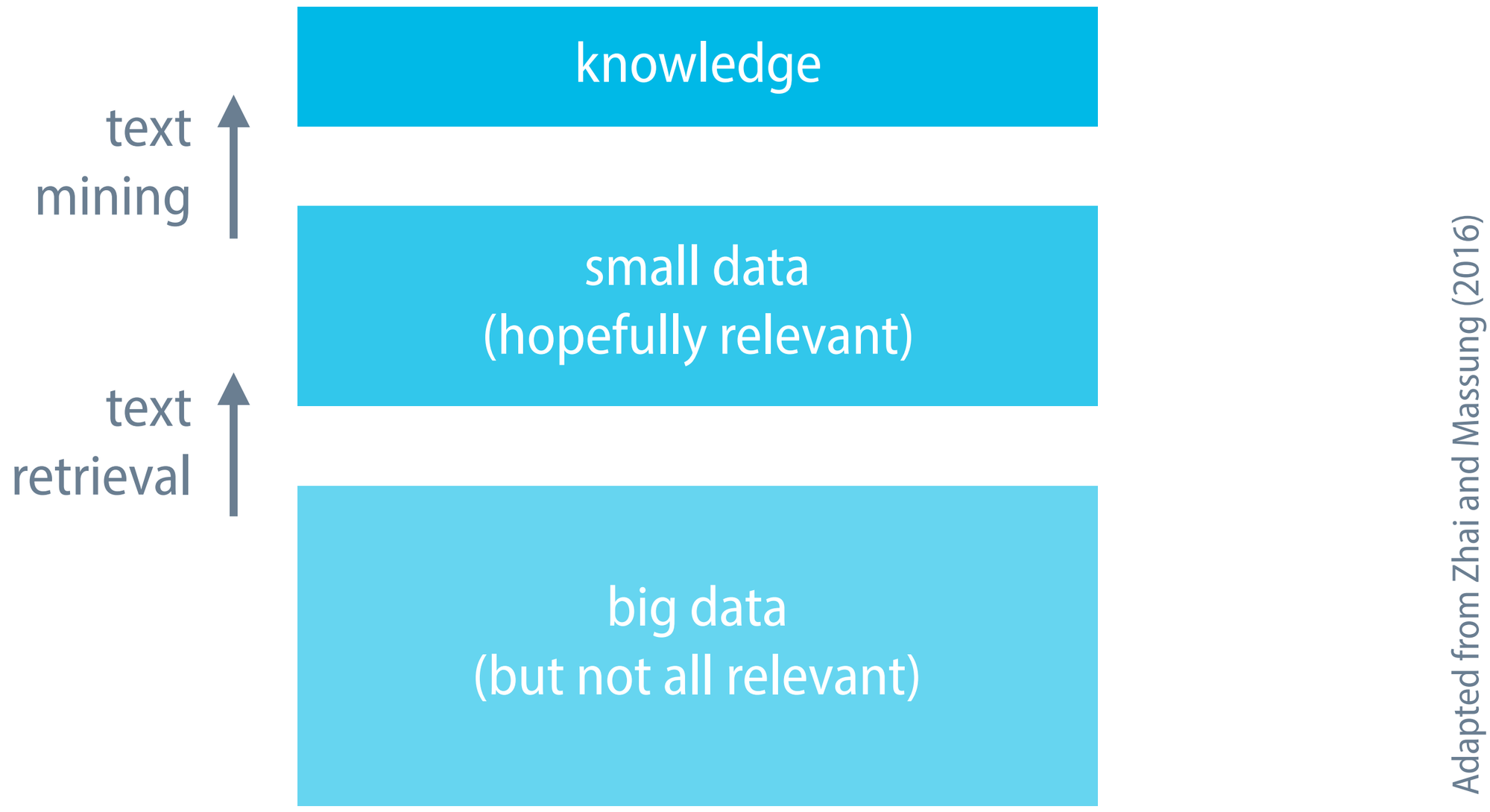
TDDE16/732A92 Text Mining (2019)

Introduction

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

Text retrieval and text mining



The Google Search index contains hundreds of billions of webpages and is well over 100,000,000 gigabytes in size.

