

732A81/TDDE16 Text Mining (HT2024)

Course Introduction

Marcel Bollmann

Department of Computer and Information Science (IDA)



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Meet the staff!



Marcel Bollmann



Riley Capshaw

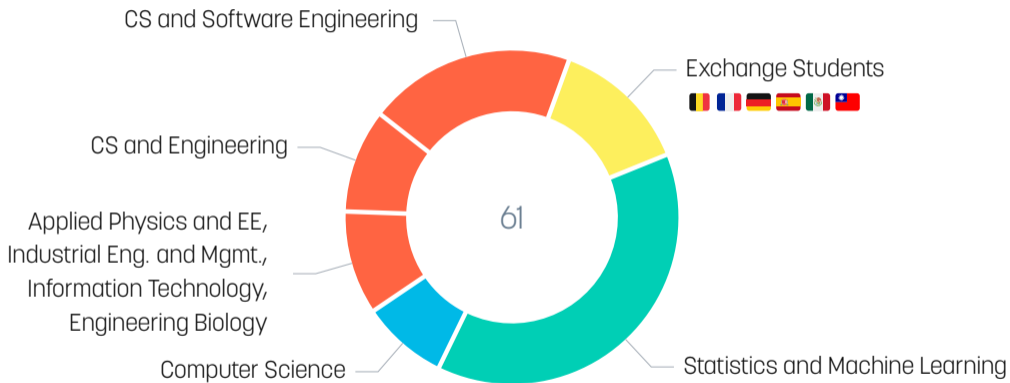


Noah-Manuel Michael



Romina Oji

Meet your fellow students!



This introduction session

1. Introduction to text mining
2. Course logistics

Introduction to Text Mining



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

Two functions of text mining

1. Accessing information

Enable the user to quickly access relevant information.

- Search engines
- Recommender systems

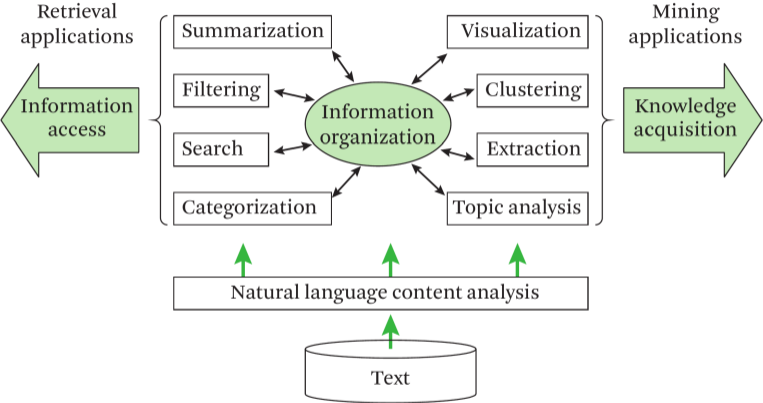
2. Extracting knowledge

Enable the user to acquire knowledge “hidden” in text.

- Information extraction
- Topic analysis

Zhai and Massung (2016)

Conceptual framework



Zhai and Massung (2016)

Search

The screenshot shows a Google search interface. The search bar contains the text "what is text mining used for". Below the search bar are navigation links for "All", "Images", "Videos", "News", "Maps", "Shopping", and "Settings". The location is set to "Sweden" and "Safe search" is set to "moderate". The search results are displayed in a list format. The first result is a snippet titled "Text mining" with a brief description: "Text mining, text data mining or text analytics is the process of deriving high-quality information from text. It involves 'the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.' Written resources may include websites, books, emails, reviews, and articles. [Wikipedia](#)". Below this snippet is a "Share Feedback" button. The second result is a link to "https://www.ibm.com > topics > text-mining" with the title "What is Text Mining? | IBM". The description for this result is: "Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights." The third result is a link to "https://monkeylearn.com > text-mining" with the title "What Is Text Mining? A Beginner's Guide - MonkeyLearn". The description for this result is: "Text minina (also known as text analysis). Is the process of transformina unstructured text into". A "Share Feedback" button is also present at the bottom right of the search results area.

what is text mining used for

Privacy, simplified. ▾

All Images Videos News Maps Shopping Settings

Sweden ▾ Safe search: moderate ▾ Any time ▾

Text mining

Text mining, text data mining or text analytics is the process of deriving high-quality information from text. It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." Written resources may include websites, books, emails, reviews, and articles. [Wikipedia](#)

[Share Feedback](#)

