MO434 - Deep Learning
Introduction to Text Analysis

Alexandre Xavier Falcão

Institute of Computing - UNICAMP

afalcao@ic.unicamp.br

Text Analysis and Natural Language Processing

- Text analysis combines Machine Learning (ML) and Natural Language Processing (NLP) to understand language present in texts, enabling human-machine interaction.

Text Analysis and Natural Language Processing

- Text analysis combines Machine Learning (ML) and Natural Language Processing (NLP) to understand language present in texts, enabling human-machine interaction.

- Applications involve machine translation, question and answering, text summarization, sentiment analysis, etc.

Text Analysis and Natural Language Processing

- Text analysis combines Machine Learning (ML) and Natural Language Processing (NLP) to understand language present in texts, enabling human-machine interaction.

- Applications involve machine translation, question and answering, text summarization, sentiment analysis, etc.

- We will divide text analysis into four lectures: text processing, representation, recurrent neural networks and transformers. Applications will be presented along with the lectures.

Agenda

- Why do we need text (pre)processing?
- What are the main NLP techniques and the importance of a corpus in text processing?
- A simple text classification application.

Why do we need text (pre)processing?

- We need to clean irrelevant parts and standardize the others to facilitate text analysis.

Why do we need text (pre)processing?

- We need to clean irrelevant parts and standardize the others to facilitate text analysis.
- A text can be divided into sentences and each sentence into tokens (entities).

Why do we need text (pre)processing?

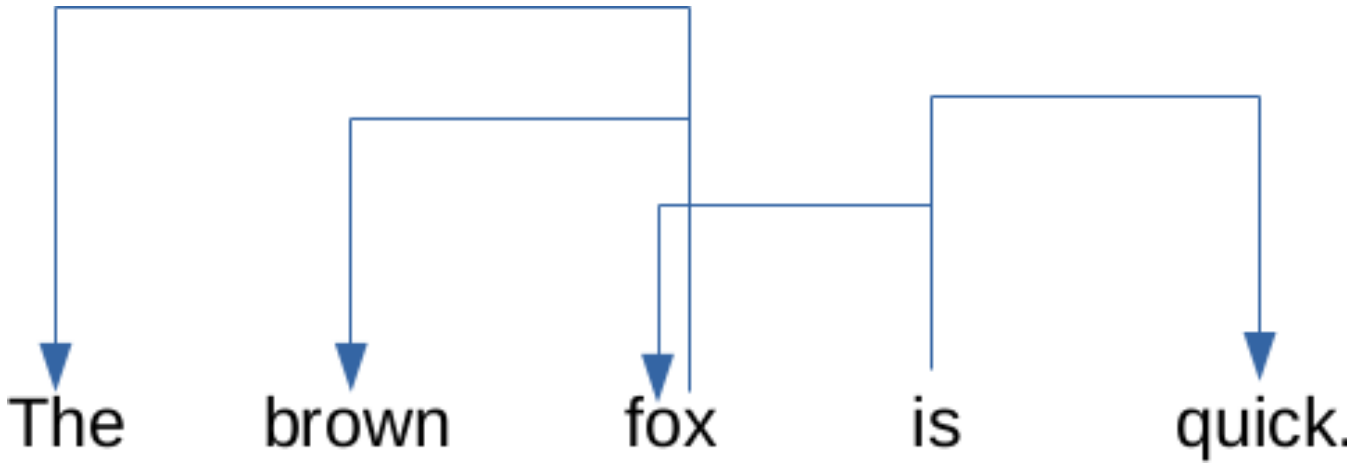
- We need to clean irrelevant parts and standardize the others to facilitate text analysis.
- A text can be divided into sentences and each sentence into tokens (entities).
- Each token has a role (part of speech - POS) and may be categorized (Named Entity Recognition - NER) by a pretrained model.

Why do we need text (pre)processing?

- We need to clean irrelevant parts and standardize the others to facilitate text analysis.
- A text can be divided into sentences and each sentence into tokens (entities).
- Each token has a role (part of speech - POS) and may be categorized (Named Entity Recognition - NER) by a pretrained model.
- Sentences can also be structured and viewed by different ways for better understanding and reasoning.

Different ways to view sentences

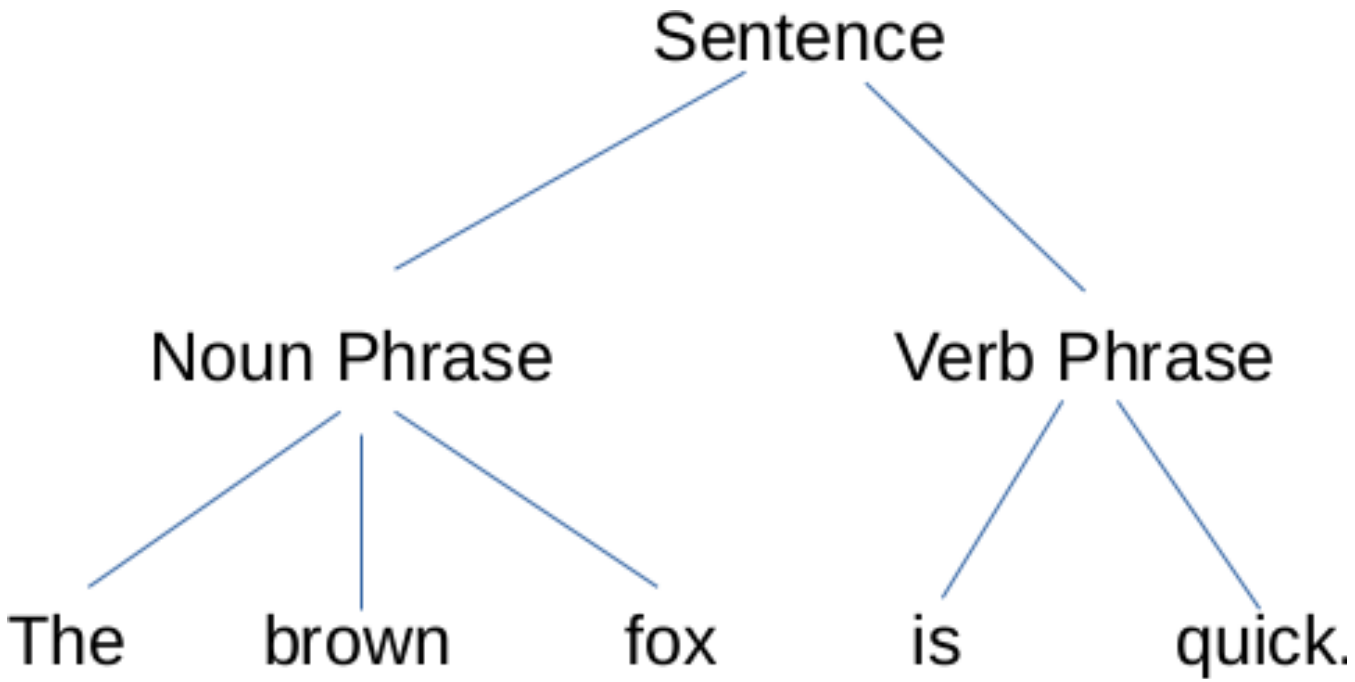
The dependency of a token with another token.



One may use that information to create knowledge graphs.

Different ways to view sentences

The constituent parts of a sentence.



Groups that contain information about the subject of the sentence.

NLP techniques and a simple application

Let's see the following notebook with NLP techniques, the importance of a corpus, and a simple application. ▸ (TEXT PROCESSING).