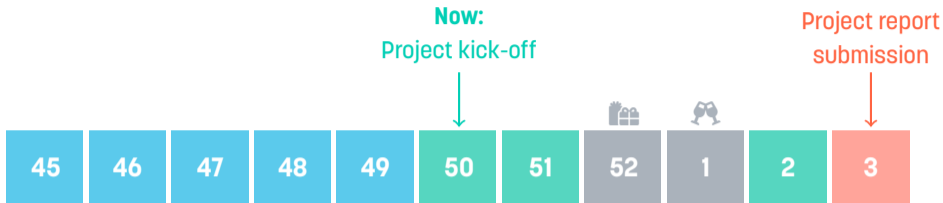


# Project kick-off

Marcel Bollmann

Department of Computer and Information Science (IDA)

# Project timeline



- Please use the chance to get **individual supervision!**
  - 19.+20.12.: Only via Zoom or e-mail
  - 21.12.–6.1.: Only via e-mail (*but please don't expect a fast response*)
  - From 7.1.: Supervision bookings possible again

# Plan for today

1. Formal Requirements

2. Example Projects

3. Practical Tips

- Project Structure
- Getting Help

4. Your Questions

# Formal Requirements



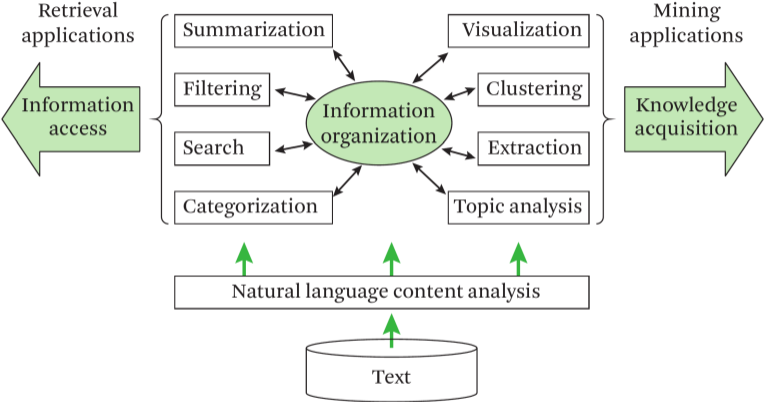
## Purpose of the project

The project module primarily tests to what extent you are able to

1. **identify, formulate and solve problems** *within the area of text mining*, and
2. **clearly present and discuss the conclusions** of a project work.

From the course memo

# Is your problem “within the area of text mining”?



Zhai and Massung (2016)

## Knowledge requirements

- You identify and formulate a **substantial text mining problem** with some help from a teacher.
- In working on your problem, you
  - implement and apply **suitable text mining methods**,
  - analyse experimental results with **appropriate evaluation methods**, and
  - summarise them with **well-developed judgements**.
- You clearly present and discuss the conclusions of your work.

From the course memo

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



# Submission

- The project report is submitted as a single PDF file **via Lisam**.
- Additionally, I will prepare a **project registration form** for you to submit:
  - Your project title & abstract. (*can be published as example projects in coming years*)
  - A link to your code repository.
    - Either on Github or on LiU's Gitlab server.
    - For private repos, you must give access to mbollmann (Github) or marbo59 (Gitlab).
  - A declaration of whether and how you used generative AI in preparing your report.

## Example projects from previous years



# Sentiment Analysis of IMDB Movie Reviews

- How well can we **predict the sentiment** of movie reviews?
  - Used the existing *IMDB movie reviews* dataset
- Trained and evaluated different types of classifiers.
  - Naive Bayes, Support Vector Machine, BERT, RoBERTa
- Performed a manual **error analysis** of reviews misclassified by BERT/RoBERTa.
  - Gain additional insights that could help improve models further

## Explainable Authorship Attribution using BERT

- Can we **identify an author** from the way they write?
  - Found an *author identification dataset* on Kaggle
- Trained and evaluated different types of classifiers.
  - Naive Bayes, Logistic Regression, BERT
- Which features does the BERT model consider **“characteristic”** for each author?
  - Used *integrated gradients method* to extract features that were most important for the BERT classifier when making its predictions

## Analyzing how the Refugee Crisis Affected Parliament Speeches using BERTopic

- Did the Syrian refugee crisis cause an increase in **immigration-related topics** being discussed in the Swedish parliament?
- Analysed the change of topics over time, as modelled by **BERTopic**.
  - BERT-based *topic modelling* technique
- Evaluated the performance using a **coherence** metric as well as manual analysis.

