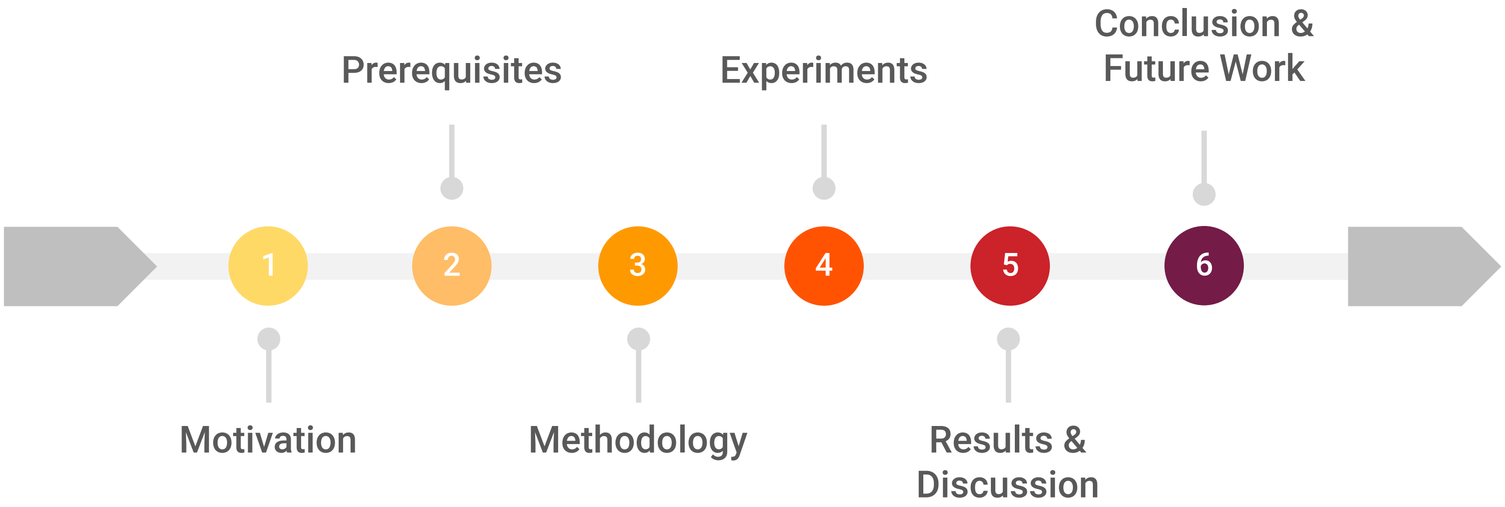
Master Thesis Recommendation Tool

Martin Stiles & Magnus Rushfeldt

Outline



1 Motivation

2 Prerequisites

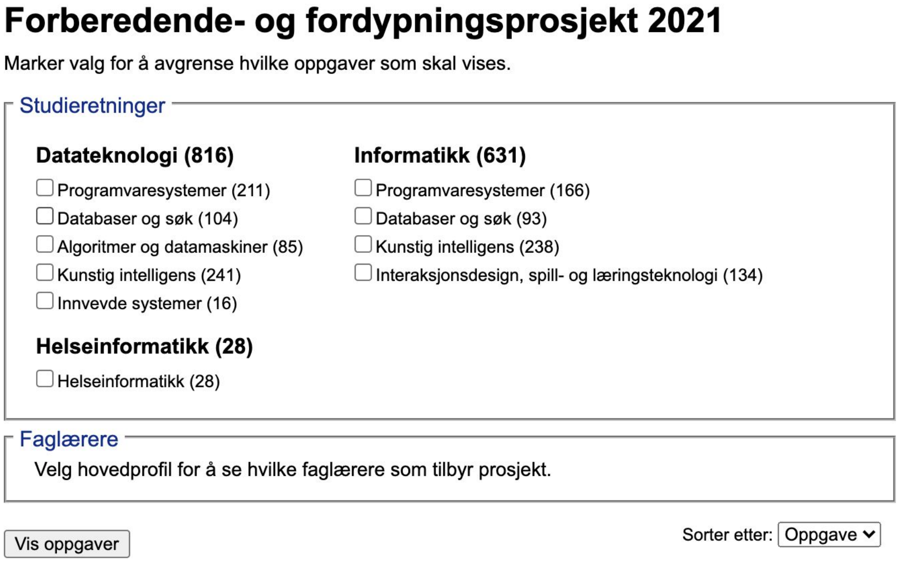
3 Methodology

4 Experiments

5 Results & Discussion

6 Conclusion & Future Work

Motivation



Forberedende- og fordypningsprosjekt 2021

Marker valg for å avgrense hvilke oppgaver som skal vises.

Studieretninger

Datateknologi (816)

□ Programvaresystemer (211)
□ Databaser og søk (104)
□ Algoritmer og datamaskiner (85)
□ Kunstig intelligens (241)
□ Innvevde systemer (16)

Helseinformatikk (28)

□ Helseinformatikk (28)

Informatikk (631)

□ Programvaresystemer (166)
□ Databaser og søk (93)
□ Kunstig intelligens (238)
□ Interaksjonsdesign, spill- og læringsteknologi (134)

Faglærere

Velg hovedprofil for å se hvilke faglærere som tilbyr prosjekt.

