Introduction

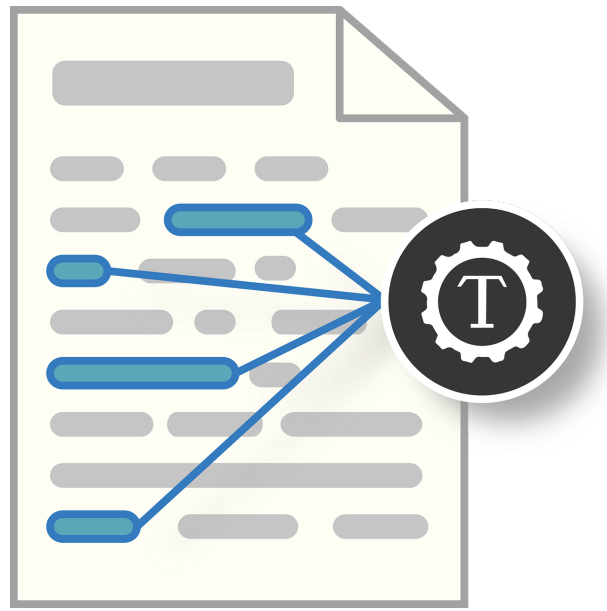
Jelke Bloem

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`

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?

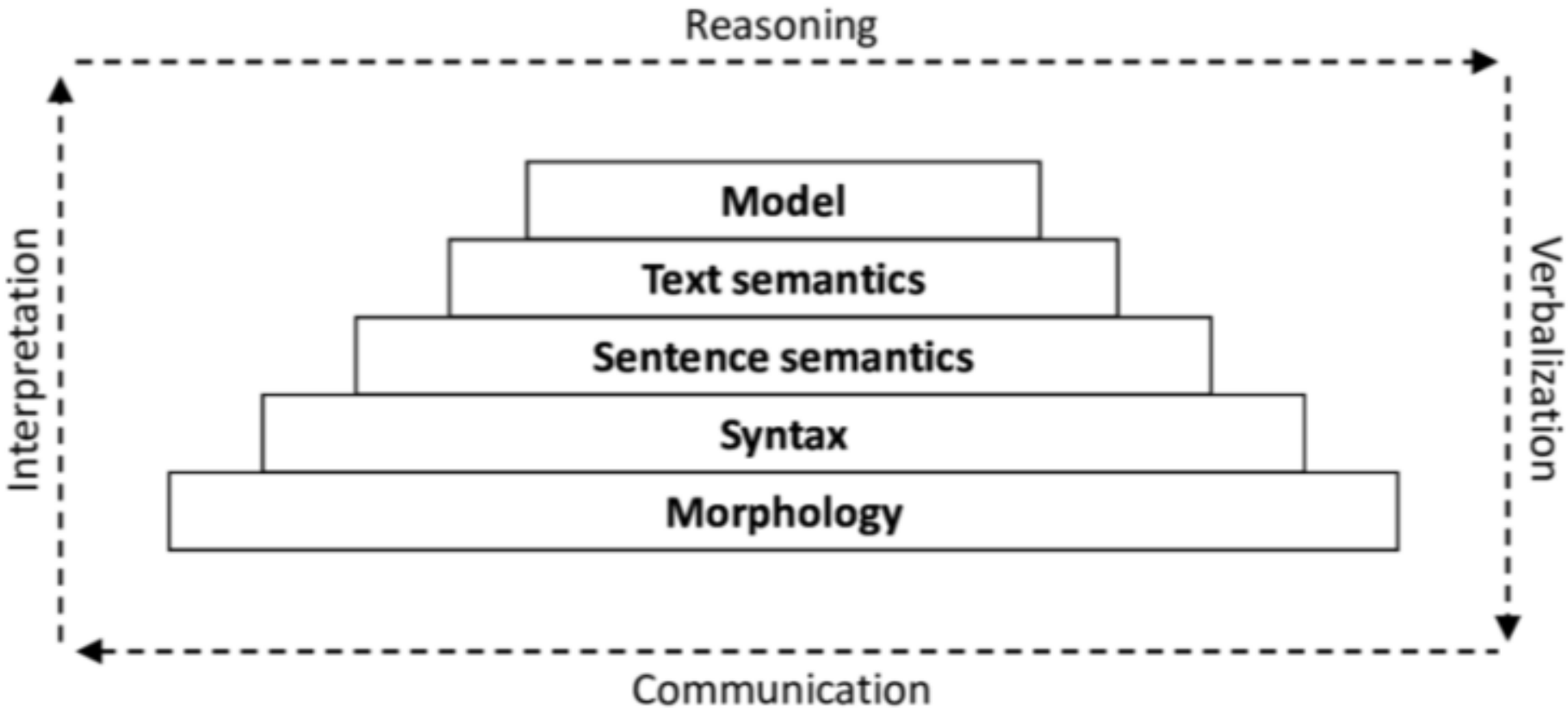