Google, How Search Works

Text data is special

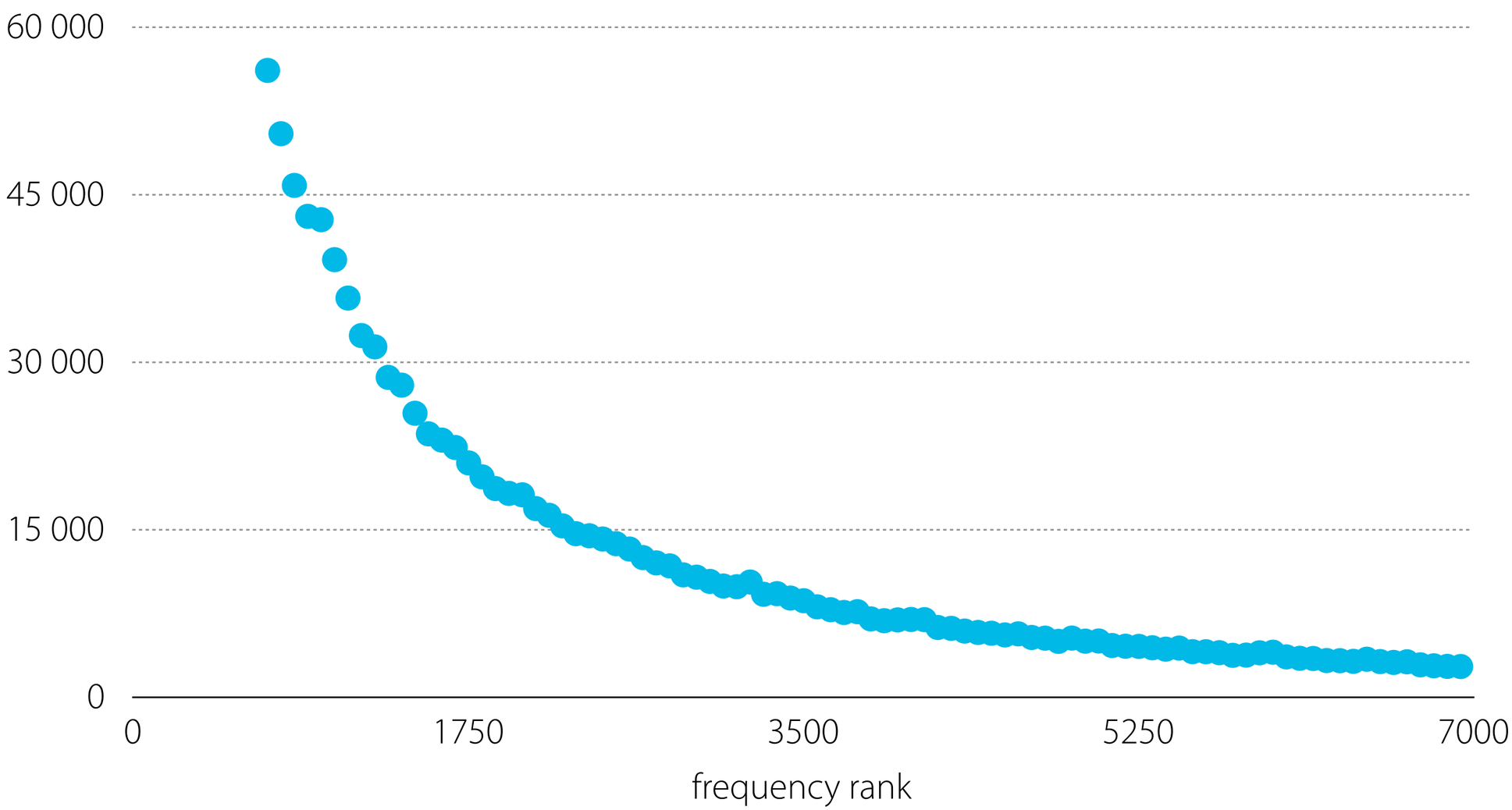
- Text data is generally produced by humans, rather than by computers or sensors.