Vis oppgaver

Sorter etter: Oppgave

idi.ntnu.no/education/fordypningsprosjekt.php

- Messy interface
- No search function
- Limited filtering options
- Loads of duplicates
- Long descriptions

Our Proposal

Recommendations

Provide ideas for thesis topics based on user input

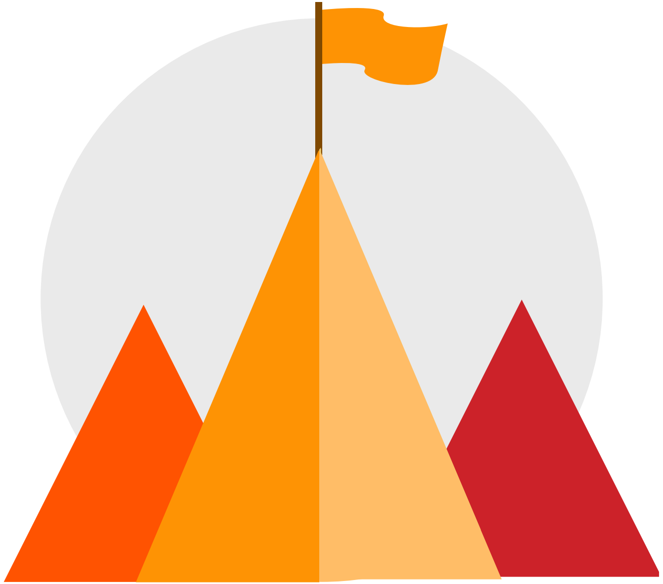
Summaries

Display the most important sentences of the theses

Keywords

A short list of words to represent a thesis topics

Milestones



01 Retrieve data

Scrape the thesis website

02 Clean data

Prepare scraped data for the models

03 Recommender

Create a method for recommendations of thesis topics

04 Summarizer

Create a simple method for summarization and keyword generation

05 Evaluation

Use real data from a survey to evaluate performance

06 Website

Make an interface for the methods

(Quick) Demo

We were unable to publish the website

😥

localhost:3000

Prerequisites

We'll assume everyone is somewhat familiar with these concepts

1 **Selenium / BS4**
Scraping

2 **NLTK**
Text manipulation: tokenizing, stop word removal, stemming, ...

3 **TextBlob**
Language detection

4 **TF-IDF vectorization**
Vectorization of textual objects

5 **Cosine similarity**
Compute distance between vectors

6 **Precision, Recall, F-score**
Measure performance

The Cold Start Problem

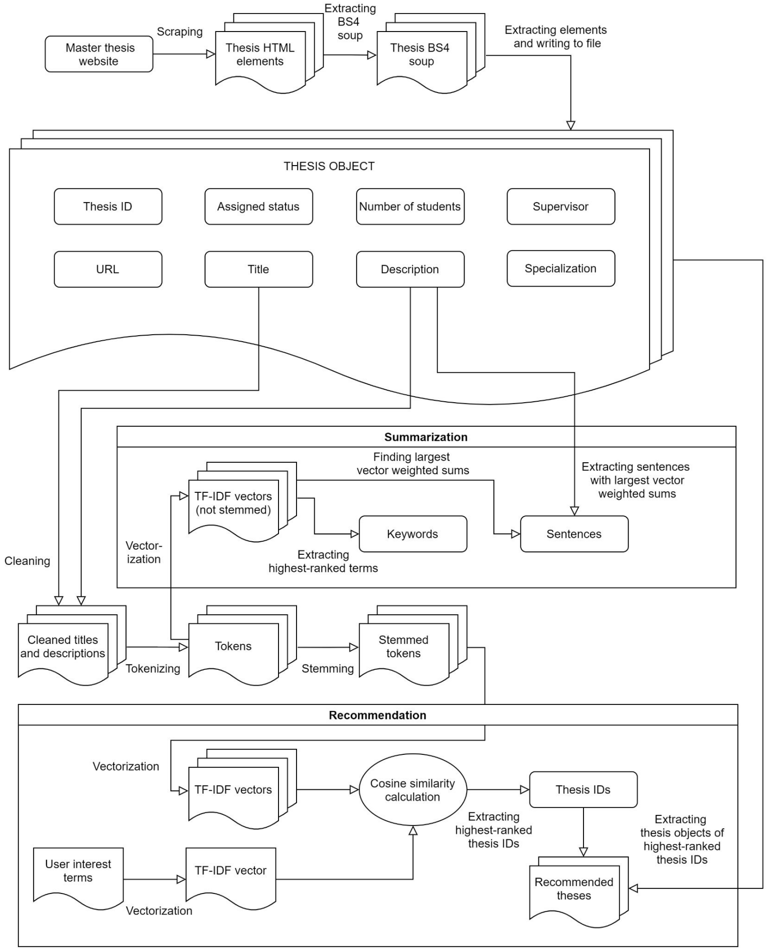
- Common in the starting phase of recommendation systems
- No user data to base the recommendations on

Narrative-driven Recommendation

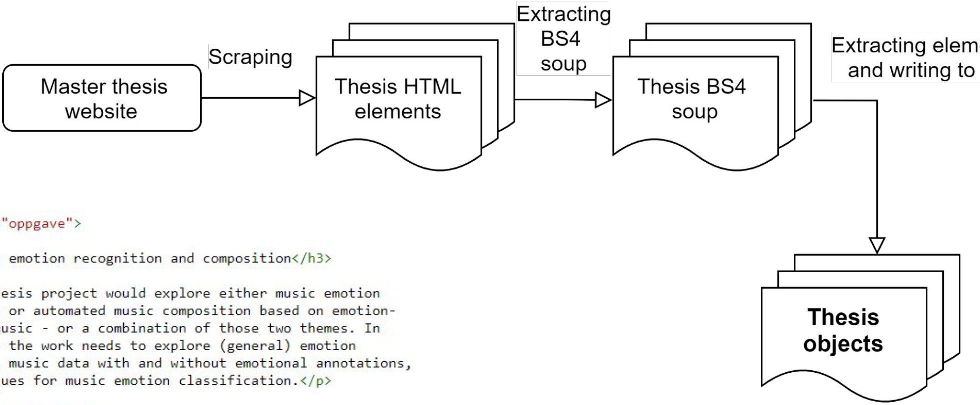
- We use narrative-driven recommendation
    - Use some kind of similarity measure + user data
- Focus on explicit user-supplied interests

The Pipeline

- Scraping
- Cleaning
- Tokenization
- Stemming
- TF-IDF vectors
- Summarization
- Recommendation