Task of *coreference resolution*

- 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,...]`

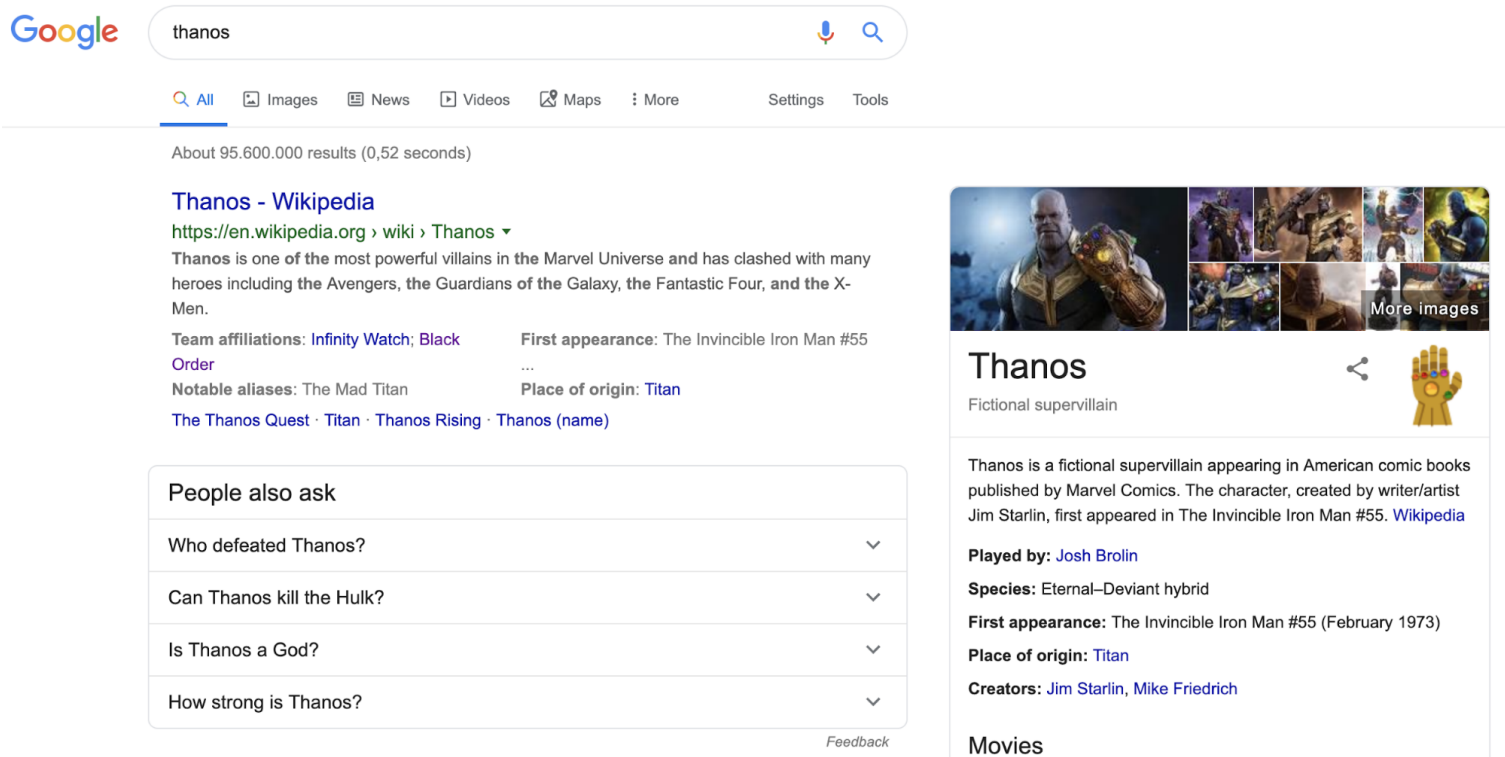
Putting words into computers is not trivial. We will see more advanced word representations in this course.

