**Computer assisted methods for topic modeling and their applications in bibliometrics**

## Rationale

Our floor does not have a constant publication record

## Original idea

With some help, I could do research on bibliometrics and publish.
So I contacted the Open University university in Holland to ask for help.

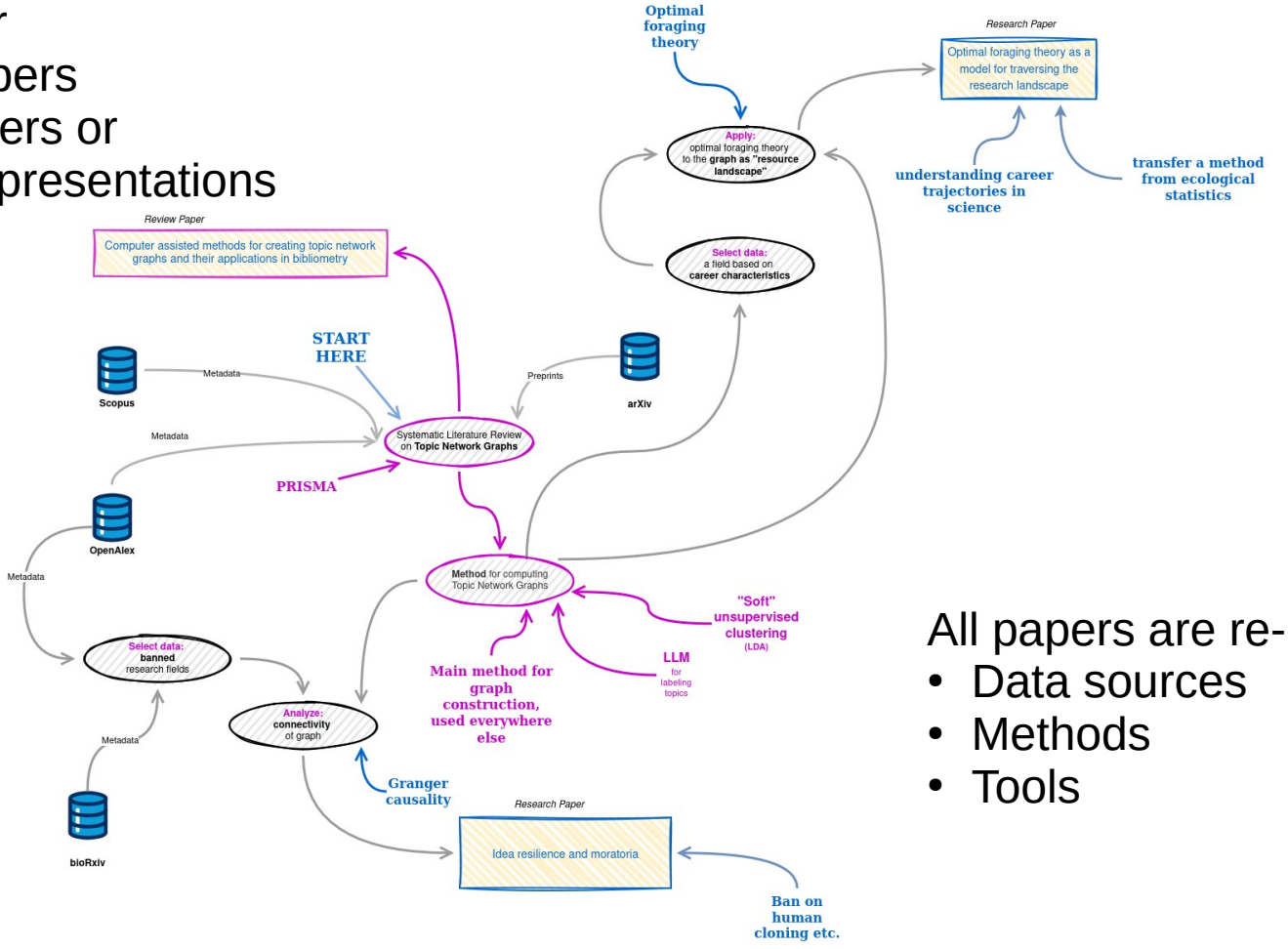
## OK, but…

The university agreed to help me, on condition that I submit a PhD proposal.
So our collaboration is formalized and my advisors get a benefit from their work.