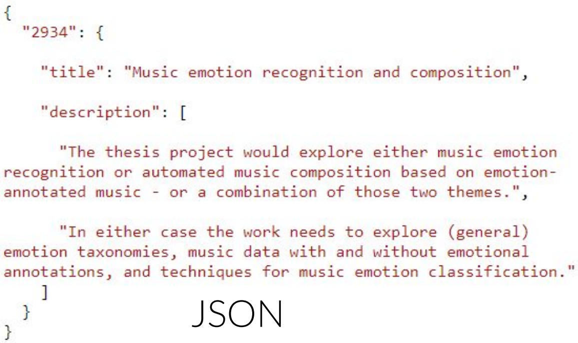


The Pipeline - Data Retrieval



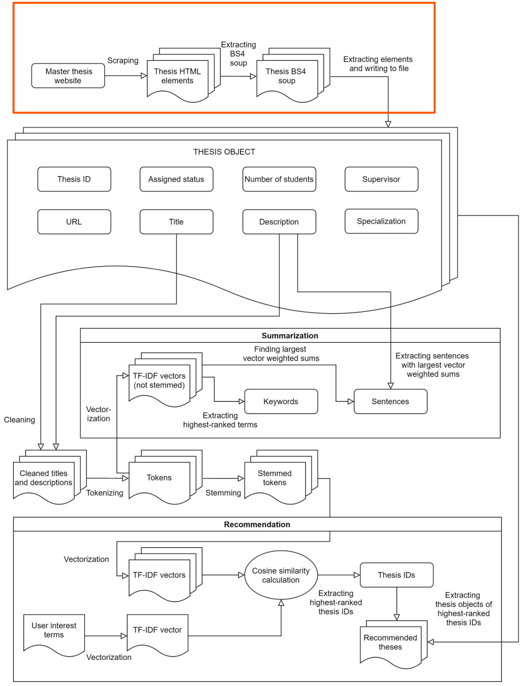
```
<div class="oppgave">

  <h3>Music emotion recognition and composition</h3>

  <p>The thesis project would explore either music emotion
recognition or automated music composition based on emotion-
annotated music - or a combination of those two themes. In
either case the work needs to explore (general) emotion
taxonomies, music data with and without emotional annotations,
and techniques for music emotion classification.</p>

    <div class="status">
      Faglærer: <a
href="https://www.ntnu.edu/employees/gamback">Björn
Gambäck</a>
      Status: <i>Valgbart</i>
      [...]
    </div>

</div>
```

HTML

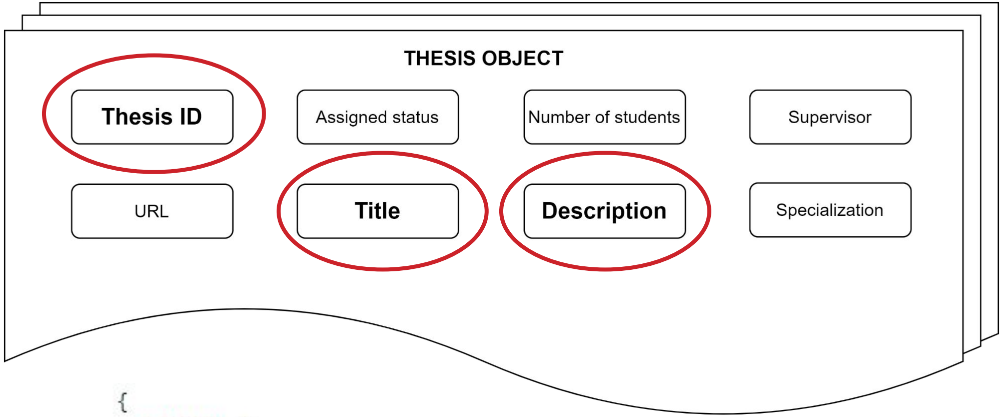


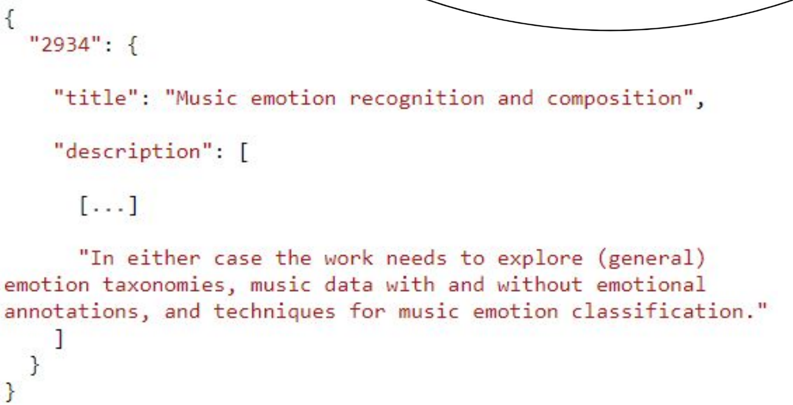
```
{
  "2934": {

    "title": "Music emotion recognition and composition",

    "description": [

      "The thesis project would explore either music emotion
recognition or automated music composition based on emotion-
annotated music - or a combination of those two themes.",

      "In either case the work needs to explore (general)
emotion taxonomies, music data with and without emotional
annotations, and techniques for music emotion classification."
    ]
  }
}
```

