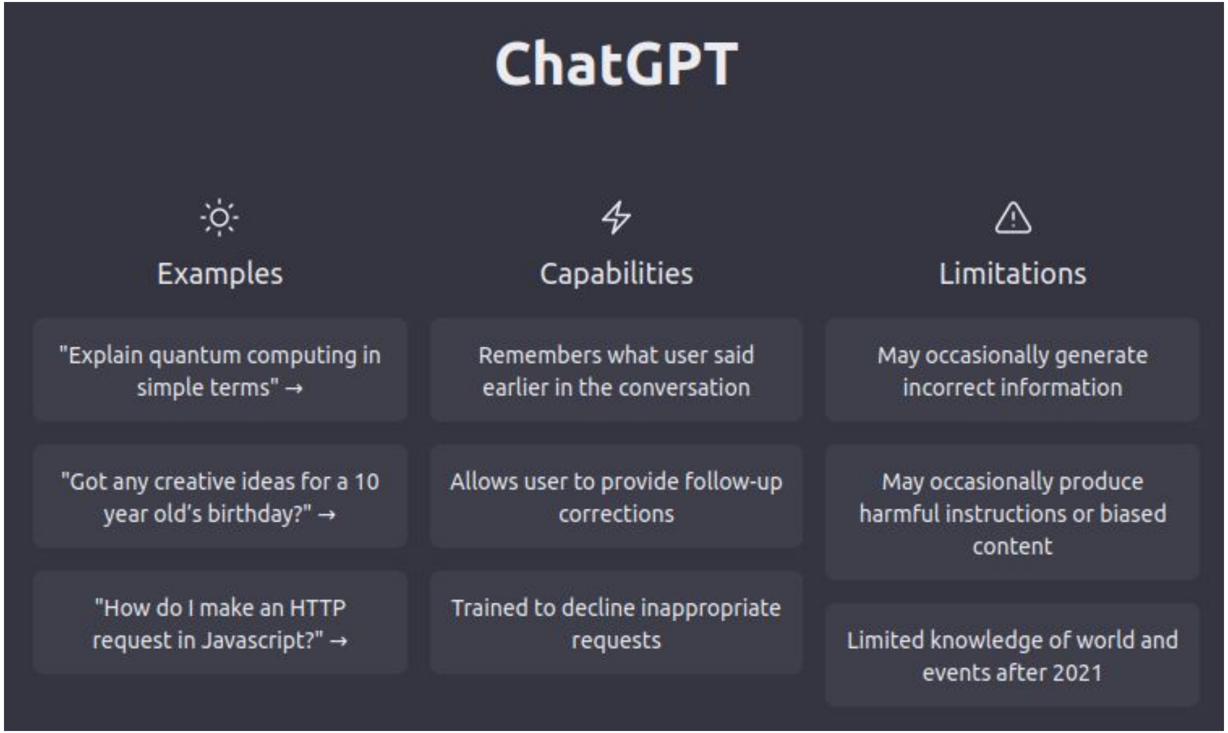
How can text mining be used in research?

Example 3: Language generation



Animation by Jay Alammar “How GPT3 Works”

Example 3: Language generation



ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

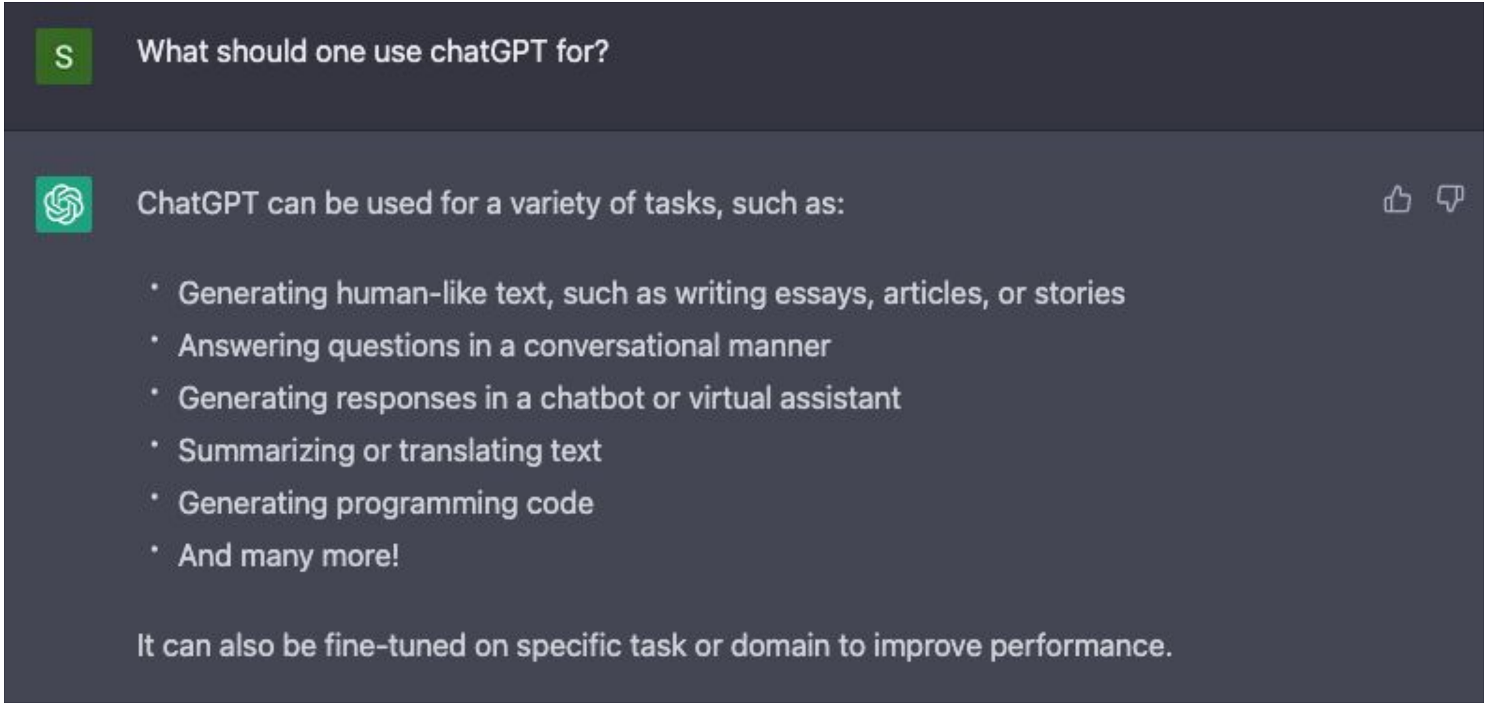
Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Example 3: Language generation