*Still, the main plan is to publish, PhD comes second.*

**Plan:**

1x Review paper
2x Research papers
+2 (at least) papers or
conference presentations

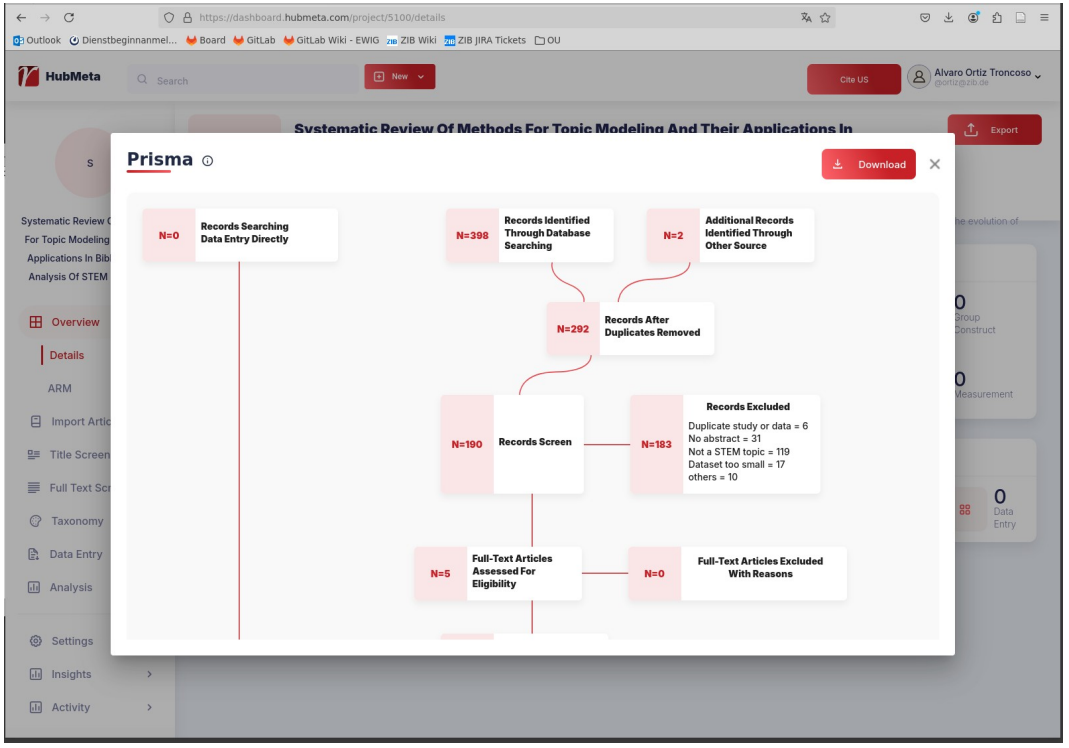



All papers are re-using common

- Data sources
- Methods
- Tools

# Literature review

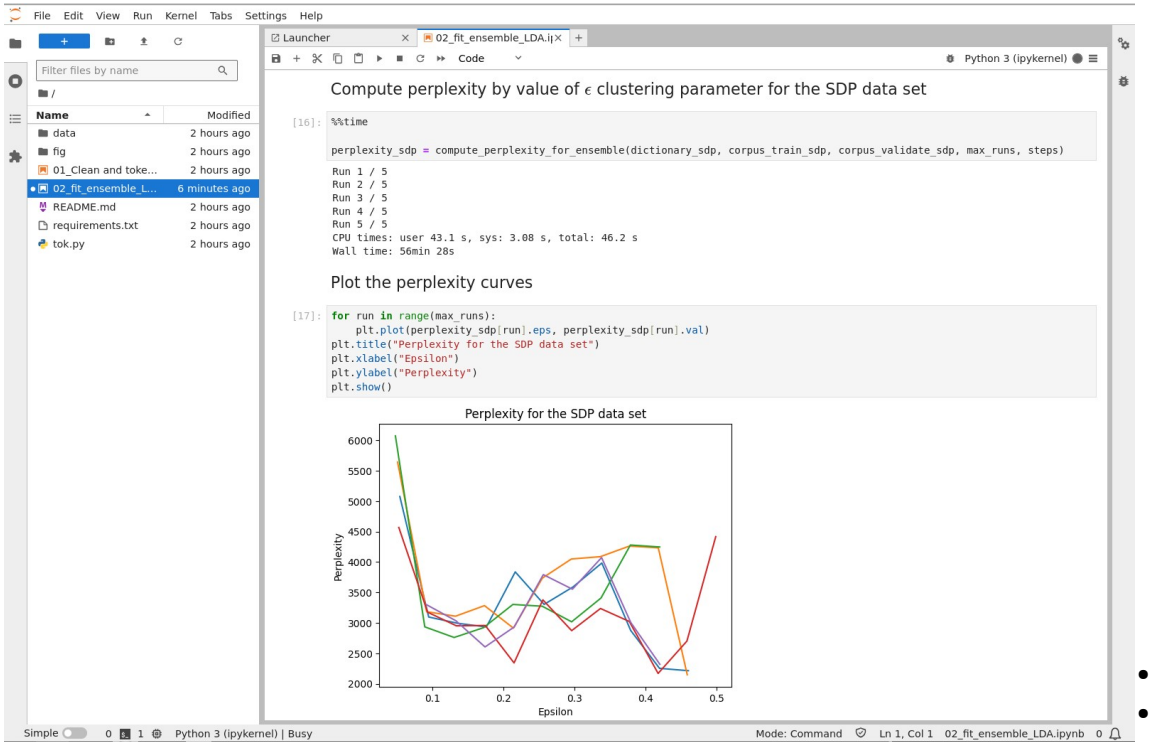
A systematic literature review on topic modelling is being conducted using tools that implement the PRISMA method.




http://hubmeta.com

# Topic modelling

Topic models are computed using LDA
**Alternative to explore: use BERT**




Where to run Jupyter notebooks on the KOBV infrastructure?