JSON

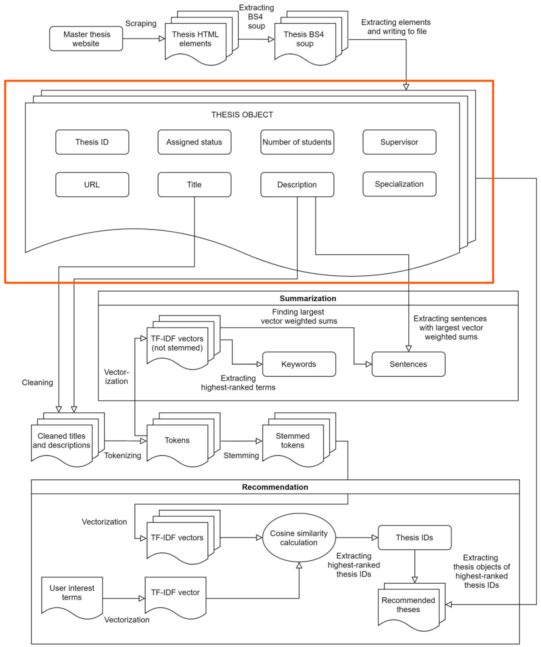


The Pipeline - Dataset

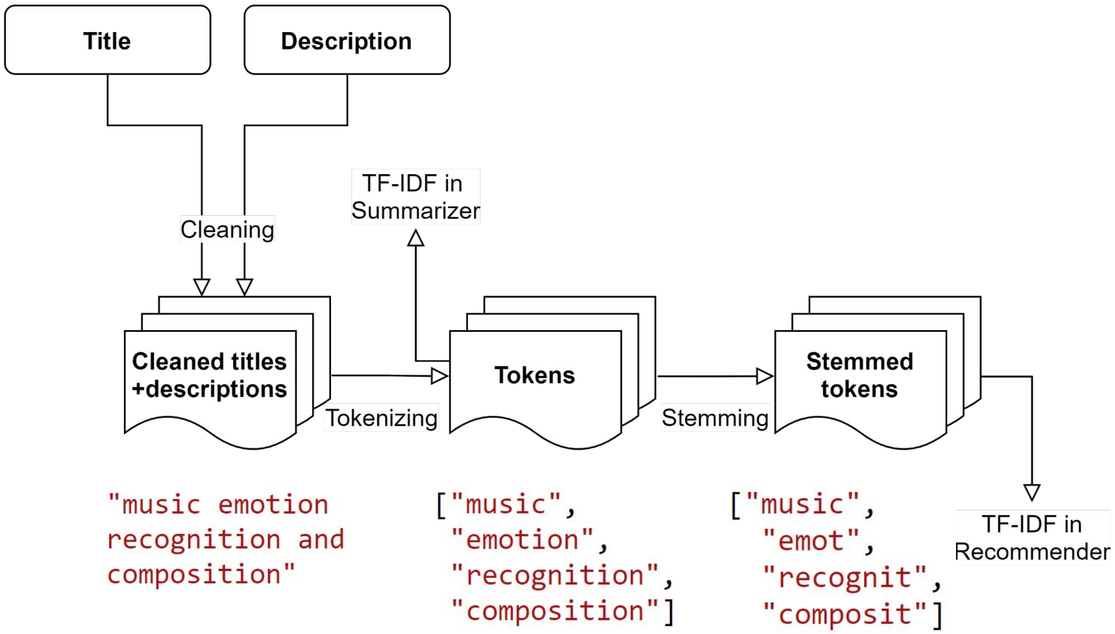




```
{
  "2934": {

    "title": "Music emotion recognition and composition",

    "description": [

      [...]

      "In either case the work needs to explore (general)
emotion taxonomies, music data with and without emotional
annotations, and techniques for music emotion classification."
    ]
  }
}
```



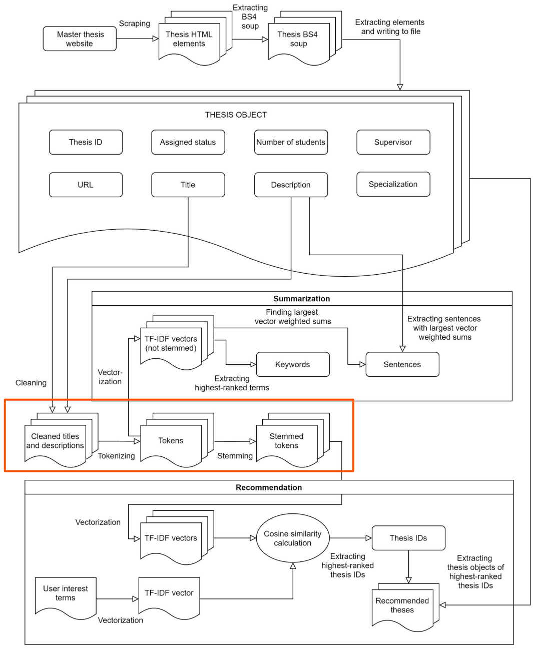
The Pipeline - Data Cleaning



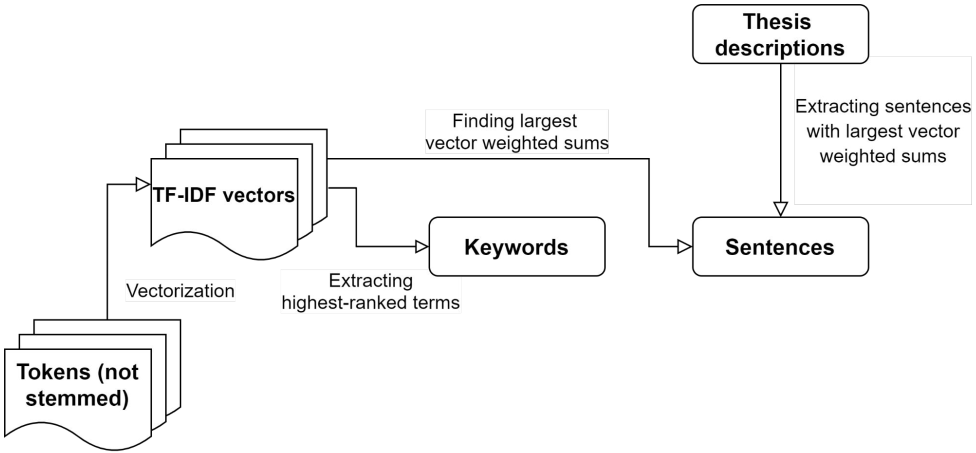
```
"music emotion
recognition and
composition"
```

```
["music",
 "emotion",
 "recognition",
 "composition"]
```

```
["music",
 "emot",
 "recognit",
 "composit"]
```