Example 1: Google web search

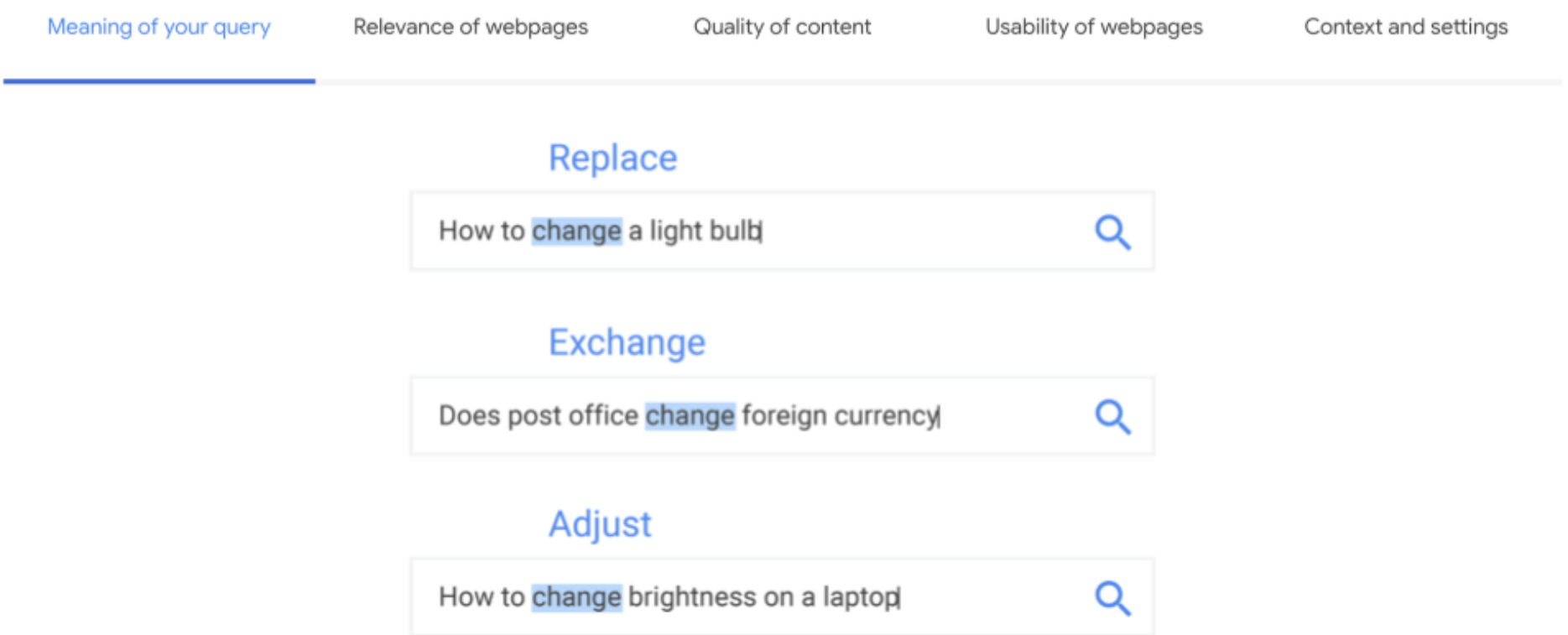
- Web crawling:
  - Many "spiders" (programs, bots) move around the Web following links. They start from lists of websites and sitemaps
  - Website owners have some control over what Google spiders do: e.g., specify what gets crawled, request a recrawl, etc. (robots.txt)
  - For every webpage, crawlers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

Example 1: Google web search

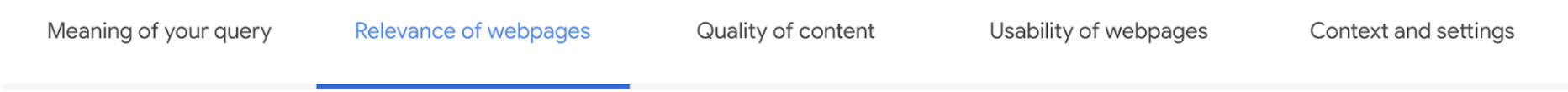
- Indexing:
  - Webpages are processed and content is extracted: text, images, structure (internal and external links), specific information
  - An index is an inverted file. E.g., text is indexed in a full-text engine, while entities (people, persons, concepts) are indexed in a knowledge graph (which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



