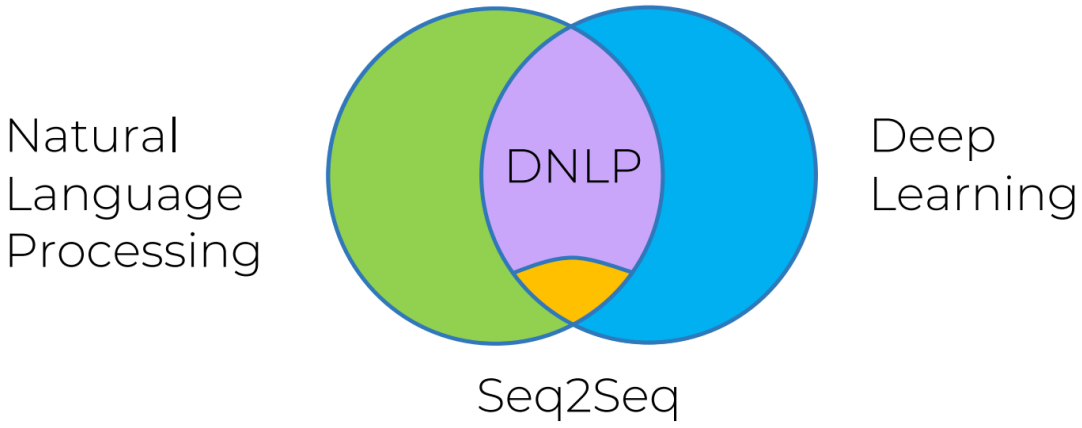
Natural Language Processing

Natural language processing (**NLP**) is an interdisciplinary subfield of computer science - specifically Artificial Intelligence - and linguistics. It is primarily concerned with providing computers the ability to process data encoded in natural language, typically collected in text corpora, using either rule-based, statistical or neural-based approaches of machine learning and deep learning.

Major tasks in Natural Language Processing are speech recognition, text classification, natural-language understanding, and natural-language generation.

Types of NLP:



Natural language processing definition

Natural language processing (NLP) is a subset of artificial intelligence, computer science, and linguistics focused on making human communication, such as speech and text, comprehensible to computers.

NLP is used in a wide variety of everyday products and services. Some of the most common ways NLP is used are through voice-activated digital assistants on smartphones, email-scanning programs used to identify spam, and translation apps that decipher foreign languages.

Natural language techniques

1. Text Processing and Preprocessing In NLP

- **Tokenization**: Dividing text into smaller units, such as words or sentences.
- **Stemming and Lemmatization**: Reducing words to their base or root forms.
- **Stopword Removal**: Removing common words (like “and”, “the”, “is”) that may not carry significant meaning.
- **Text Normalization**: Standardizing text, including case normalization, removing punctuation, and correcting spelling errors.

NLP encompasses a wide range of techniques to analyze human language. Some of the most common techniques you will likely encounter in the field include:

- **Sentiment analysis:** An NLP technique that analyzes text to identify its sentiments, such as “positive,” “negative,” or “neutral.” Sentiment analysis is commonly used by businesses to better understand customer feedback.
- **Summarization:** An NLP technique that summarizes a longer text, in order to make it more manageable for time-sensitive readers. Some common texts that are summarized include reports and articles.
- **Keyword extraction:** An NLP technique that analyzes a text to identify the most important keywords or phrases. Keyword extraction is commonly used for search engine optimization (SEO), social media monitoring, and business intelligence purposes.
- **Tokenization:** The process of breaking characters, words, or subwords down into “tokens” that can be analyzed by a program. Tokenization undergirds common NLP tasks like word modeling, vocabulary building, and frequent word occurrence.

Deep learning

Deep learning is the subset of machine learning methods based on neural networks with representation learning. The adjective "deep" refers to the use of multiple layers in the network. Methods used can be either supervised, semi-supervised or unsupervised.

Deep-learning architectures such as deep neural networks, deep belief networks, recurrent neural networks, convolutional neural networks and transformers have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

Early forms of neural networks were inspired by information processing and distributed communication nodes in biological systems, in particular the human brain. However, current neural networks do not intend to model the brain function of organisms, and are generally seen as low-quality models for that purpose.

Types of deep learning models

Deep learning models use a variety of constructions and frameworks to achieve specific tasks and goals. Some types of deep learning models include:

- **Convolutional neural networks:** You can use convolutional neural networks for image processing and recognition.
- **Recurrent neural networks:** You can use recurrent neural networks for speech recognition and natural language processing.
- **Long short-term memory networks**: You can use long short-term memory networks for sequential prediction tasks, such as language modeling.

