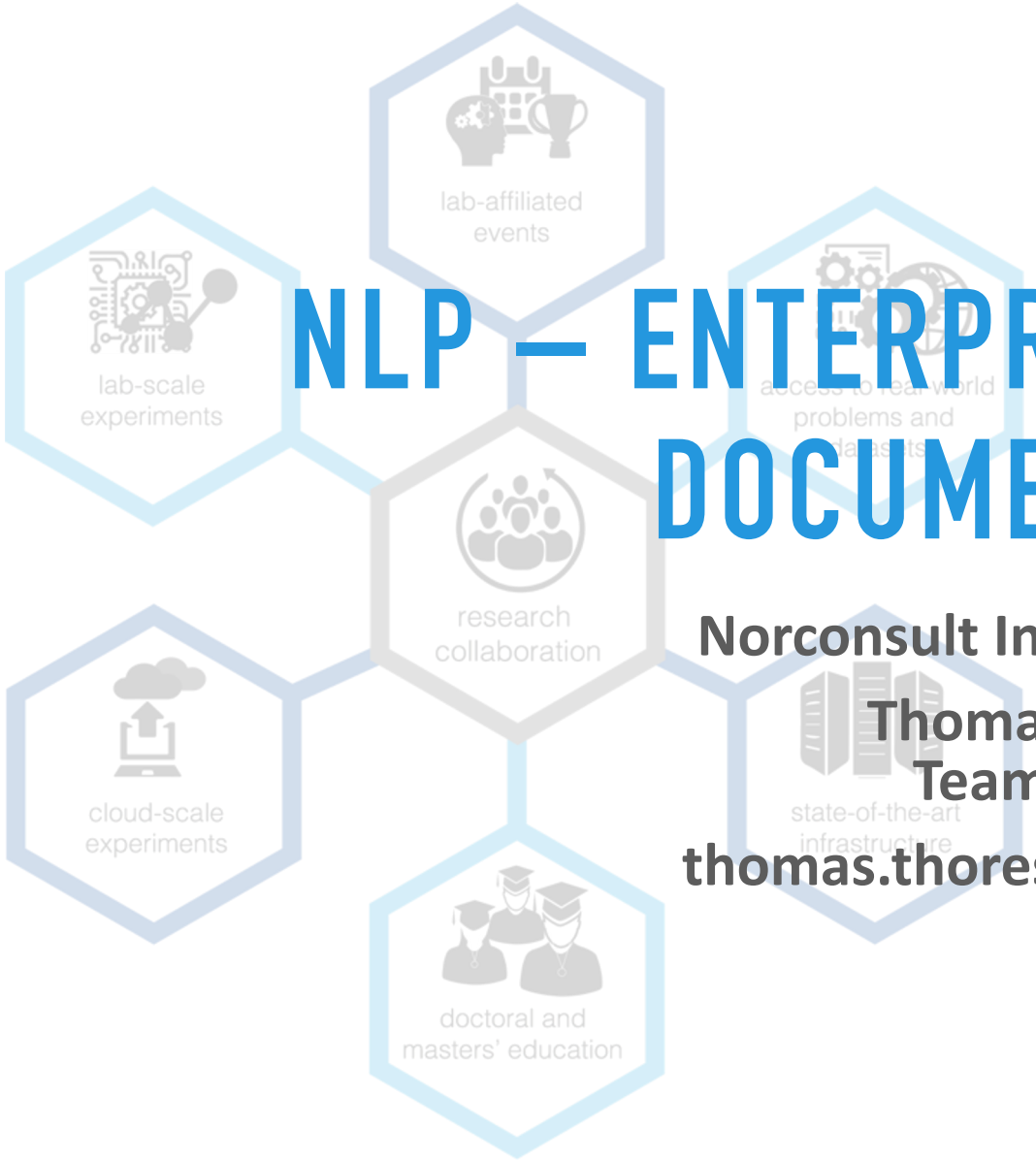


NLP – ENTERPRISE-SCALE SMART DOCUMENT SEARCH

Norconsult Informasjonssystemer

Thomas H. Thoresen
Team lead ML/AI
thomas.thoresen@norconsult.com

norwegian open ai lab

Who are you?