Example 1: Google web search



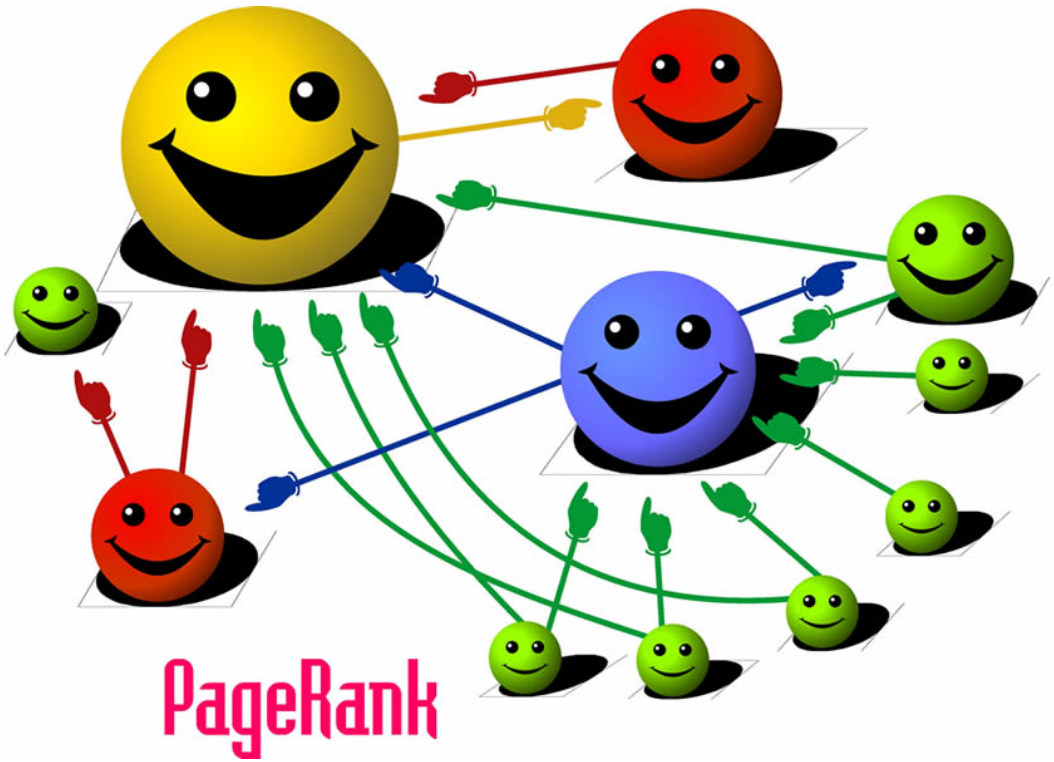
Example 1: Google web search



- 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 ...