  contrast with e.g. image data

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

  so-called unstructured data

Zipf’s law

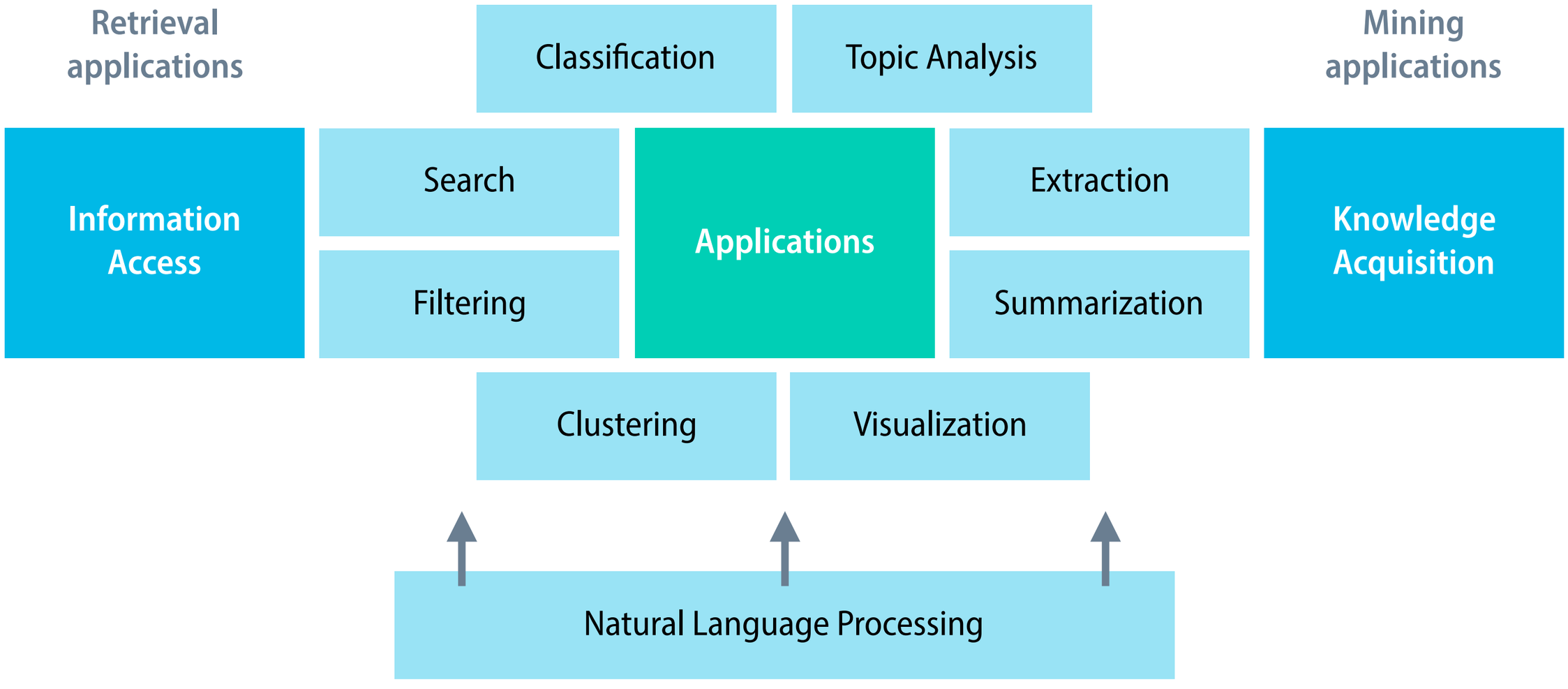


Word frequency data from the Contemporary American English Corpus

Typical applications of text mining

- **Search.** Take a user's query and return relevant documents.
- **Filtering.** Filter a stream of incoming documents.
- **Classification.** Sort documents into predefined categories.
- **Clustering.** Discover groups of similar text documents.
- **Topic Analysis.** Identify topics in a document collection.
- **Visualization.** Visually display patterns in text data.
- **Information Extraction.** Extract entities and relations between them.
- **Summarization.** Generate a summary of a document collection.

Conceptual framework for text mining



Adapted from Zhai and Massung (2016)

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)

Two perspectives

- **Natural Language Processing**
  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 alumnus co-founded educational technology company Coursera.

