Terminology: text preprocessing with spaCy

Practical 1, Master 2 NLP

2021/2022

1 Preliminary

spaCy (`https://spacy.io`) is a free, open-source library for advanced Natural Language Processing (NLP) in Python. We will use it in the context of this course to preprocess texts and extract from them candidate terms.

1. Install spaCy on your machine
2. Read the tutorials on "Linguistic features" and "Rule-based matching" (`https://spacy.io/usage/spacy-101`)

2 Text preprocessing

1. Get a scientific article in English in the specialized domain you are the most familiar with (ex. linguistics, computer science, NLP). Copy-and-paste the text in a text file.
2. Preprocess the corresponding text using spaCy part-of-speech tagger.
3. Check the result.

3 Extracting candidate terms

1. Using the "token matcher" tool of spaCy, extract candidate terms from the text with a pattern composed of a sequence of two nouns.
2. Check the result.
3. Add more patterns in order to extract more candidates.