Natural Language Processing

Natural Language Processing

Natural language processing (NLP) refers to the branch of data science concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the speech-recognition, part of speech tagging (POS), word sense disambiguation, sentiment analysis, text classification, natural language generation

Text is unstructured data so, in order to be processed by a computer, it need to be properly pre-processed to extract numerical features out of it

Some useful NLP terminology

- **Document**: some text.
- **Corpus**: a collection of documents.
- **Dictionary**: a list of all the words in a document
- **Token**: a well-defined unit inside a document. It can be a word or a sub-word.
- **Lemma**: the root or base form of a word (e.g. played, playing => play)
- **Vector**: a mathematically convenient representation of a document.
- **Model**: an algorithm for transforming vectors from one representation to another.

Text pre-processing: Tokenisation and Vectorisation

Text pre-processing often consists of these steps:

- **tokenizing** strings and giving an integer id for each possible token, for instance by using white-spaces and punctuation as token separators.
  - optionally tokenization can be followed by **stop-word removal** and **stemming** or **lemmatization**
- some form of **vectorisation**, such as**:**
  - **counting** the occurrences of tokens in each document (bag-of-words approach)
  - **normalizing** and weighting with diminishing importance tokens that occur in a plurality of samples. (TF-IDF approach)
  - **word embeddings** (Word2Vec, GloVe, FastText, etc.)

Example of NLP pipeline