<https://www.ibm.com > topics > text-mining>

What is Text Mining? | IBM

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

<https://monkeylearn.com > text-mining>

What Is Text Mining? A Beginner's Guide - MonkeyLearn

Text minina (also known as text analysis). Is the process of transformina unstructured text into

[Share Feedback](#)

Filtering

YOU HAVE A DONATION OF 2,000,000.00 GBP OF COVID19 SUPPORT

▼ Von: "UNCO"  

--

YOU HAVE A DONATION OF 2,000,000.00 GBP OF COVID19 SUPPORT FROM UNITED
WORLD CHARITY ORGANIZATIONS, FOR CHARITABLE WORK IN YOUR COUNTRY,
TO RECEIVE YOUR DONATION AMOUNT KINDLY SEND YOUR FULL DETAILS TO VIA
unitedworldcharityorg@gmail.com



Recommender systems

The screenshot shows the Goodreads website interface. At the top, there's a navigation bar with the Goodreads logo, 'Home', 'My Books', 'Browse', 'Community', and a search bar. Below this, the page title is 'Recommendations > Fiction Genre'. A text block explains that recommendations are based on books added in the genre and that other readers with similar interests have enjoyed them. It includes a link 'How to improve your recommendations...' and a timestamp 'updated: 0 minutes ago'. A 'View: covers | list' link is also present. The main content area displays a grid of book covers. Each cover has a 'Want to Read' button and a star rating system. A right-hand sidebar contains a 'Recommendations by Shelf' section (empty), a 'Recommendations by Genre' section with a table, and a 'More Actions' section with several options.

100 New Horror Books Picks for every kind of reader >

goodreads Home My Books Browse Community Search books

Recommendations > Fiction Genre

Here are some books we recommend based on the books you've added in this genre. Other readers with similar interests have enjoyed them. [How to improve your recommendations...](#)

updated: 0 minutes ago

View: covers | list

Book Title	Author	Want to Read	Rating
ANDY WEIR		Want to Read	★★★★★
PROJECT HAIL MARY		Want to Read	★★★★★
A PSALM FOR TIME		Want to Read	★★★★★
WILD BUILT	BODY EXAMINERS	Want to Read	★★★★★
THE THREE-BODY PROBLEM	LIU CUNGLIANG	Want to Read	★★★★★
THE PAPER MONSOON	KEN LIU	Want to Read	★★★★★
AXIOMATIC	GREG EGAN	Want to Read	★★★★★
SUSANNA CLARKE			
Douglas Adams			
ASIMOV			
RECURSION			
CARL SAGAN			

Recommendations by Shelf

You have no recommendations based on your bookshelves yet.

Recommendations by Genre

Fantasy	50
Fiction	50
Poetry	50
Science Fiction	50

More Actions

- Recommendations from Members
- Give Recommendations
- Ask for Recommendations
- Books Marked as 'Not Interested'

Categorization: Sentiment analysis



I love it so much! The mic works great!!!! I use it for online live classes, cosplay, and to look cute!! The lightup feature really works great! The app also works great too! The sound sounds amazing too! I just wish it had a case for when I travel.

positive

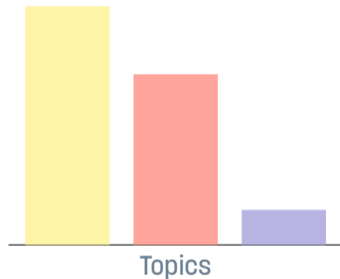
Not durable. The cord came apart from the audio adjuster. The saddest part is that happens only two months after it was purchased, and no force was applied. Definitely, I will not purchase and I do not recommend the item.

negative

Adapted from Amazon

Topic analysis

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**.



Blei (2012)

Summarization



The screenshot shows the Goodreads page for the book "Spring Snow" by Yukio Mishima. The page includes a search bar, a book cover, a "Buy on Amazon" button, and a star rating. The main text is a review by user Alex Fonseca, dated February 27, 2018. The review discusses the novel's themes of class divisions, western influence, and reincarnation. It mentions that the main character is the son of a wealthy family and that the novel is the first volume of a tetralogy. The review also notes that the book is well-written and engaging, and that the reviewer is uncertain if it compels them to read the entire tetralogy but mentions that the second volume, "Runaway Horses," is highly rated.

goodreads Home My Books Browse Community Search books

Search review text Filters

Displaying 1-10 of 128 reviews

February 27, 2018

Alex Fonseca 100 reviews 128 reviews

More to read

Buy on Amazon

☆☆☆☆ Rate this book

Spring Snow

Mishima (1925-1972) was a classic Japanese author. He was a fierce anti-communist who led a band of rebels hoping to restore the Emperor. He committed that suicide when the plot failed. His best known work is a tetralogy. The Sea of Fertility of which this book is the first volume.