The Pipeline - Summarization



Thesis descriptions

Extracting sentences with largest vector weighted sums

Finding largest vector weighted sums

TF-IDF vectors

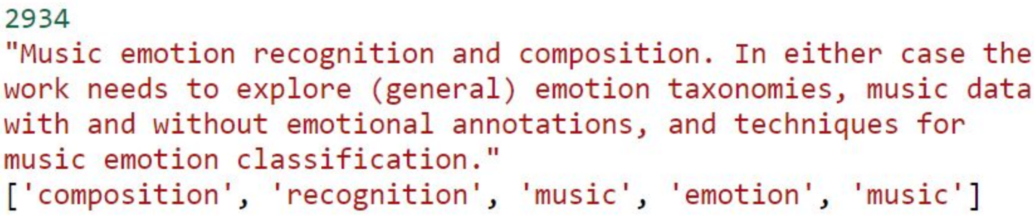
Keywords

Sentences

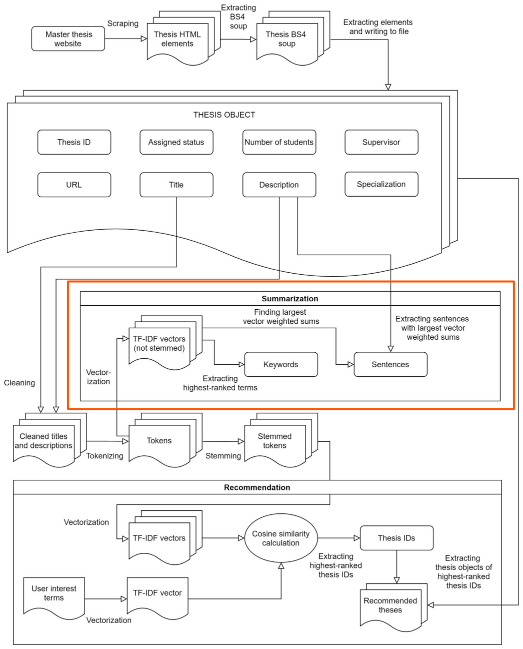
Vectorization

Extracting highest-ranked terms

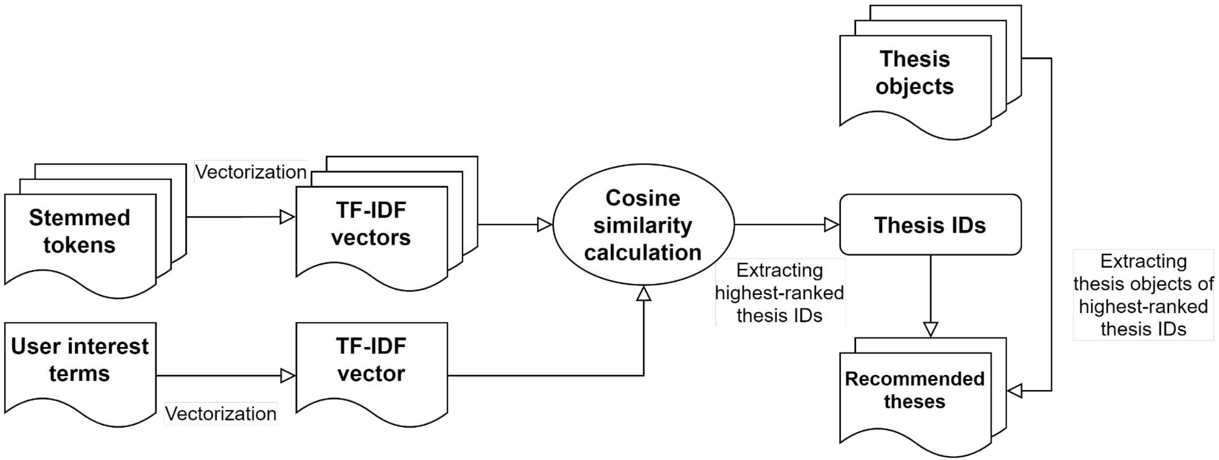
Tokens (not stemmed)

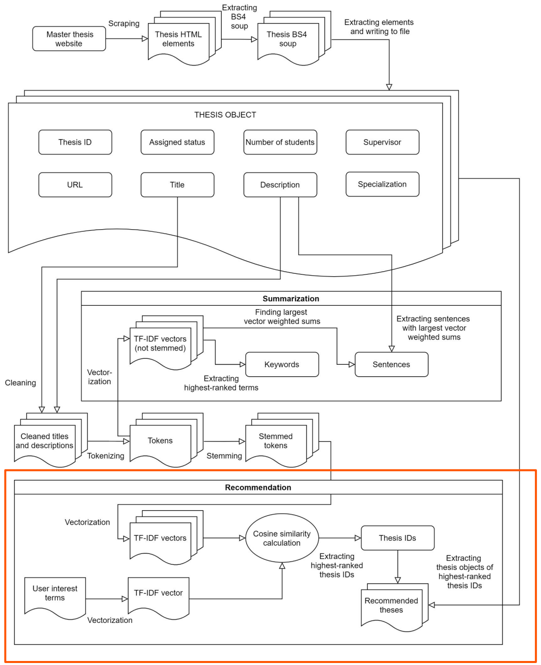


```
2934
"Music emotion recognition and composition. In either case the
work needs to explore (general) emotion taxonomies, music data
with and without emotional annotations, and techniques for
music emotion classification."
['composition', 'recognition', 'music', 'emotion', 'music']
```