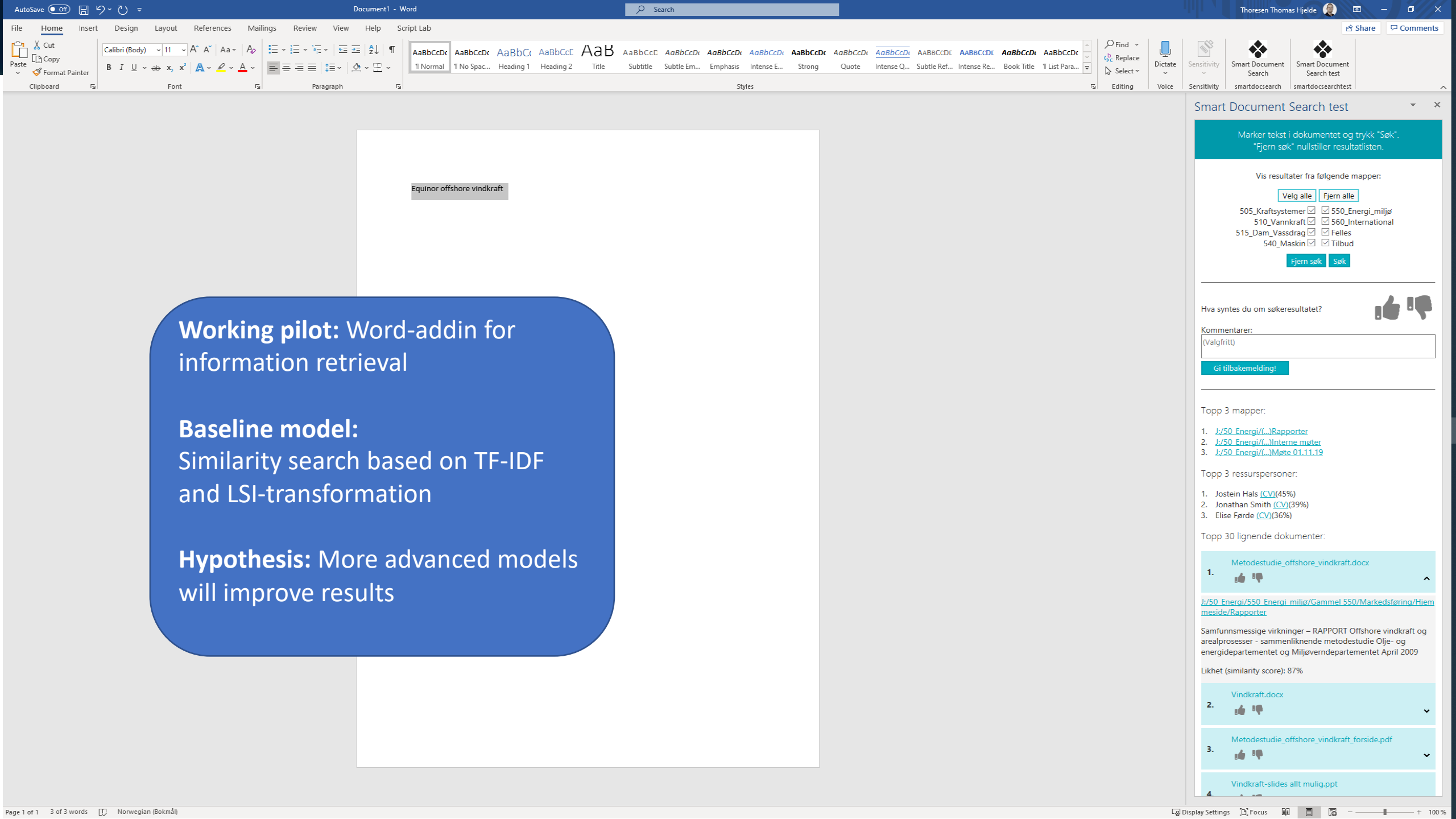
- Norconsult AS (Mother company)
- Norconsult Informasjonssystemer (NOIS) – IT Division
- Fundator – IT Consulting department
- ML/AI team – 14 consultants
- Thomas H. Thoresen – Team Lead ML/AI
- Consulting experience on ML/AI projects for
  - Equinor
  - DNV GL
  - Lånekassen
  - Digitaliseringsdirektoratet (Difi)
  - BNBank
  - Statens Vegvesen
  - Norconsult AS
  - ++
- MSc Computer Science from NTNU – AI programme

