PHILOSOPHY AND HISTORY OF SCIENCE WITH COMPUTATIONAL MEANS

PROF. DR. GERD GRAßHOFF

Natural Language Processing

In this tutorial, we will learn about the basics of Natural Language Processing (NLP). The most popular suite of libraries for NLP is the Natural Language Toolkit (NLTK), released in 2001. We will start with spaCy, an open-source software library for advanced NLP released in 2015 and a very efficient tool.

Setup

Open the Anaconda Navigator, go to 'Environments,' click on 'base (root),' and choose 'Open Terminal.' Write:

```
conda install -c conda-forge spacy
```

Press the 'y' standing for 'yes' to proceed.

Now that you have installed spaCy, we need to download the models for the languages that we need. For example, for English:

```
python -m spacy download en
```

Press the up-arrow key, and the last input will appear again. In this case, you can erase the 'en' and write 'de' for downloading the German model.

```
python -m spacy download de
```

Press again the up-arrow key, delete 'download de' and replace it with 'validate' to check if the installation was successful.

```
python -m spacy validate
```

Basics

We start by importing spaCy:

```
import spacy
```

Then, load the language libraries. For English:

```
nlp_en = spacy.load('en_core_web_sm')
```

For German:

```
nlp_de = spacy.load('de_core_news_sm')
```

The function nlp() returns raw text as a Doc object that holds information about the tokens, their linguistic features and their relationships.

We use the name that we gave to the library. In the brackets we put some text in quotation marks or the name of the variable that refers to text that we want to analyze.

```
doc = nlp_en()
```

The following returns the analysis of the tokens, their linguistic features and their relationships.

```
for token in doc:
    print(token, token.pos_, token.dep_)
```

Check the spaCy documentation for the specification of the annotations.

https://spacy.io/api/annotation

And to learn more features, you can also check the Notebook of this tutorial.