The Pipeline - Recommendation



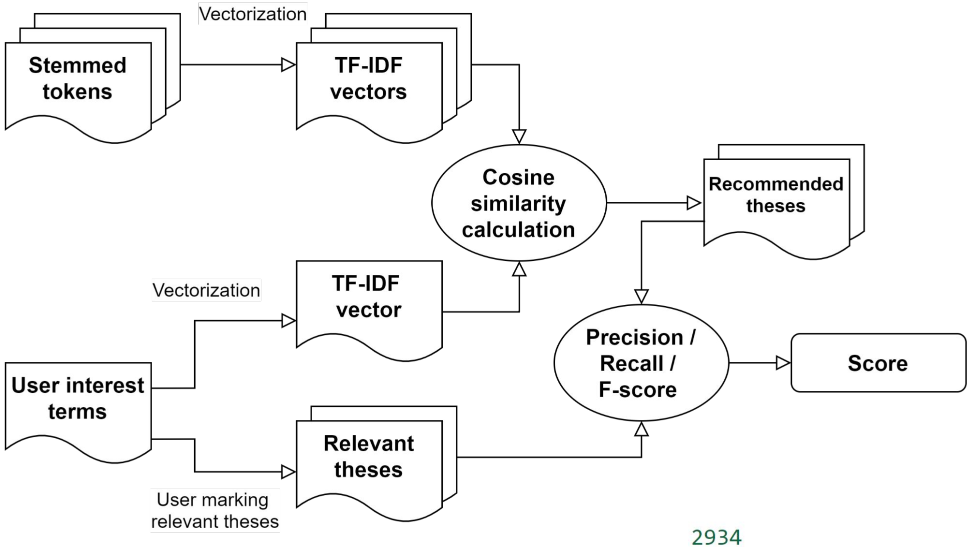


```
"music recognition automated"
```

```
2934
"Music emotion recognition and
composition"

2848
"Automatic music transcription with
deep learning"

2907
"Organizational decision-making in
the age of AI"
```

Experiments

Recommender Evaluation



"music recognition automated" →

4 relevant theses

R 2934 "Music emotion recognition and composition"

R 2848 "Automatic music transcription with deep learning"

2907 "Organizational decision-making in the age of AI"

→

Precision = 2/3
Recall = 2/4
F-score = 0.40

Survey

- We asked our friends that study computer science to provide us with:
  - Their interests
  - Relevant thesis topics (out of a subset of 40 thesis topics)

```
{
  "query": "frontend games agile web strategy",
  "relevant_ids": [3086, 3075, 2550, 3071, 2972, 3000, 2965, 2737, 2961, 2996, 2738, 2924],
  "language": "en"
}
```

- We got six such objects
- Not a lot of data, but it’s something!

Results and Discussion

Results – Recommender

We expect P to be high for LOW values of n
We expect R to be high for HIGH values of n

	Query 1	Query 3	Query 4	Query 5	Average
n = 5	P = 0.80 R = 0.29 F = 0.42	P = 1.00 R = 0.45 F = 0.63	P = 0.40 R = 0.20 F = 0.27	P = 0.40 R = 0.50 F = 0.44	P = 0.73 R = 0.39 F = 0.49
n = 10	P = 0.80 R = 0.57 F = 0.67	P = 0.90 R = 0.82 F = 0.86	P = 0.60 R = 0.60 F = 0.60	P = 0.20 R = 0.50 F = 0.29	P = 0.63 R = 0.63 F = 0.62
n = 15	P = 0.67 R = 0.71 F = 0.69	P = 0.67 R = 0.91 F = 0.77	P = 0.40 R = 0.60 F = 0.48	P = 0.13 R = 0.50 F = 0.21	P = 0.48 R = 0.71 F = 0.56

1 “AI computer vision deep learning transformers autonomous vehicles"

3 “Frontend games agile web security”

4 “Algorithms gpgpu graphics low-level non-hardware”

5 “music art nature history fashion”

Discussion

Performs quite well for simple queries
→ Commonly used words
→ About similar subjects

	Query 1	Query 3	Query 4	Query 5	Average
n = 5	**P = 0.80** R = 0.29 F = 0.42	**P = 1.00** R = 0.45 F = 0.63	P = 0.40 R = 0.20 F = 0.27	P = 0.40 R = 0.50 F = 0.44	P = 0.73 R = 0.39 F = 0.49
n = 10	**P = 0.80** R = 0.57 F = 0.67	P = 0.90 R = 0.82 **F = 0.86**	P = 0.60 R = 0.60 F = 0.60	P = 0.20 R = 0.50 F = 0.29	P = 0.63 R = 0.63 F = 0.62
n = 15	P = 0.67 **R = 0.71** F = 0.69	P = 0.67 **R = 0.91** F = 0.77	P = 0.40 R = 0.60 F = 0.48	P = 0.13 R = 0.50 F = 0.21	P = 0.48 R = 0.71 F = 0.56

1	“AI computer vision deep learning transformers autonomous vehicles"
3	“Frontend games agile web security”
4	“Algorithms gpgpu graphics low-level non-hardware”
5	“music art nature history fashion”

Discussion

Struggles with queries consisting of uncommon words
→ A thesaurus is a possible solution

	Query 1	Query 3	Query 4	Query 5	Average
n = 5	P = 0.80 R = 0.29 F = 0.42	P = 1.00 R = 0.45 F = 0.63	P = 0.40 R = 0.20 **F = 0.27**	P = 0.40 R = 0.50 F = 0.44	P = 0.73 R = 0.39 F = 0.49
n = 10	P = 0.80 R = 0.57 F = 0.67	P = 0.90 R = 0.82 F = 0.86	P = 0.60 R = 0.60 **F = 0.60**	P = 0.20 R = 0.50 F = 0.29	P = 0.63 R = 0.63 F = 0.62
n = 15	P = 0.67 R = 0.71 F = 0.69	P = 0.67 R = 0.91 F = 0.77	P = 0.40 R = 0.60 **F = 0.48**	P = 0.13 R = 0.50 F = 0.21	P = 0.48 R = 0.71 F = 0.56

1 “AI computer vision deep learning transformers autonomous vehicles"

3 “Frontend games agile web security”

4 “Algorithms gpgpu graphics low-level non-hardware”

5 “music art nature history fashion”

Discussion

Struggles with queries containing many different subjects
→ A hard case to deal with

