

# NATURAL LANGUAGE PROCESSING

## TWO POSSIBLE WORKFLOWS

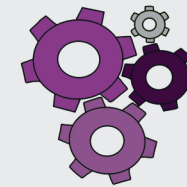
### BAG OF WORDS



Load **data**



	world	quiz
Doc A	4	0
Doc B	0	5
⋮		



Tokenize text

Pre-process tokens (stemming, lemmatization, remove stops words, ...)

Generate a **document feature matrix** (DFM)

- Use labeled data to **train your model** and to classify texts
- Use **dictionary-based approaches** for topic models or sentiment analysis
- ...

### BERT

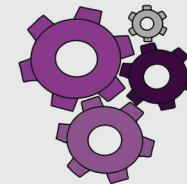


Load **data**

Load **pre-trained tokenizer**

Load **pre-trained model**

[CLS] rory lives in a world of books [SEP]



Tokenize text with **pre-trained tokenizer** (pre-processing happens under the hood)

Use labeled data to **fine-tune pre-trained models** for text classification, sentiment analysis, translation, question answering, ...




 Tutorial: **Tokenizer pipeline**  
<https://huggingface.co/docs/tokenizers/pipeline>



 Tutorial: **Fine-tune pre-trained model**  
<https://huggingface.co/docs/transformers/training>



 Different tasks at Huggingface  
<https://huggingface.co/tasks>