732A81/732A92/TDDE16 Text Mining (2022)

Course introduction

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING UNIVERSITY

Meet the team!

Ali Basirat

Ehsan Doostmohammadi

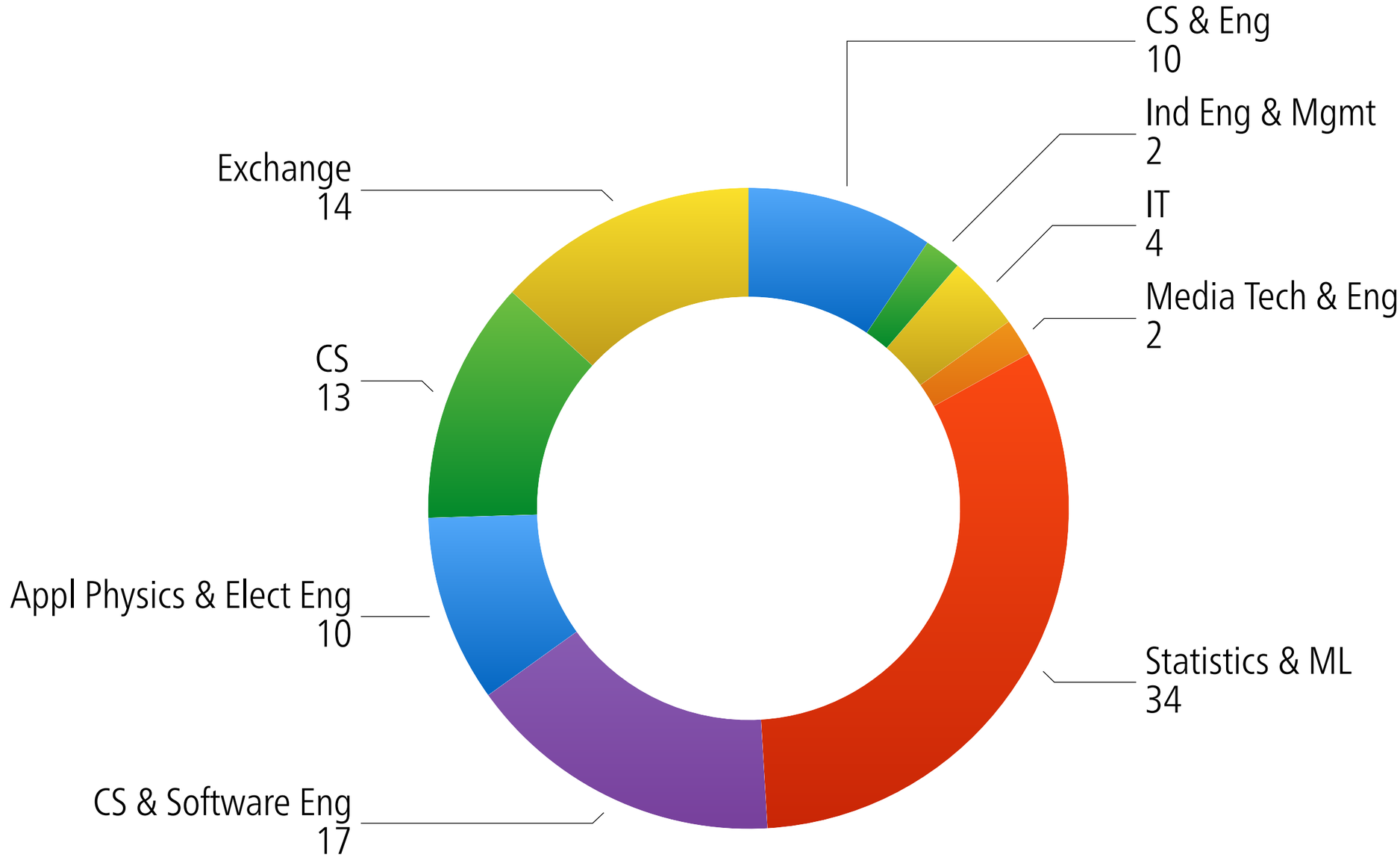
Jenny Kunz

Marco Kuhlmann

Oskar Holmström

Riley Capshaw

Meet your fellow students!



This session

- Introduction to text mining (15 minutes)
- Course logistics (15 minutes)
- Some simple examples of text mining (15 minutes)

Do you have any questions?

- **Synchronously**

  During the session, in the break, in the lab, …

- **Asynchronously**

  Email, course team, chat

  marco.kuhlmann@liu.se – marku61

- **Project-related questions**

  Schedule a meeting via the link on the course website.

Introduction to text mining

Text Mining is the process of accessing information in and extracting knowledge from large volumes of text.