	Query 1	Query 3	Query 4	Query 5	Average
n = 5	P = 0.80 R = 0.29 F = 0.42	P = 1.00 R = 0.45 F = 0.63	P = 0.40 R = 0.20 F = 0.27	P = 0.40 **R = 0.50** F = 0.44	P = 0.73 R = 0.39 F = 0.49
n = 10	P = 0.80 R = 0.57 F = 0.67	P = 0.90 R = 0.82 F = 0.86	P = 0.60 R = 0.60 F = 0.60	P = 0.20 **R = 0.50** **F = 0.29**	P = 0.63 R = 0.63 F = 0.62
n = 15	P = 0.67 R = 0.71 F = 0.69	P = 0.67 R = 0.91 F = 0.77	P = 0.40 R = 0.60 F = 0.48	P = 0.13 **R = 0.50** **F = 0.21**	P = 0.48 R = 0.71 F = 0.56

1 “AI computer vision deep learning transformers autonomous vehicles"

3 “Frontend games agile web security”

4 “Algorithms gpgpu graphics low-level non-hardware”

5 “music art nature history fashion”

Discussion

Average - we are quite happy with the results

	Query 1	Query 3	Query 4	Query 5	Average
n = 5	P = 0.80 R = 0.29 F = 0.42	P = 1.00 R = 0.45 F = 0.63	P = 0.40 R = 0.20 F = 0.27	P = 0.40 R = 0.50 F = 0.44	**P = 0.73** R = 0.39 F = 0.49
n = 10	P = 0.80 R = 0.57 F = 0.67	P = 0.90 R = 0.82 F = 0.86	P = 0.60 R = 0.60 F = 0.60	P = 0.20 R = 0.50 F = 0.29	P = 0.63 R = 0.63 **F = 0.62**
n = 15	P = 0.67 R = 0.71 F = 0.69	P = 0.67 R = 0.91 F = 0.77	P = 0.40 R = 0.60 F = 0.48	P = 0.13 R = 0.50 F = 0.21	P = 0.48 **R = 0.71** F = 0.56

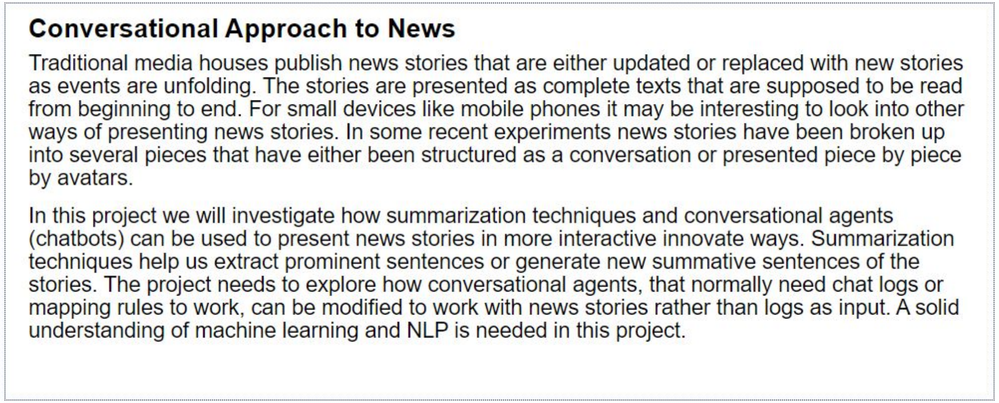
1 "AI computer vision deep learning transformers autonomous vehicles"

3 "Frontend games agile web security"

4 "Algorithms gpgpu graphics low-level non-hardware"

5 "music art nature history fashion"