Currently on kobv-fan01
What about gtc-gpu201?

- https://github.com/aot29/arxiv_exploratory/
- https://github.com/aot29/openalex_exploratory/

# Naming the topics using LLM

```
You are an AI assistant specialised in Natural Language Processing.
What short, concise and human-readable label best describes the topic
characterised by these terms: {terms}? Output only the label"
```

```
Processing topic 14 / 15
CPU times: user 9min 28s, sys: 1.6 s, total: 9min 30s
Wall time: 37.7 s
```

```
topics_ensemble_cscl.sort_values(by='Topic')
```

| | Topic | First 5 keywords | Label |
|---|---|---|---|
| 0 | 0 | translation, machine, data, nmt, neural | Neural Machine Translation |
| 1 | 1 | question, answer, answering, task, reasoning | Natural Language Processing (NLP) |
| 2 | 2 | llm, task, large, performance, prompt | Natural Language Processing (NLP) |
| 3 | 3 | speech, data, task, recognition, training | Speech Recognition |
| 4 | 4 | bias, gender, data, task, based | Debiasing NLP Models |
| 5 | 5 | dialogue, task, state, system, human | Conversational AI |
| 6 | 6 | style, knowledge, task, transfer, text | Natural Language Processing (NLP) |
| 7 | 7 | evaluation, human, metric, task, summarization | Text Summarization |
| 8 | 8 | topic, approach, document, method, word | Natural Language Processing (NLP) |
| 9 | 9 | event, argument, extraction, task, method | Natural Language Processing (NLP) |
| 10 | 10 | code, task, generation, training, dataset | Large Language Models (LLMs) |
| 11 | 11 | grammar, parsing, based, syntactic, parser | Natural Language Processing (NLP) |
| 12 | 12 | graph, method, word, approach, network | Natural Language Processing (NLP) |
| 13 | 13 | news, social, medium, information, data | Fake News Detection |
| 14 | 14 | speaker, speech, based, method, task | Speech Synthesis |

## LlaMa: fast but inaccurate

```
Processing topic 14 / 15
CPU times: user 7min 46s, sys: 8.16 s, total: 7min 55s
Wall time: 7min 55s
```

```
topics_ensemble_cscl.sort_values(by='Topic')
```

| | Topic | First 5 keywords | Label |
|---|---|---|---|
| 0 | 0 | translation, machine, data, nmt, neural | Neural Machine Translation (Nmt) - Training Performance Explanation |
| 1 | 1 | question, answer, answering, task, reasoning | Question Answering And Information Retrieval (Qair) |
| 2 | 2 | llm, task, large, performance, prompt | Large-Scale Llm (Language Model) Performance |
| 3 | 3 | speech, data, task, recognition, training | Speech Recognition System |
| 4 | 4 | bias, gender, data, task, based | Gender Bias In Large-Scale Nlp Systems |
| 5 | 5 | dialogue, task, state, system, human | Dialogue System |
| 6 | 6 | style, knowledge, task, transfer, text | Text-Based Transfer Learning For Sentence-Level Tasks |
| 7 | 7 | evaluation, human, metric, task, summarization | Automatic Text Summarization Evaluation |
| 8 | 8 | topic, approach, document, method, word | Topic Modeling With Neural Lda And Text Data Analysis |
| 9 | 9 | event, argument, extraction, task, method | Event-Based Argument Mining |
| 10 | 10 | code, task, generation, training, dataset | Machine Learning Model Training And Evaluation |
| 11 | 11 | grammar, parsing, based, syntactic, parser | Natural Language Processing (Nlp) Techniques |
| 12 | 12 | graph, method, word, approach, network | Text-Based Graph Neural Networks For Information Retrieval And Question Answering |
| 13 | 13 | news, social, medium, information, data | Social Media News Detection |
| 14 | 14 | speaker, speech, based, method, task | Speech-Based Methods For Multi-Utterance Text Proposal In Human-Voice Prosody Labeling |

## Mistral: more precise, but also much slower

Where to go from here?

Collaborate with my advisors / co-authors from the Open University in Holland

***Research questions:***

- What is the current state of topic modeling, in terms of techniques and applications?
- What temporal patterns describe the trajectories of scientists' careers?
- Which interactions exist between idea resilience and moratoria in science?
- Find a way to compare the Topic Graph and the Citation Graph
- *Further research questions are currently being examined for feasibility*