Two functions

- **Information Access**

  Enable the user to access relevant information in time.

  search engines (pull), recommender systems (push)

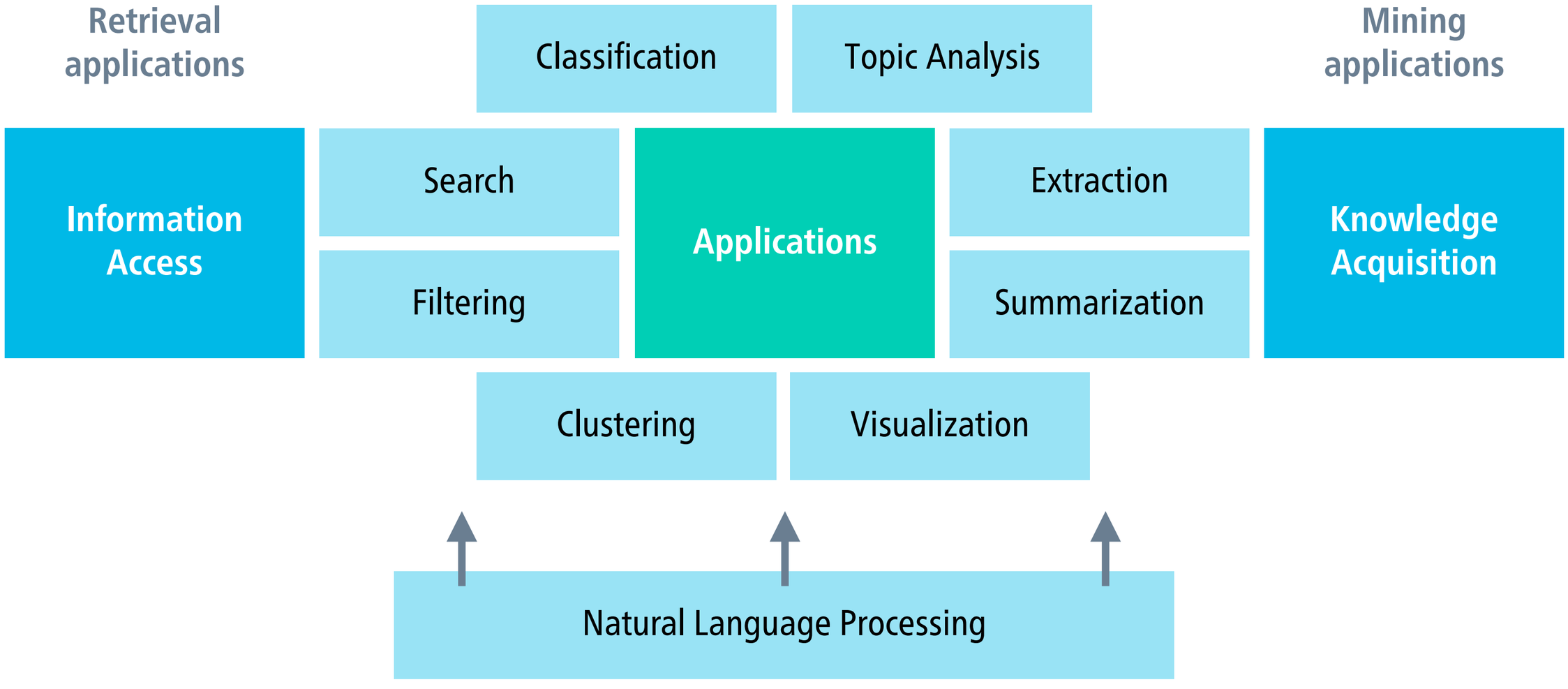
- **Knowledge Acquisition**

  Enable the user to acquire knowledge ‘hidden’ in text.

  information extraction, topic analysis

Adapted from Zhai and Massung (2016)

Conceptual framework for text mining



Adapted from Zhai and Massung (2016)

Two perspectives

- **Natural Language Understanding**

  Make limited inferences based on the natural language text.

  information extraction

- **Data Mining**

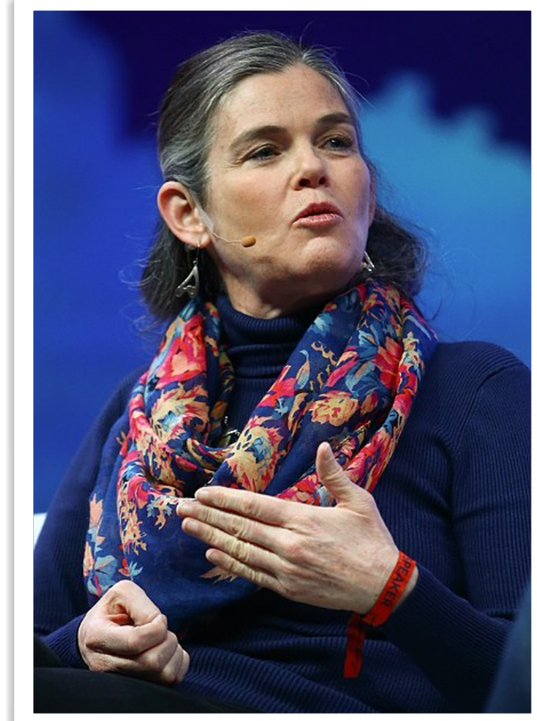
  Discover and extract interesting patterns in the text data.

