Software to be applied

- NLTK
- Retriev
- PyTerrier
- BERTopic
- ElasticSearch
- …

1st batch of tasks

- 3 notebooks are online
- Get familiar with their functions and check how they work
- They will be used in the next tasks

2nd batch of tasks

- Find your dataset
- Create a TopicModel with your dataset
- Index and search your Dataset

Obtain a Dataset

- Obtain a Dataset from either
  - datasets Or pyTerrier
  - 2 Notebooks are provided
  - Make sure you chose a dataset with at least 10000 documents and at least one field which contains text
  - Add your name and the name of the dataset into the in the Learnweb. Each student should use a different dataset

Look at the content of the data

- BERTopic
- Create a topic model from your text data
- If there are performance issues, you can limit the system to the first 2000 documents
- Create another topic model with a lower number of topics

Index your collection with retriev

- Index one text column as document from your Dataset with two different weighting formulas and two different stemmers
- Run the same query and print the results for the four combinations