Class divisions and changing values in Japan due to western influence are major themes. The main character is the son of a very wealthy family. How wealthy? They have 40 servants and the boy doesn't know all these names even though some of them who have worked there for years. The family lives in a multi-story compound which includes a western-style home in which they entertain and occasionally have western meals. The boy's own room is in a Japanese-style house but it's decorated with western furniture. His mother often dresses and wears her hair in western style. And yet, he has a friend at school whose family, he feels is, more "western in outlook" than his, despite their family's lack of western trappings in furniture, food and dress.



The boy is very good-looking, delicately, melancholy, icy and emotionless. He's bright but he figures he'll get into a university for rich kids because he's not wanting time studying for exams to get into an academically prestigious school, he's 18 when the story starts. The end of the Russo-Japanese War, 7 years ago, is a key backdrop to the story. So we know it's around 1902.

So the boy's family has money, but it's not one of Japan's traditional 28 noble families, like the one next door. So his father creates an alliance between those two families and the boy spends much time at the neighboring residence absorbing the noble sensibilities. His father supplies the money, the other family supplies the prestige.

The plot involves around a love story between this boy and the daughter of the neighboring household. They have known each other all their lives and she has loved him since they were children. But his feelings toward her are unrequited, still again, he's returned her and perhaps he doesn't care for her. Finally she gives up on him and becomes engaged to a son of a noble family, actually a member of the Emperor's household.

At this point (she's 21, he's 16), and after the engagement has been approved by the Emperor himself, finally he decides he loves her and begins to pursue her. They begin a sexual relationship and she becomes pregnant. It kind of any of this gets out, it would be the equivalent of a national scandal. When the boy's father learns what is going on, after spending his whole life teaching the emperor and the nobles, he says he is appalled, is pushing it away. Never having (that) a hand to his son before, he beats him with a good cane.

The difficult romance gives the author a chance to discuss the theme of the light of reason vs. the darkness of passions. There's also quite a bit of discussion about Buddhism and reincarnation. But we know all this can only end in tragedy. His friend counsels him that he is breaking his own almost, as if he seems to open outside.



There is good writing, such as this passage that I loved: "On a warm spring day a galloping horse was only too clearly a seeming animal of flesh and bone. But a horse racing through a wilderness became one with the very elements, wrapped in the whirling dust of the earth wind, the best embodied the joy breath of winter."

It's a good story, I don't know if it entices me to read the whole tetralogy, but the second volume in the series, Runaway Horses, is really highly rated as Spring Snow. (The other two are The Temple of Dawn and The Decisive of the Angels) (Probably the author's best known work in English is not part of the tetralogy, it's The Sailor Who Fell from Grace with the Sea.)

The review discusses Yukio Mishima's novel "Spring Snow," the first volume of the tetralogy "The Sea of Fertility." [...] **The review highlights the themes of reason vs. passion and incorporates elements of Buddhism and reincarnation. While the book is well-written and engaging, the reviewer is uncertain if it compels them to read the entire tetralogy** but mentions that the second volume, "Runaway Horses," is highly rated.

Source: Jim Fonseca on Goodreads and ChatGPT

Visualization

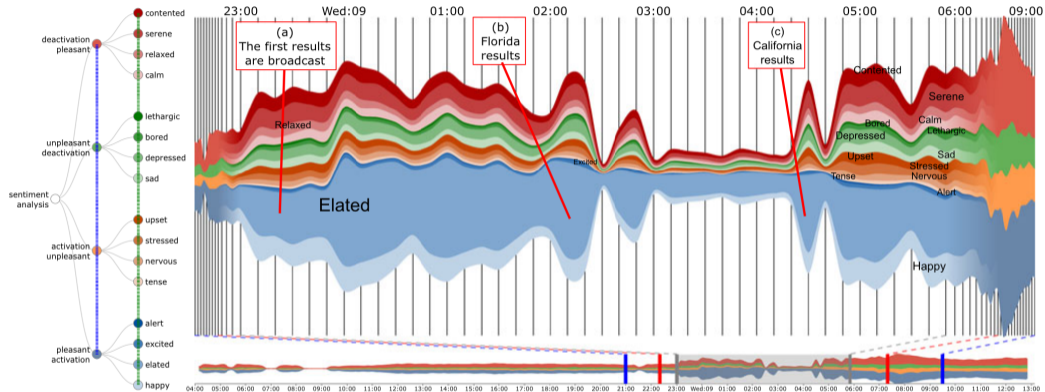


Fig. 19. Evolution of sentiments expressed in tweets on the US 2016 presidential election day: Results.