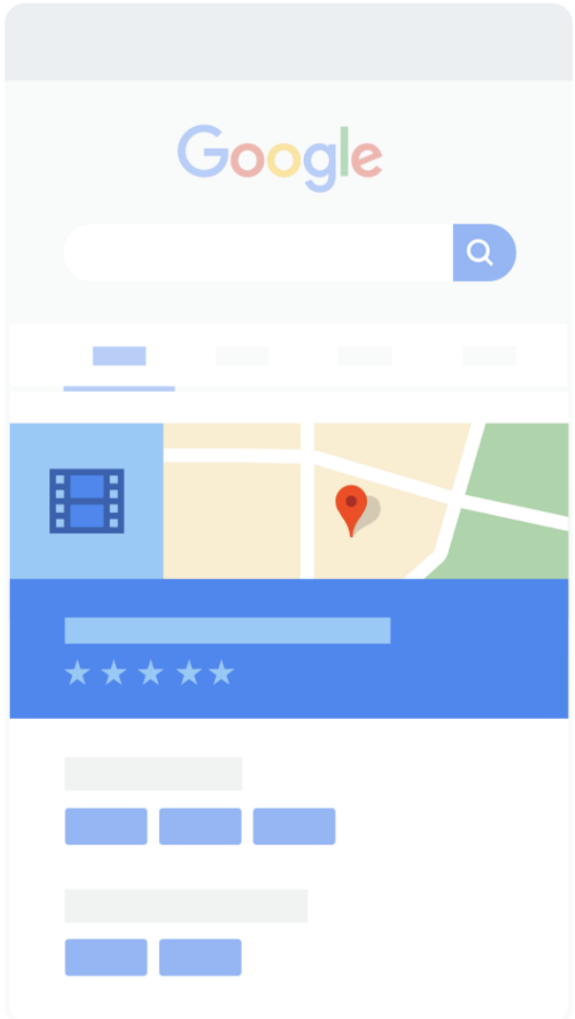
Example 1: Google web search

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..