norwegian open ai lab

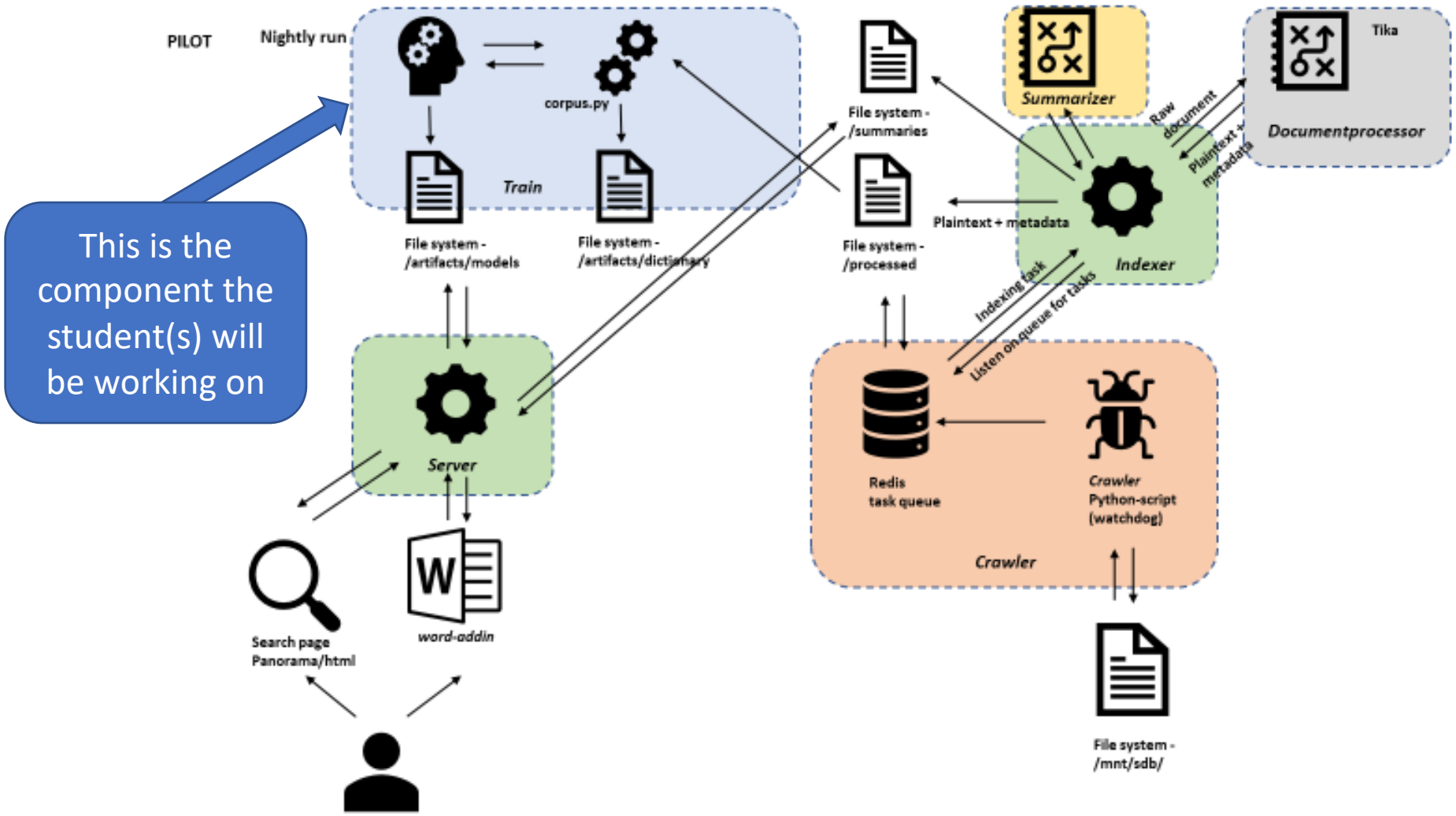
Problem Description

- NOAS has copious amounts of documents stored on network disks
- A lot of the information has accumulated through years and across geographic and subject domains, making it difficult to reuse and locate information as needed.
- The reuse of existing knowledge in documents has the potential to both improve the quality of our work in projects, as well as reduce the amount of time spent on unbillable tasks, such as preparing bids for tenders.

norwegian
open ai lab


Document1 - Word
Search
Thoresen Thomas Hjelde
Equinor offshore vindkraft
Working pilot: Word-addin for information retrieval
Baseline model:
Similarity search based on TF-IDF and LSI-transformation
Hypothesis: More advanced models will improve results
Smart Document Search test
Marker tekst i dokumentet og trykk "Søk".
"Fjern søk" nullstiller resultatlisten.
Vis resultater fra følgende mapper:
Velg alle
Fjern alle
505_Kraftsystemer
510_Vannkraft
515_Dam_Vassdrag
540_Maskin
550_Energi_miljø
560_International
Felles
Tilbud
Fjern søk
Søk
Hva syntes du om søkeresultatet?
Kommentarer:
(Valgfritt)
Gi tilbakemelding!
Topp 3 mapper:
1. J:/50_Energi/(...)Rapporter
2. J:/50_Energi/(...)Interne møter
3. J:/50_Energi/(...)Møte 01.11.19
Topp 3 ressurspersoner:
1. Jostein Hals (CV)(45%)
2. Jonathan Smith (CV)(39%)
3. Elise Førde (CV)(36%)
Topp 30 lignende dokumenter:
1. Metodestudie_offshore_vindkraft.docx
J:/50_Energi/550_Energi_miljø/Gammel 550/Markedsføring/Hjemmeside/Rapporter
Samfunnsmessige virkninger – RAPPORT Offshore vindkraft og arealprosesser - sammenliknende metodestudie Olje- og energidepartementet og Miljøverndepartementet April 2009