Cuenca et al. (2018)

Visualization

- Not covered in this course, but...

👍 Many examples to be found on:

<https://textvis.lnu.se/>

👍 Information Visualization (iVis)

Group here at LiU

Text Visualization Browser

A Visual Survey of Text Visualization Techniques (IEEE PacificVis 2015 short paper)
Provided by ISOVIS group

[About](#) [Summary](#) [Add entry](#) [Other surveys](#)

The screenshot displays the Text Visualization Browser interface. On the left is a sidebar with the following sections:

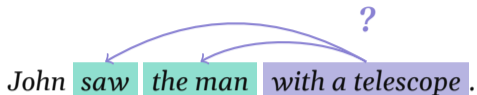
- Techniques displayed:** 440
- Search:** A search input field with a clear button (X).
- Time filter:** A range from 1976 to 2019 with a histogram below it.
- Analytic Tasks:** A grid of icons for various tasks like summarization, classification, and clustering.
- Visualization Tasks:** A grid of icons for tasks like visualization, interaction, and navigation.
- Data:** A field for the data source.

The main area on the right is a grid of 440 small thumbnail images, each representing a different text visualization technique. These thumbnails show a variety of visualizations including word clouds, network graphs, treemaps, and various charts.

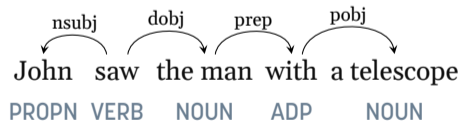
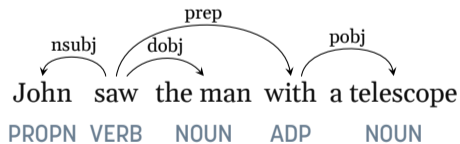
Text data is special

- Text data is generally **produced by humans**, rather than computers or sensors.
 - Contrast with, e.g., image data!
- Text data is generally **meant for humans**, rather than computers or sensors.
 - So-called unstructured data
- Language is often **subjective** and **ambiguous**.

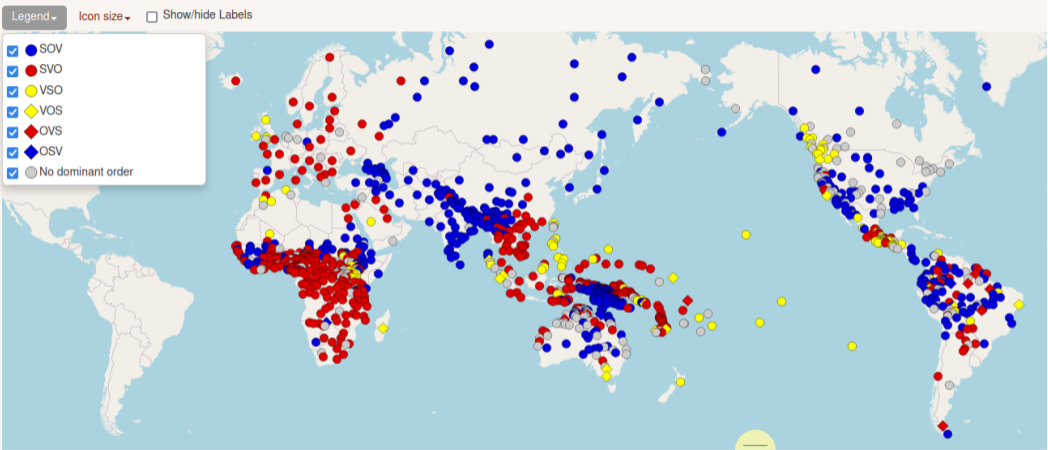
Language is ambiguous



- **Implicit** structure...