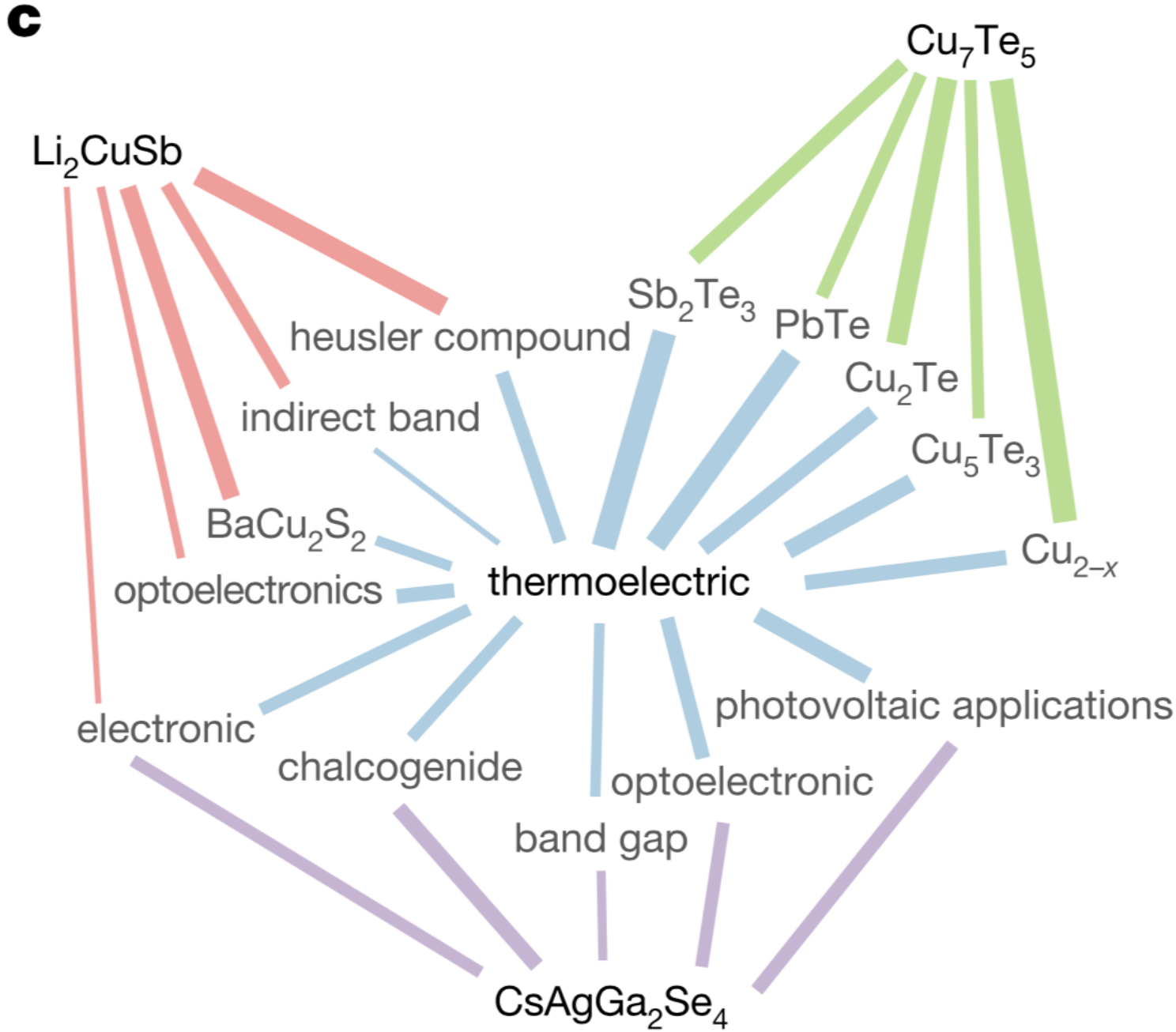
LETTER

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

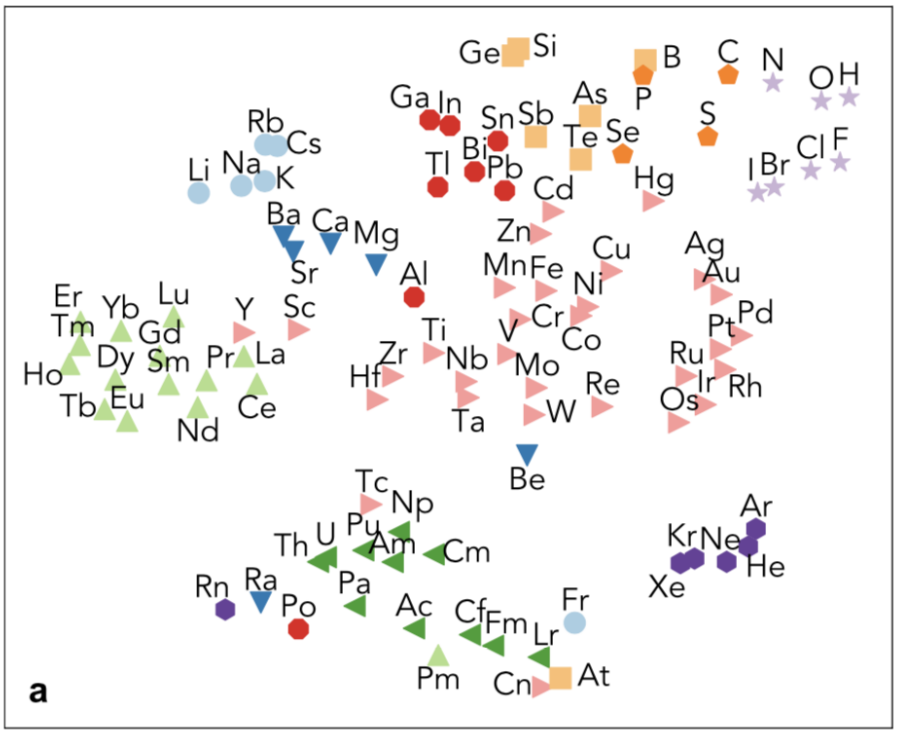