Bag-of-words model

Bag-of-Words Model

The **bag-of-words model** is a model of text which uses a representation of text that is based on an unordered collection (or "bag") of words. It is used in natural language processing and information retrieval (IR). It disregards word order (and thus any non-trivial notion of grammar) but captures multiplicity. The bag-of-words model has also been used for computer vision.

Example implementation

The following models a text document using bag-of-words. Here are two simple text documents:

```
(1) John likes to watch movies. Mary likes movies too.
(2) Mary also likes to watch football games.
```

Based on these two text documents, a list is constructed as follows for each document:

```
"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"
```

Representing each bag-of-words as a JSON object, and attributing to the respective JavaScript variable:

```
BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
BoW2 = {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};
```

Each key is the word, and each value is the number of occurrences of that word in the given text document.

The order of elements is free, so, for example {"too":1,"Mary":1,"movies":2,"John":1,"watch":1,"likes":2,"to":1} is also equivalent to *BoW1*. It is also what we expect from a strict *JSON object* representation.

Note: if another document is like a union of these two,

```
(3) John likes to watch movies. Mary likes movies too. Mary also likes to watch football games.
```

its JavaScript representation will be:

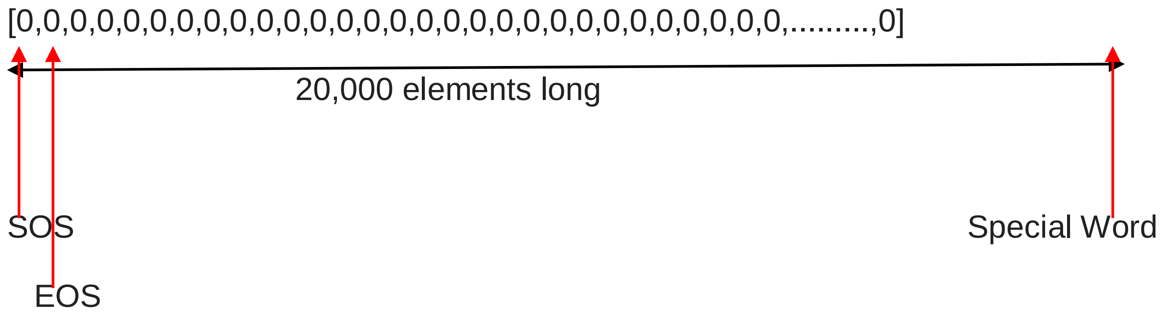
```
BoW3 =
{"John":1,"likes":3,"to":2,"watch":2,"movies":2,"Mary":2,"too":1,"also":1,"football":1,"games":1};
```

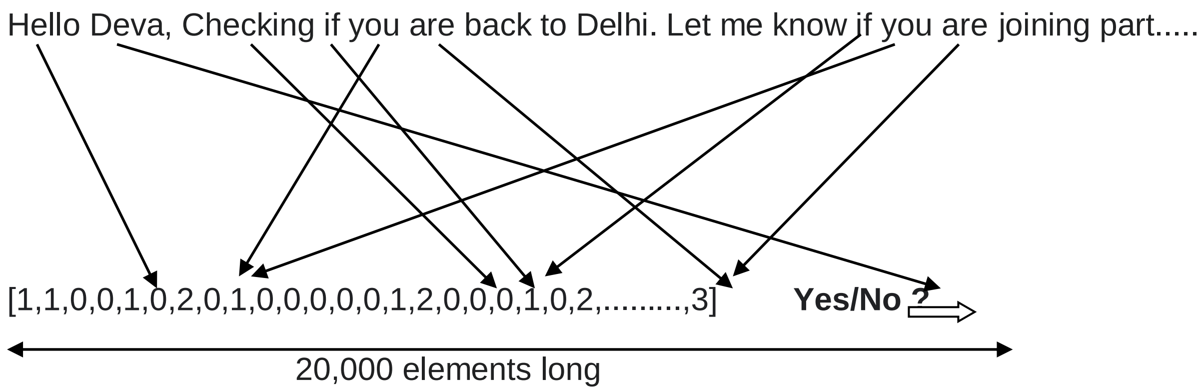
So, as we see in the bag algebra, the "union" of two documents in the bags-of-words representation is, formally, the disjoint union, summing the multiplicities of each element.

> $BoW3 = BoW1 \biguplus BoW2$.

English Words:

Estimating the number of words in the English language is a complex process. The Oxford English Dictionary estimates that there are around 170,000 words in current use, with an additional 47,000 obsolete and 20,000-30,000 words used by each individual person.





Training Data:

Hey mate, have you read about Hinton's capsule network? → No

Did you like that recipe I sent you last week? → Yes

Hi Deva, are you coming to dinner tonight? → Yes

Dear Deva, would you like to service your car with us again? → No

Are you coming to Australia in December? → Yes

→ ---