## Generating free-text explanations with GPT-2/GPT-3

- Natural language inference (NLI) is the task of predicting whether two sentences **entail or contradict** each other.
  - *“An adult holds a stick.”* contradicts *“An adult walks away empty-handed.”*
- Can we **generate free-text explanations** for these predictions?
  - *“Holding a stick implies using hands, so it is not empty-handed.”*
- Compared **two generative models** using different techniques and performed **human evaluation** on the generations.
  - GPT-2 with fine-tuning vs. GPT-3 with few-shot prompting

## Family tree extraction for Tolkien's world

- Can we automatically **extract family trees** for Tolkien's characters using encyclopedia entries about them?
  - Scraped the *Lord of the Rings Wikia* site
  - Implies **detecting** character names and the **relations** between them
- Evaluated using the **“infoboxes”** section of each character page as ground truth.

## Project abstracts from previous years

- A selection of **project abstracts** from previous years is on the course website.

### Note

It is perfectly fine to work on the same topic as a project from previous years!

- Perhaps you can find a better approach, or do a different analysis, or...
- You are unlikely to have *exactly* the same ideas as another student.



# Practical Tips



## A common project structure

1. Identify your problem (ca. 8 hours)
2. Design your approach (ca. 32 hours)
3. Evaluate your approach (ca. 32 hours)
4. Produce your report (ca. 16 hours)

## Phase 1: Identify your problem

- Is there a specific **task** you want to work on?
  - *“relation extraction to construct family trees of Tolkien’s characters”*
- Is there a (limited-scale) **research question** you want to answer?
  - *“Did the Syrian refugee crisis cause an increase in immigration-related topics?”*
- One way to get inspiration is to **look for datasets** first.
  - [Kaggle](#), [HuggingFace Datasets](#), [Riksdagens öppna data](#), many “shared tasks” such as [RepEval 2017](#), ...

## Phase 2: Design your approach


1. Select one or more **datasets**.
2. Write code to **process** the data and prepare models/experiments.
  - Use existing libraries, e.g. [spaCy](#), [Gensim](#), [HuggingFace Transformers](#), ...
  - No requirement on the programming language!
3. Choose a method for **evaluating** your results.
  - Consider both evaluation *measures* as well as *baselines*.
  - Think about this **before** implementing your approach!

## Phase 2: Design your approach – tips & tricks (I)




- It's a good idea to **review previous work** related to your problem.
  - Search the [ACL Anthology](#) (*many NLP & text mining papers*)
  - Search via [Semantic Scholar](#)
- This can give you new ideas as well as pointers to existing **code and datasets!**
  - [Papers with Code](#) specifically collects academic papers with code implementations.
- Collect references as you go so you can **cite your sources** appropriately!
  - Includes code that you re-used.

## Phase 2: Design your approach – tips & tricks (II)

If your project uses **machine learning** in any way, consider:

-  **How to avoid machine learning pitfalls, by Michael A. Lones**
  - Before you start to build models
  - How to reliably build models
  - How to robustly evaluate models
  - How to compare models fairly
  - How to report your results

## Phase 2: Design your approach – tips & tricks (III)

- **Web scraping** is the process of automatically extracting data from websites.
- Involves fetching a web page, parsing it, and finally extracting data from it.
  -  **BeautifulSoup**
  -  **Scrapy**
  -  **Trafilatura**
- May violate the terms of use of some websites and/or constitute copyright infringement.
  - Countermeasure include blocking the scraper's IP address.

Dwarf - Dwarf Fortress Wiki — Mozilla Firefox

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Page **Discussion** [Read](#) [Edit](#) [View history](#)

**DWARF FORTRESS WIKI**

Main page  
Centralized discussion  
Community portal  
Recent changes  
Announcements  
Random page  
Random page by namespace  
Help

Tools  
What links here  
Related changes  
Special pages

**v50 Steam/Premium information for editors**

- v50 information can now be added to pages in the **main namespace**. v0.47 information can still be found in the **DF2014** namespace. [See here](#) for more details on the new versioning policy.
- [Use this page](#) to report any issues related to the migration.

This notice may be cached—the current version can be found [here](#).