Unsupervised word embeddings capture latent knowledge from materials science literature

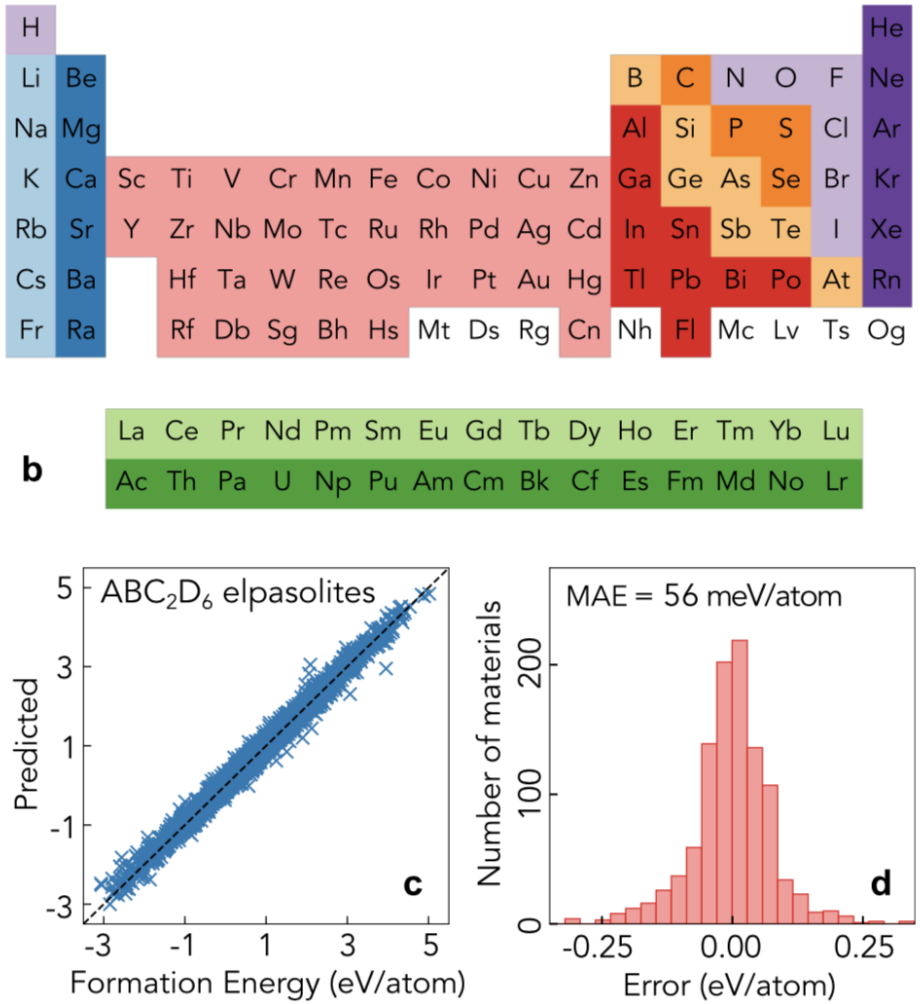
Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