Languages are different



The World Atlas of Language Structures

Course logistics



Background survey



<https://forms.office.com/e/6NWtispcF1>

Course topics and timeline

Week	Topic	Scheduled
45	Information Retrieval	Lecture + 2 Lab sessions
46	Text Classification	Lecture + 2 Lab sessions
47	Information Extraction	Lecture + 2 Lab sessions
48	Clustering & Topic Modelling	Lecture + 2 Lab sessions
49	Text Summarization	Lecture + 2 Lab sessions
50	Project Work	Kick-off session
*	Project Work	Individual supervision





Course website






<https://liu-nlp.ai/text-mining/>

Examination

Computer labs

-  3 credits
-  Notebooks (Done in pairs)
-  Reflection (Done individually)
-  Pass/Fail

Project

-  3 credits
-  Written project report
(Done individually)
-  732A81 ECTS
TDDE16 U345

Sign up for a lab group!

732A81



TDDE16



- Please sign up by **Wednesday (EOD)**! I will transfer the groups to Lisam after that.
 - You won't be able to see the lab submissions in Lisam before that.

Working on the labs

- Labs come in form of **Jupyter Notebooks**.
 - We have tested that they run on the lab computers.
- If you want to run them on your own computer, we provide a list of required packages on the course website.

```
1 pip install -r requirements.txt
```

- We **cannot provide support** if you run into technical issues on your own computers!

Reflection questions

- Should be done **after** completing the notebooks.
- Written & submitted **individually**.
 - You may discuss with your lab partner, but the text you produce must be your own!
- Answers should be brief; **max. one page** total.

Assignment due dates



Mondays, 23:59,
the week after the labs







18.01.2024
(last examination date)




- Meeting the first due date gives you **timely feedback** (*within a week*) and an extra assessment opportunity.
- **We do not grade (re)submissions between the deadlines!**

Examination

Computer labs

-  3 credits
-  Notebooks (Done in pairs)
-  Reflection (Done individually)
-  Pass/Fail

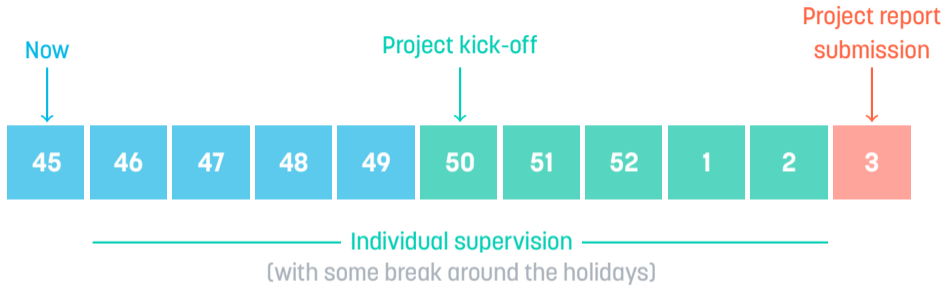
Project

-  3 credits
-  Written project report
(Done individually)
-  732A81 ECTS
TDDE16 U345

Project work – Knowledge requirements

- You are free to work on **a topic of your choice** within text mining.
 - Identifying and formulating a problem to work on is part of the project work.
- You can book **individual supervision meetings** with me throughout the teaching period to discuss ideas, questions, problems related to the project.
 - 15-minute time slots; please come prepared!
 - Booking link is on the course website.

Project timeline

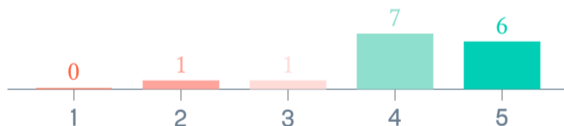


- Start brainstorming potential topics throughout the lectures & labs!

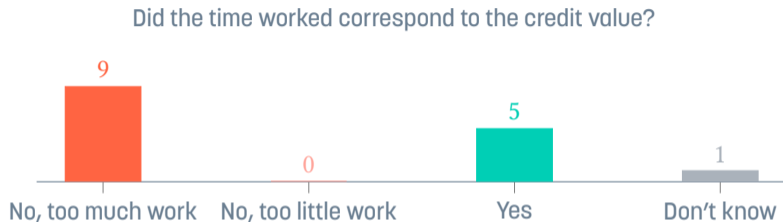
Previous course evaluation

- **103** students took the course in HT2023.
- **15** students submitted a course evaluation (→ 14.6%).

What is your overall evaluation of the course?



Criticism: Too much work!



- Revised project assessment criteria, more time for supervision this year
- Started to revise the labs, but work is ongoing

Questions?

In person

- During the session
- In the break
- In the lab

Asynchronously

- Email

Project-related

- Email
- Schedule a meeting via the booking link on the website

✉ marcel.bollmann@liu.se — marbo59