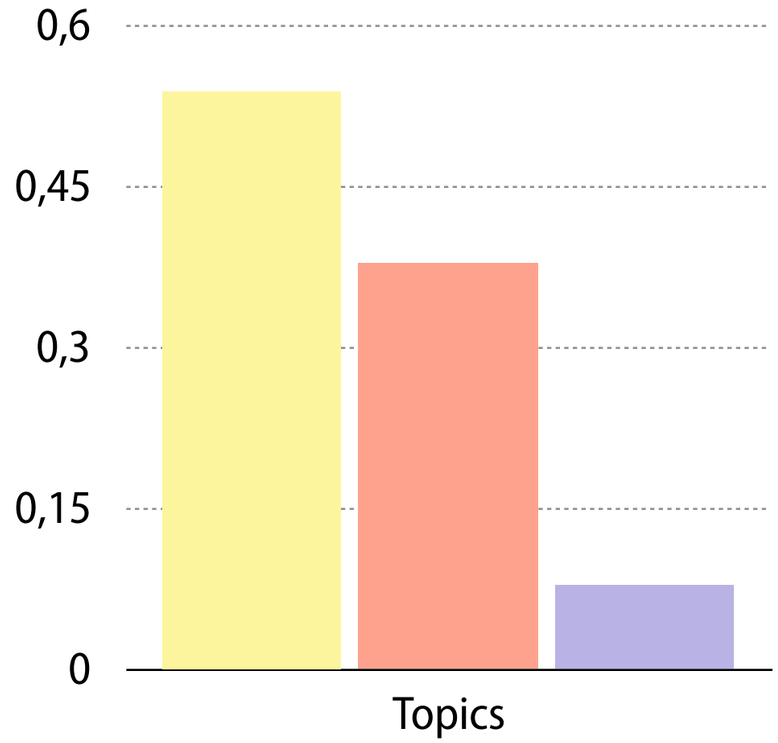
Source: MacArthur Foundation

SPARQL query against DBPedia

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:almaMater dbr:Stanford_University.
  dbr:Coursera dbo:foundedBy ?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)

Topic models

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Three stages

- Retrieving and textual data

  Information Retrieval

- Analysing the linguistic structure of the data

  Natural Language Processing

- Building statistical models from the data

  Statistical Modelling

Course organisation

Course outline

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

	Monday	Tuesday	Wednesday	Friday
W45	LEC Course introduction	LEC Information Retrieval	LAB Information Retrieval	Individual Supervision
W46	Individual Supervision	LEC Text Classification	LAB Text Classification	Individual Supervision
W47	Individual Supervision	LEC Clustering and Topic Analysis	LAB Clustering and Topic Analysis	Individual Supervision
W48	Individual Supervision	LEC Natural Language Processing	LAB Natural Language Processing	Individual Supervision
W49	Individual Supervision	LEC Information Extraction	LAB Information Extraction	LEC Project kick-off
W50	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W51	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W02		Individual Supervision	Individual Supervision	Individual Supervision
W03	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision

Examination

	Computer labs	Text Mining Project
ECTS credits	3 credits	3 credits
to be done	in pairs	individually
grading	Pass/Fail	U345, ECTS
form of hand-in	notebooks	written project report

Changes compared to previous session

- The 2018 session received very favourable ratings.

  732A92: 4.20 (5/26), TDDE16: 4.90 (10/34)

- In the 2019 session, we put even more focus on the project:
  - more but shorter labs, including new lab on classification
  - expanded project instructions
  - more time slots for individual feedback

TDDE09
Course Information
Syllabus
Examination
Timetable
All Messages
Contact

MATERIALS
Lectures
Labs
Project

INTERNAL
IDA internal
Student Pages
Emergency

TDDE09 Natural Language Processing (6 ECTS)

VT1 2019

Welcome to the course website for TDDE09 Natural Language Processing!

Natural Language Processing (NLP) develops techniques for the analysis and interpretation of natural language – a key component of smart search engines, personal digital assistants, and many other innovative applications. The goal of this course is to provide you with a theoretical understanding of and practical experience with the advanced algorithms that power modern NLP. The course focuses on methods that involve machine learning on text data.

Latest News...

2019-01-14 Welcome to the course!

The course website has now been updated for the 2019 session. (Those parts which will be updated during the session are clearly marked as such.) The first lecture will take place on Monday 2019-01-21 08:15-10 in U14.

Page responsible: **Marco Kuhlmann**
Last updated: 2019-01-14

Please visit your course website!

https://www.ida.liu.se/~732A92/

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

Example projects

- topic classification for cooking recipes
- topic analysis for the TV series *Friends*
- mood classification of songs based on lyrics
- predicting gender and age from blogs
- sentiment classification of Amazon reviews