  topic modelling

Adapted from Zhai and Massung (2016)

JEOPARDY!®

This Stanford University alumna co-founded educational technology company Coursera.



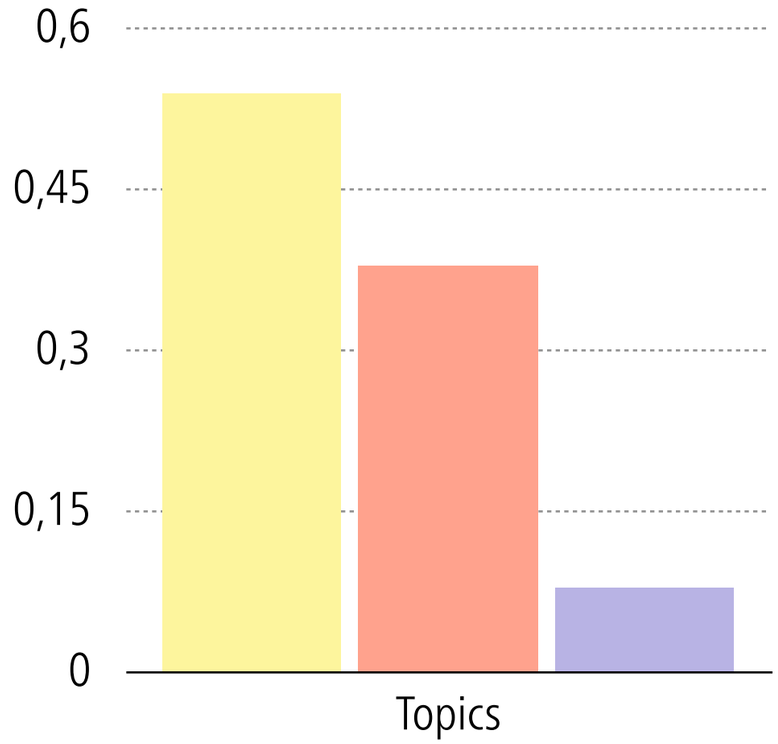
Collision Conf, CC BY 2.0, via Wikimedia Commons

SPARQL query against DBPedia

```
SELECT DISTINCT ?x WHERE {
  ?x dbp:education dbr:Stanford_University.
  dbr:Coursera dbp:founder ?x.
}
```

Topic models

How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes.



Source: Blei (2012)