# Dwarf


xTATTEREDx · +FINE+ · \*SUPERIOR\* · =EXCEPTIONAL= · =MASTERWORK=

**v50.10** · v0.47.05

This article is about the current version of DF.  
Note that some content may still need to be updated.  
[More Info](#) · [V](#)

*A short, sturdy creature fond of drink and industry.*

Dwarves (singular: Dwarf) are "intelligent" alcohol-dependent humanoid creatures that are the featured race of



Inspector Console Debugger Network Style Editor Performance Memory Storage Accessibility Application AdBlock

Search HTML

```

<div id="page-quality-rating" style="text-align: right;margin: 0.8em 0;font-size:75%;"/></div>
<table class="infobox" style="float: right;border: 1px solid #aaa;margin: 0 0 1em 6px;border-spacing: 0;width: 200px;"></table>
<table class="version-table" style="text-align: center;margin: 3px auto 5px;padding: 0.25em;border: 2px solid #add8e6;background: #f5f5f5;min-width:27em;"></table>
<div>
  <div>
    <div>
      A short, sturdy creature fond of drink and industry.
    </div>
  </div>
  </div>
  <div class="thumb tright"></div>
</div>

```

Filter Styles · Filter Styles · Filter Styles

element { }

Inherited from div#mw-content-text

@media screen {

div#content-ltr {

direction: ltr;

font-family: sans-serif;

font-size: 14px;

line-height: 22.4px;



## Web scraping example: Trafilatura

```
1 >>> from trafilatura import fetch_url, extract
2 >>> downloaded = fetch_url("http://dwarffortresswiki.org/index.php/Dwarf")
3 >>> text = extract(downloaded, include_tables=False)
4 >>> print(text)
5 - A short, sturdy creature fond of drink and industry.
6 This is a masterfully-designed engraving of a Dwarf and a battle axe.
7 Dwarves (singular, Dwarf) are "intelligent", alcohol-dependent, humanoid
8 creatures that are the featured race of fortress mode, as well as ...
```

## Phase 3: Evaluate your approach

- **Run your experiments** and **interpret your results**.
  - Quantitative measures, *e.g.*, accuracy, precision, recall, topic coherence, perplexity, ...
  - Qualitative analysis, *e.g.*, human evaluation, manual inspection of interesting test cases, ...
- Remember that most evaluation methods are **relative**!
  - Can't conclude that “X works well”, but “X works *better* or *worse* than Y”
- Consider [!\[\]\(467d80e979964f7f8c752fb22248b5b7\_img.jpg\) Colab](#) or [!\[\]\(b71552d33dbf62adf5e5199a70ee02bf\_img.jpg\) Kaggle](#) if you need computing resources.

## Phase 4: Produce your report

- Use the **required template** and revisit the **suggested structure**.
- Consider **drafting sections** already during the previous phases!
  - e.g. Data & Method during phase 2
- Consider **how to present** your project in a clear and understandable way.
  - e.g. giving examples of data or inputs/outputs, summarizing important figures in a table, visualizing complex concepts in a figure, visualizing results in a graph/plot, ...

## How to get help

- Pitch your project idea to me!
- I am offering **one-on-one meetings** for you to get feedback.
  - Book a time slot through the link on the website
- I am also available **via e-mail**.
  - Use whichever option works best for you and the question(s) you have!

# Your Questions



“ *How can we find a project idea?* ”

- Check the **project abstracts** on the course website
- Check the **example projects** in this presentation
- Check slide **15**

“*Can we use methods or topics from outside the course?*”

- **Yes!** As long as it fits the **framework for text mining** and does natural language analysis in some way, you can work on any topic you want.

“Are we restricted to certain packages or frameworks?”

- **No!** You can use any tools, libraries, frameworks, and programming languages you want.



“*How much time to spend on coding vs. report writing?*”

- That's a difficult question to answer in general. See slide [14](#) for a rough estimate. Don't underestimate the report writing part, though, and ideally start drafting early!

“*How much data do we need? What if data collection takes a significant amount of time?*”

- You need enough data to draw **meaningful conclusions**. How much that is depends on what you want to do with it; ask me if unsure! There's certainly no need to work with huge datasets of millions of documents, though.
- Spending effort on gathering data can **contribute to a higher grade** for your project.

“*Are there any good web-scraping techniques?*”

- Yes! See the pointers on slide 19.

“ *What is the length expectation of the paper?* ”

- 3–5 pages of content in the ACL template. Please check the [project report formalities](#) for more details!

“ *What is the minimum amount of sources we should use?* ”

- There is no minimum requirement.
- For a **passing grade**, make sure to cite all sources of text, data, or ideas that you use in your project.
- For a **higher grade**, you need to cite scientific articles, but quality is more important than quantity. Citing one paper and connecting it really well to your own work is better than citing ten papers without any meaningful discussion.

“ *Does the project have to have scientific novelty?* ”

- **No!**

“What are the points that you specifically check for the highest grade?”

- Please check the **project report formalities** for more details on the assessment criteria. For the **highest grade**, your report should be very good in each of the three aspects mentioned there.
  - But keep in mind that *it's not a checklist*, and you don't need to achieve *everything* there to get the highest grade.
- If unsure, ask! Tell me that you are aiming for a higher grade and what your plans are to achieve this, and I can give you feedback on how well your plans fit the criteria.