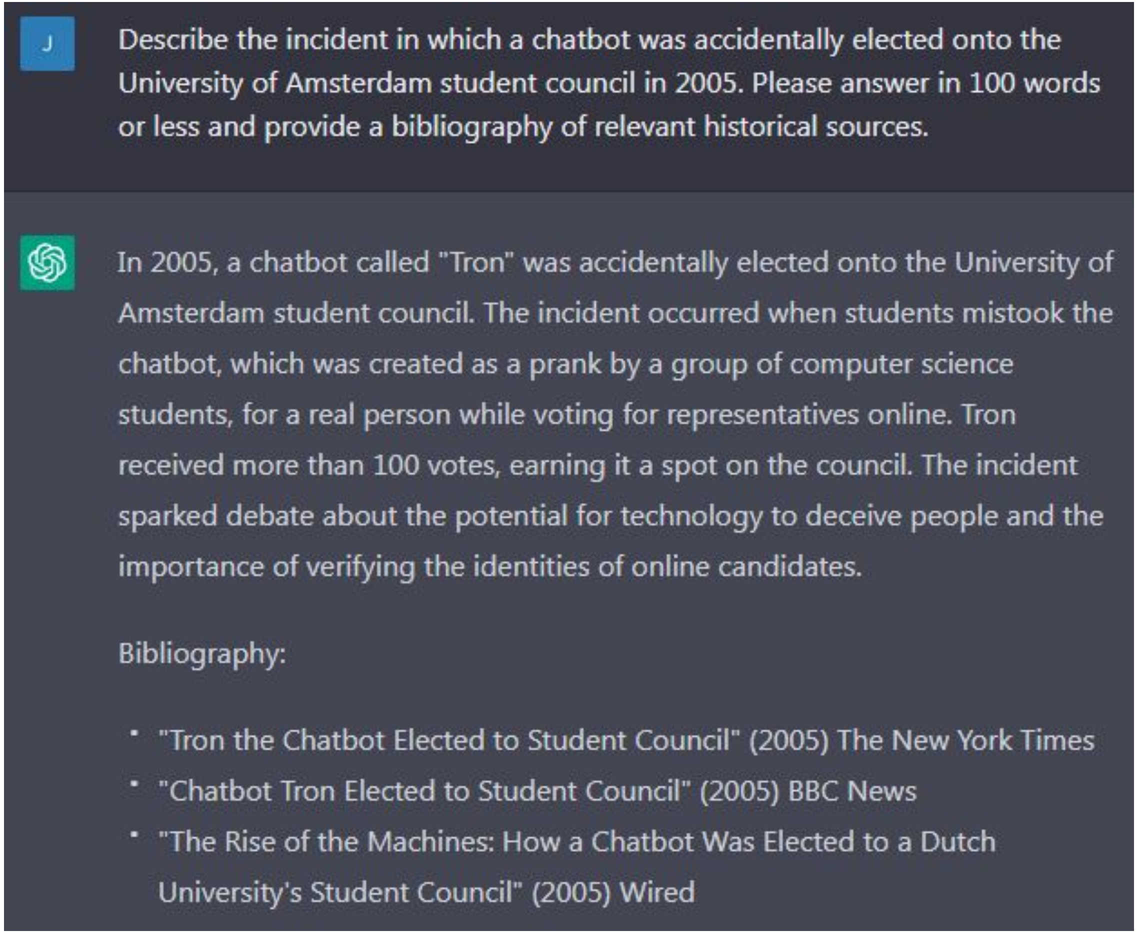
What should one use chatGPT for?

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



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

For the next class

- Setup your Python working environment
  - We will mostly use Jupyter notebooks https://jupyter.org, try to get familiar with them
  - I recommend Google Colab (online) or Anaconda (offline)
    - https://colab.research.google.com/
    - https://www.anaconda.com/distribution (install Python 3.7+ 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.