Likhet (similarity score): 87%
2. Vindkraft.docx
3. Metodestudie_offshore_vindkraft_forside.pdf
4. Vindkraft-slides allt mulig.ppt

Service-based architecture



norwegian open ai lab

Data

- All documents belonging to a specific part of Norconsult's Energy-division
- ~310 000 documents parsed to raw .txt-files
- Total size of raw documents: 612 GB
- After parsing: 69GB
- Some metadata as well, example below:

{"author": "Ola Nordmann", "last_author": "Kari Nordmann", "content_type": "application/pdf", "last_modified": " 2019-11-18T10:09:54Z ", "last_save_date": "", "creation_date": "2019-11-18T10:00:13Z", "x_tika_content": "<All document content here>", "path": "/data/tilbud/Beskrivelse av omfang.pdf", "filename": "Beskrivelse av omfang.pdf", "id": "6b2dab4c-2218-4ebf-b2bc-86c375e704d7"}

norwegian open ai lab

Data – labels

- Have established system for getting feedback from users easily and will provide a set of labeled queries / results based on this.
- Number of labeled queries / results not clear yet, but 500-100 is a fair estimate.

norwegian
open ai lab

Tasks / Challenges

- Work on improving the TF-IDF+LSI baseline for information retrieval.
- Review state-of-the art for semantic search with both small queries and paragraphs as queries.
- Some opportunities to explore include:
  - SIF
  - P-SIF
  - Finetuning word vectors to domain-specific corpus
  - Language models such as BERT etc.
- Implement and evaluate models/systems against labelled dataset and baseline.
- Package the output of the models into an API to allow for easy use, evaluation and extensibility.