# NLP for the Web
# Lab 3

**Prof. Dr. Iryna Gurevych, Dr. Thomas Arnold**
Gisela Vallejo, Stefan Thaut

`spaCy`: as discussed in class, this is a useful open-source library that enables the user to perform several NLP tasks with high quality results. It is not only helpful for beginners in NLP but also for advance programmers, who want to integrate NLP features in real products. Given that we have been working from scratch with Python data structures and that it is not allowed to use any other libraries other than internal packages from Python, for this exercise you should only use spaCy, and Numpy and Pandas if needed. Please follow the instructions as given below and in case of questions use our Discussion forum in Moodle.

## Task 1 - 1 point

Using spaCy functions:

a) Tokenize the sentences in file *yelp_polarity.txt* (please ignore class labeling "1"/"0" in the document).

b) Calculate the occurrence (absolute) of each token lowercase (For this you can also use Numpy or Pandas).

c) Extract (print) all tokens with occurrences ≥ 5, excluding punctuation.

d) Using the top 5 most frequent tokens, give an example to explain the internal structure of spaCy (Explain the three components - Please refer to Jupyter Notebook *spaCy_Exercise.ipynb*).

## Task 2 - 2 points

Using the extracted tokens from **Task 1**:

a) Explore at least 5 lexical attributes of these tokens (other than lemma, pos-tag, dependency), and write a small comment for each of them.

b) After observing relevant attributes, write (a) pattern(s) to create groups of similar tokens (it's up to you which similarity metric/criteria you use).

c) Explain your answer and discuss the reason of the chosen pattern.

## Task 3 - 2 points

Using the **whole document** (*yelp_polarity.txt*):

a) Write a function using patterns to extract proper nouns with more than one token.

b) Write an additional function to figure out which verbs and nouns share the same lemma.

c) Explain your functions.

Please upload in Moodle your working Jupyter-Notebook and commented code `before next lab` session (Nov 14th, 4:14pm). Submission format: ExerciseX_YourName.ipynb, e.g. *Exercise3_GiselaVallejo.ipynb*