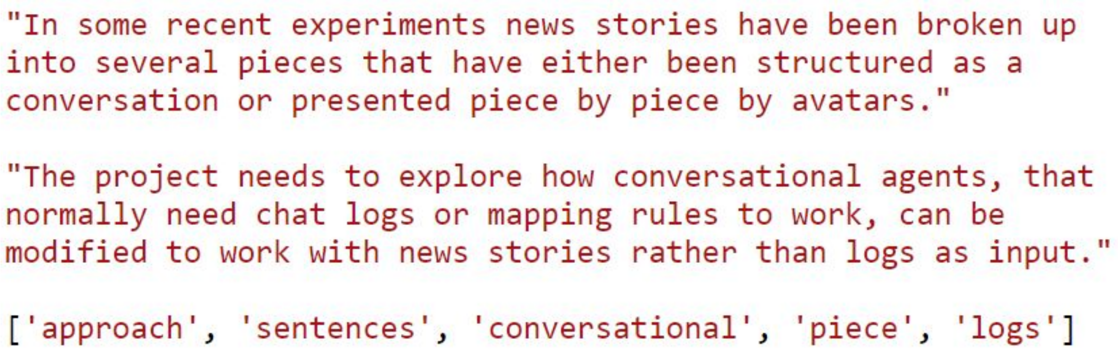
Discussion – Summarizer



Conversational Approach to News

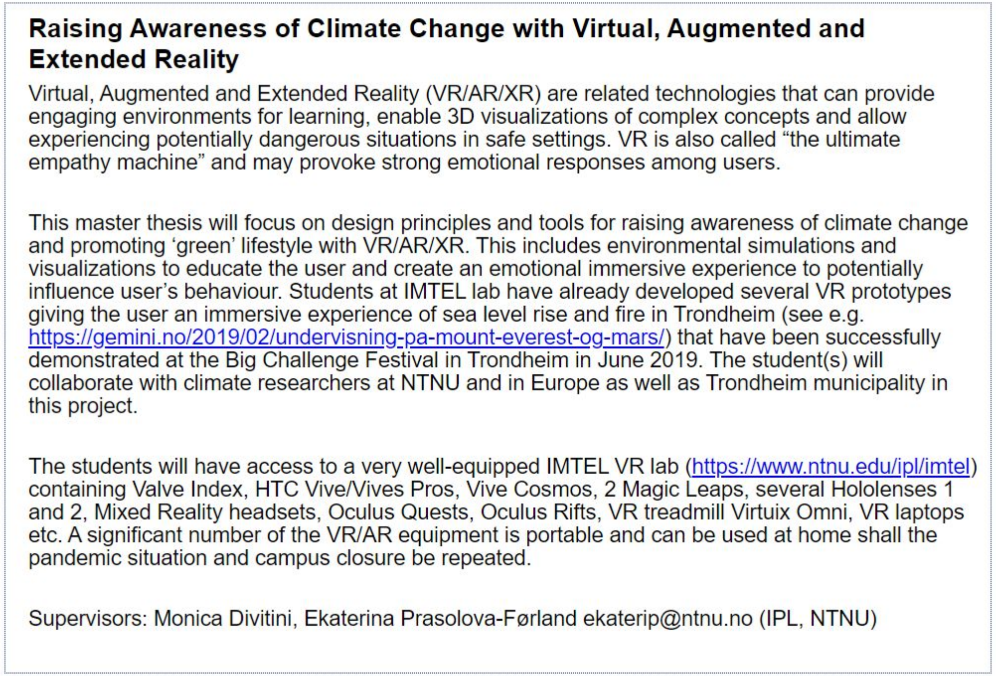
Traditional media houses publish news stories that are either updated or replaced with new stories as events are unfolding. The stories are presented as complete texts that are supposed to be read from beginning to end. For small devices like mobile phones it may be interesting to look into other ways of presenting news stories. In some recent experiments news stories have been broken up into several pieces that have either been structured as a conversation or presented piece by piece by avatars.

In this project we will investigate how summarization techniques and conversational agents (chatbots) can be used to present news stories in more interactive innovate ways. Summarization techniques help us extract prominent sentences or generate new summative sentences of the stories. The project needs to explore how conversational agents, that normally need chat logs or mapping rules to work, can be modified to work with news stories rather than logs as input. A solid understanding of machine learning and NLP is needed in this project.



```
"In some recent experiments news stories have been broken up
into several pieces that have either been structured as a
conversation or presented piece by piece by avatars."

"The project needs to explore how conversational agents, that
normally need chat logs or mapping rules to work, can be
modified to work with news stories rather than logs as input."

['approach', 'sentences', 'conversational', 'piece', 'logs']
```

Discussion – Summarizer



Raising Awareness of Climate Change with Virtual, Augmented and Extended Reality

Virtual, Augmented and Extended Reality (VR/AR/XR) are related technologies that can provide engaging environments for learning, enable 3D visualizations of complex concepts and allow experiencing potentially dangerous situations in safe settings. VR is also called “the ultimate empathy machine” and may provoke strong emotional responses among users.

This master thesis will focus on design principles and tools for raising awareness of climate change and promoting ‘green’ lifestyle with VR/AR/XR. This includes environmental simulations and visualizations to educate the user and create an emotional immersive experience to potentially influence user’s behaviour. Students at IMTEL lab have already developed several VR prototypes giving the user an immersive experience of sea level rise and fire in Trondheim (see e.g. https://gemini.no/2019/02/undervisning-pa-mount-everest-og-mars/) that have been successfully demonstrated at the Big Challenge Festival in Trondheim in June 2019. The student(s) will collaborate with climate researchers at NTNU and in Europe as well as Trondheim municipality in this project.

The students will have access to a very well-equipped IMTEL VR lab (https://www.ntnu.edu/ipl/imtel) containing Valve Index, HTC Vive/Vives Pros, Vive Cosmos, 2 Magic Leaps, several Hololenses 1 and 2, Mixed Reality headsets, Oculus Quests, Oculus Rifts, VR treadmill Virtuix Omni, VR laptops etc. A significant number of the VR/AR equipment is portable and can be used at home shall the pandemic situation and campus closure be repeated.

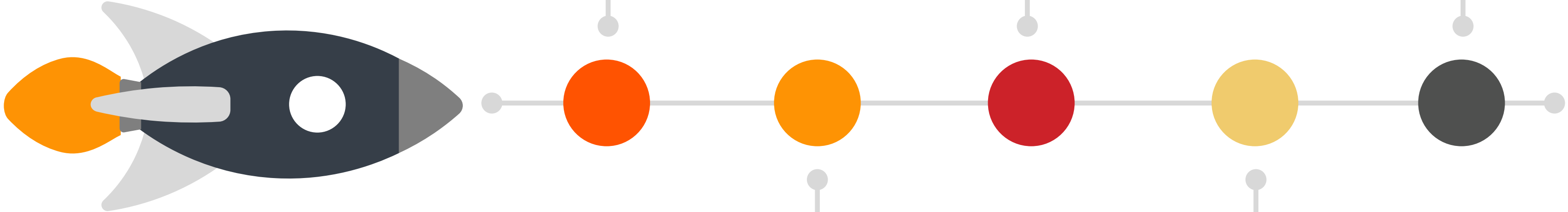
Supervisors: Monica Divitini, Ekaterina Prasolova-Førland ekaterip@ntnu.no (IPL, NTNU)

```
"https://gemini.no/2019/02/undervisni
ng-pa-mount-everest-og-mars/) that
have been successfully demonstrated
at the Big Challenge Festival in
Trondheim in June 2019."

"The students will have access to a
very well-equipped IMTEL VR lab
(https://www.ntnu.edu/ipl/imtel)
containing Valve Index, HTC
Vive/Vives Pros, Vive Cosmos, 2 Magic
Leaps, several Hololenses 1 and 2,
Mixed Reality headsets, Oculus
Quests, Oculus Rifts, VR treadmill
Virtuix Omni, VR laptops etc."

['ntnu', 'user', 'ipl', 'augmented',
'awareness']
```

Conclusion

- More time must be put into the cleaning process
- The summarizer has much potential for improvement
- The key words are practically useless
- We are generally very happy with the recommendation tool
- Test data is very limited

Future work



Thesaurus

Expand uncommon user interest terms

Preprocessing

Improve data cleaning and structuring

TF-IDF

Use proper DF in summarization

Evaluation

Improve schemes for evaluation

NER

Use entities in addition to tokens

Thank you!

Questions? 🤔