Text data is special

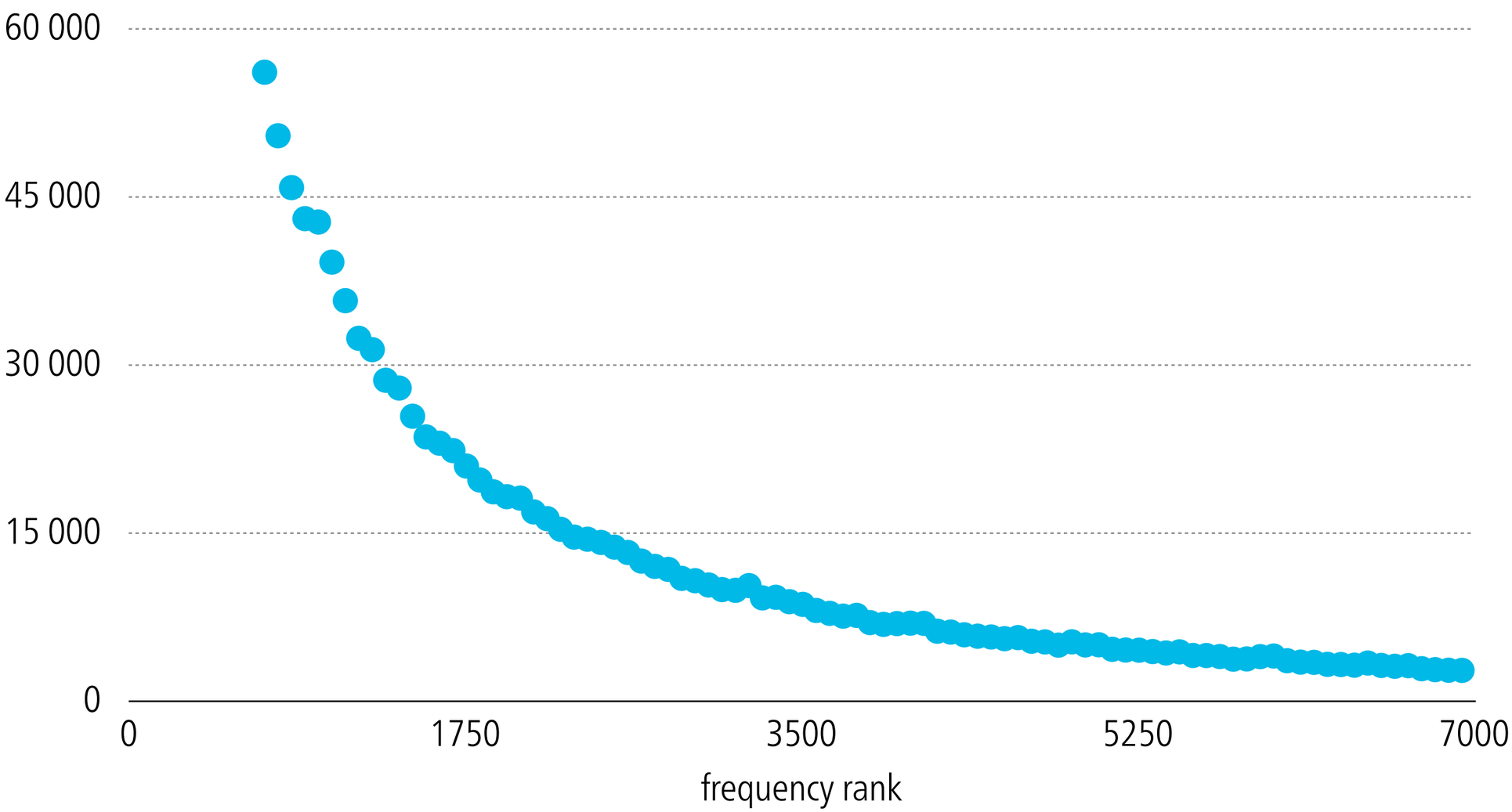
- Text data is generally produced by humans, rather than computers or sensors.

  contrast with e.g. image data

- Text data is generally meant to be consumed by humans, rather than computers or sensors.

  so-called unstructured data

Zipf's law



Word frequency data from the Contemporary American English Corpus

This session

- Introduction to text mining (15 minutes)
- Course logistics (15 minutes)
- Some simple examples of text mining (15 minutes)

Course logistics

Intended learning outcomes

1. implement and apply text mining methods
2. analyse and summarise the results of text mining experiments
3. identify, formulate and solve text mining problems
4. clearly present and discuss the conclusions of her or his work

Course contents

- Topic 1: Information Retrieval
- Topic 2: Text Classification
- Topic 3: Text Clustering and Topic Modelling
- Topic 4: Word embeddings
- Topic 5: Information Extraction
- Text Mining Project (you!)

	Monday	Tuesday	Wednesday	Friday
W44		LEC Information Retrieval	LAB Information Retrieval	LAB Information Retrieval
W45	LEC Text Classification	Individual Supervision	LAB Text Classification	LAB Text Classification
W46	Individual Supervision	LEC Clustering & Topic Modelling	LAB Clustering & Topic Modelling	LAB Clustering & Topic Modelling
W47	LEC Word Embeddings	Individual Supervision	LAB Word Embeddings	LAB Word Embeddings
W48	Individual Supervision	LEC Information Extraction	LAB Information Extraction	LAB Information Extraction
W49	Individual Supervision	LEC Project kick-off	Individual Supervision	Individual Supervision
W50	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W51	Individual Supervision	Individual Supervision	Individual Supervision	
W02	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision

Examination

	Computer labs	Project
ECTS credits	3 credits	3 credits
To be done	in pairs	individually
Deliverables	notebooks + diagnostic test	written project report
Grading	Pass/Fail	ECTS, U345

Detailed information on your course website

https://www.ida.liu.se/~732A81/

https://www.ida.liu.se/~TDDE16/

Student feedback and changes

- Students seem to be generally happy with the course. The 2021 session received favourable ratings, but the response rate was low.

  732A92: 4.50 (6/38), TDDE16: 4.56 (9/58)

- For this session, we changed the format of the diagnostic test:
  - assessment in pairs → individual assessment
  - one assessor → two assessors

This session

- Introduction to text mining (15 minutes)
- Course logistics (15 minutes)
- Some simple examples of text mining (15 minutes)

Some simple examples

Launch on Binder