Example 2: Material science research



Example 2: Material science research

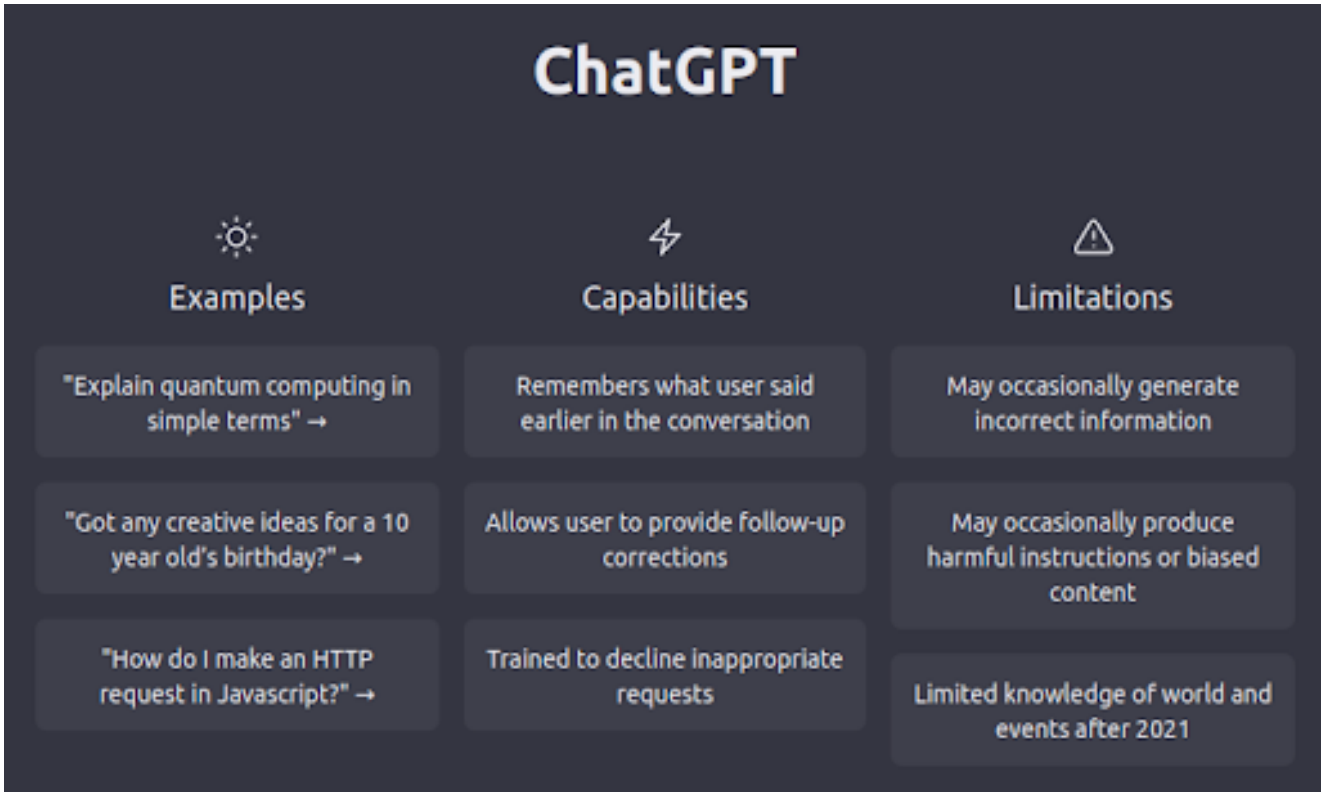




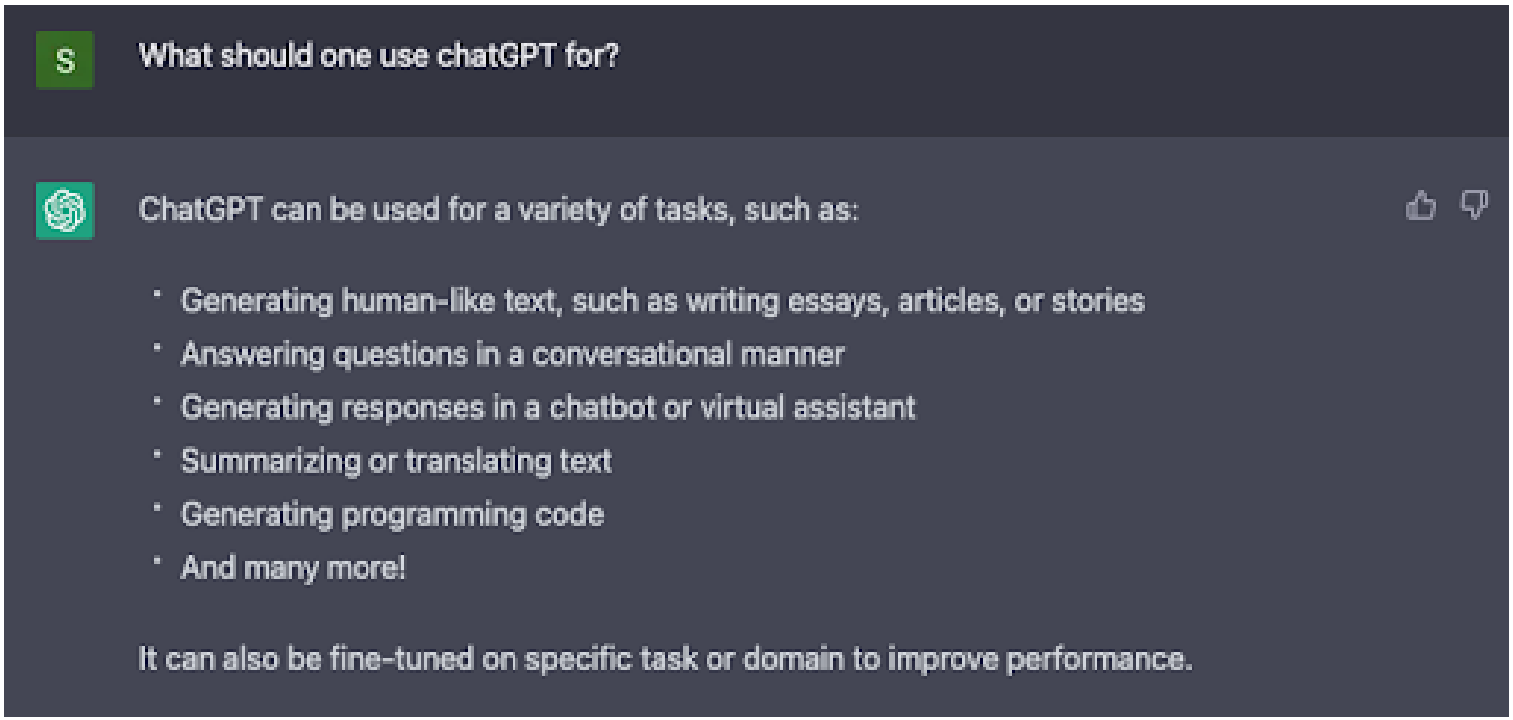
Example 2: Material science research

How can text mining be used in research?

Example 3: Language generation



Example 3: Language generation



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.

$\rightarrow$ 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.

$\rightarrow$ Group Project - 35%

$\rightarrow$ In-class presentation - 10%

Logistics

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