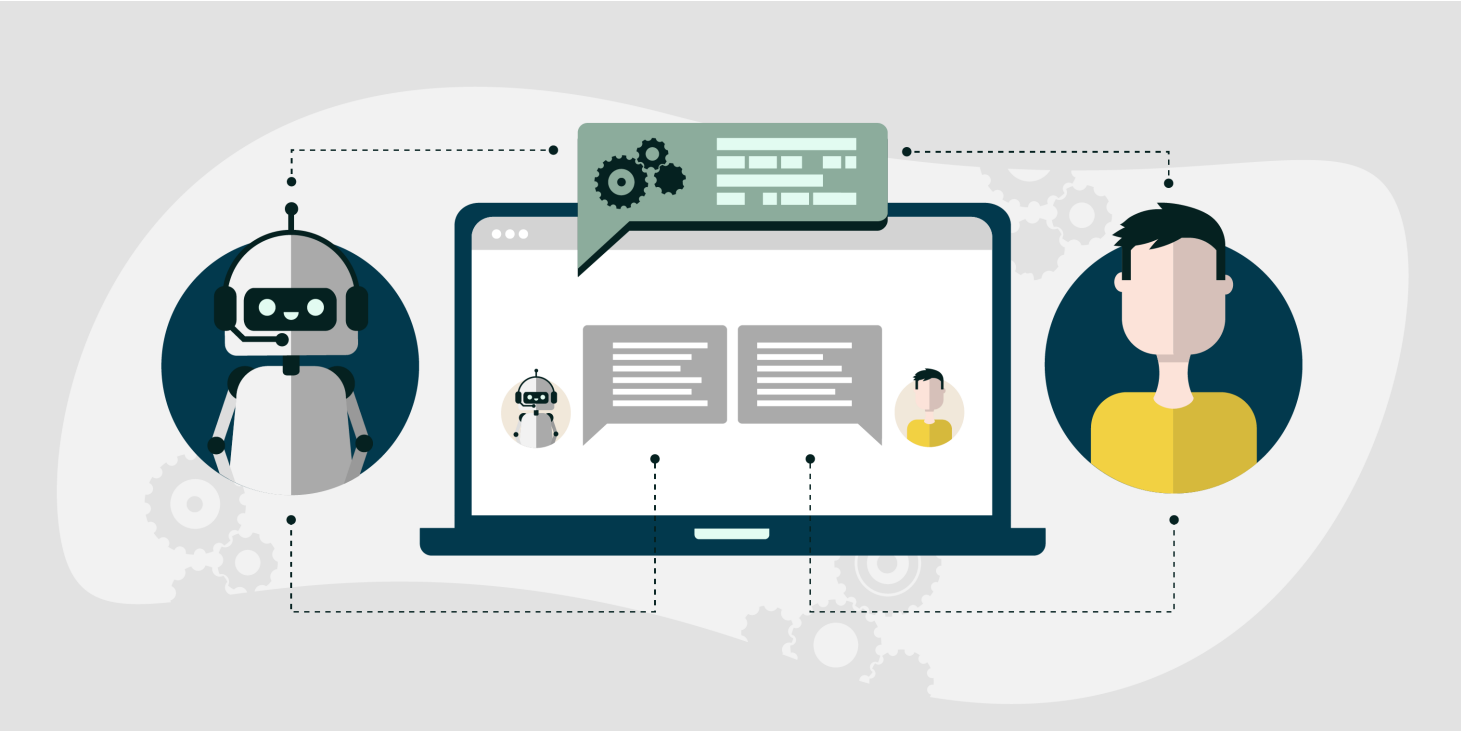
Introduction to Natural Language Processing

What is Natural Language Processing?



What is Natural Language Processing?

- Extracting meaningful information from natural language text
- Examples:
  - Sentiment Analysis
  - Chat Bots
  - Automatic Translation
  - Speech Recognition

Steps in Natural Language Processing

1. Data Cleaning (remove punctuation, capitalization etc)
2. Tokenization (separating each word into its own entity)
3. Removing 'stop words' (such as and, or, like, ...)
4. Lemmatization / Stemming (removing common prefixes or suffixes, replacing words by their roots - such as mapping 'gone', 'going', 'goes', 'went' all into 'go')

Bag of Words

- 'Bag Of Words': text is represented as a "bag" of words without paying attention to grammar or word order, but keeping number of repetitions.
- Vectorization will generate vectors which indicate the presence of tokens in different text instances.
- From here on, we can apply models we already know!

Natural Language Processing

Note: this workshop is using a very basic algorithm in text classification. In many cases, the positions of words in a sentence have additional meaning, and there are more advanced algorithms to handle this. Some example: Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer 3 (GPT-3)