

NLP Workshop

Get Your Hands Dirty

NLP Workshop

**Alejandro Nardo,
Dominic Schweizer
Noe Thalheim**

Forschungsstelle Digitale Nachhaltigkeit

15. August 2020, Bern

The spaCy logo, featuring the word 'spaCy' in a large, blue, lowercase, sans-serif font.

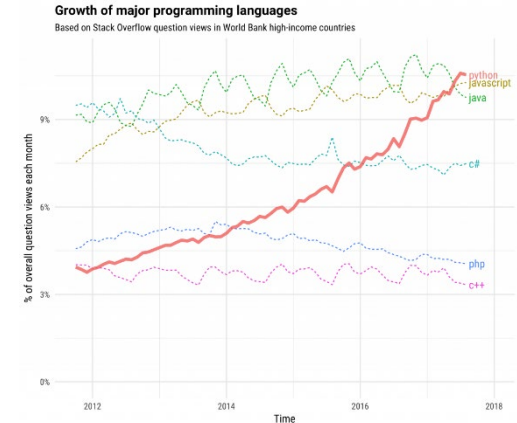
Ablauf

- Theorie
 - Tools
 - POS Tagging
 - Entity Recognition
- Hands-On
 - Basics (POS &ER)
 - Model Training
- Aufsetzen Anaconda



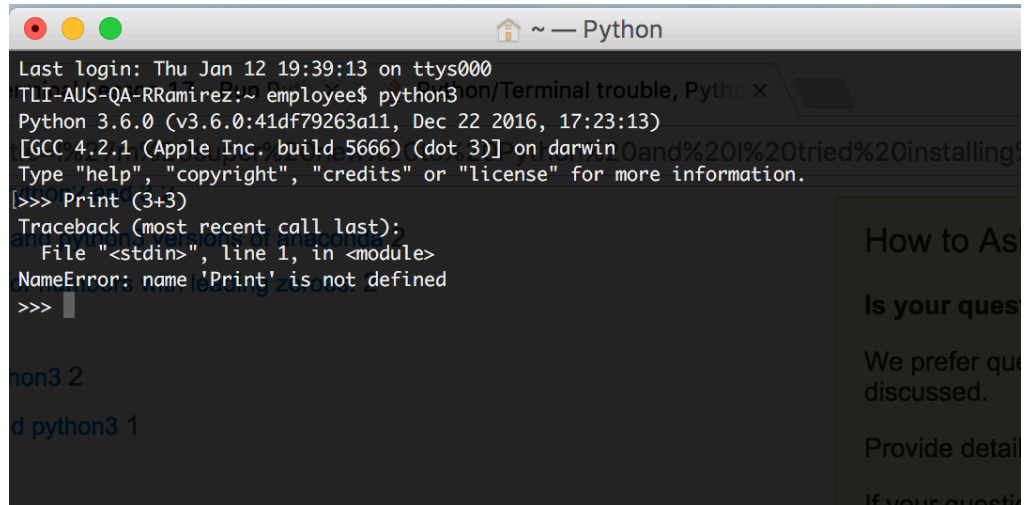
Tools

- Verschiedene Tools & Sprachen
 - Excel
 - R
 - Python
- Focus Python
 - Verbreitet
 - Viele Bibliotheken und Werkzeuge
 - Open-Source



Python

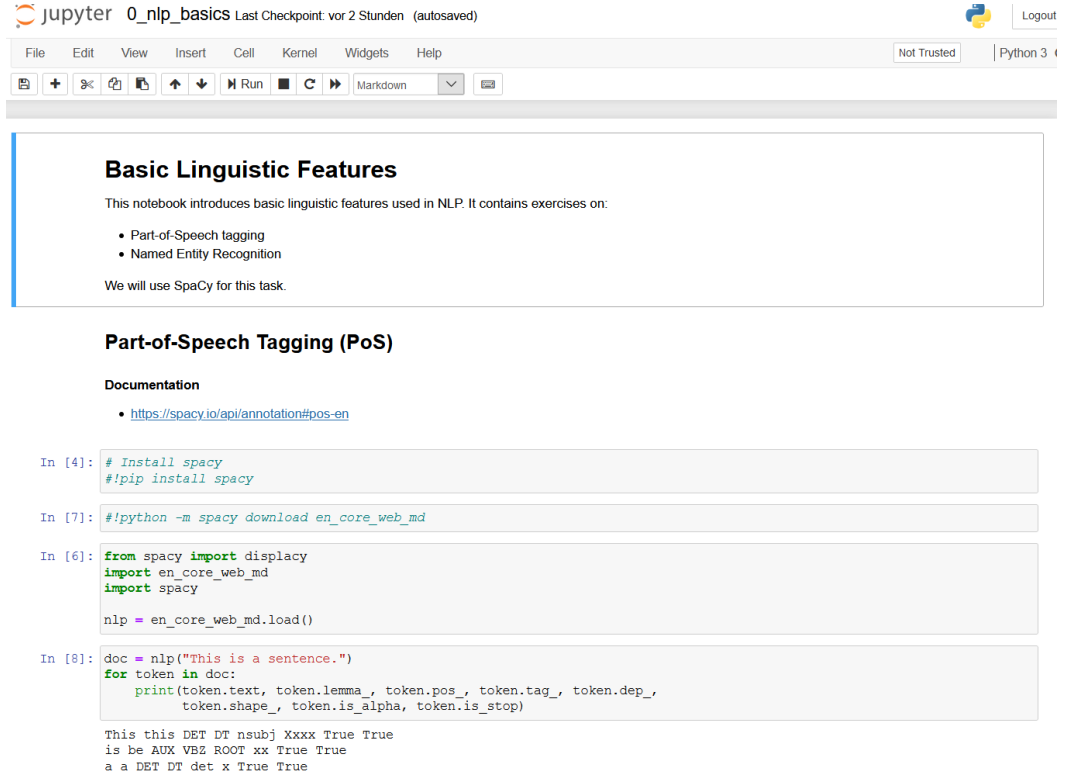
- Einrichten aufwändig
- Terminals & Kommandozeilen



```
Python
Last login: Thu Jan 12 19:39:13 on ttys000
TLI-AUS-QA-RRamirez:~ employee$ python3
Python 3.6.0 (v3.6.0:41df79263a11, Dec 22 2016, 17:23:13)
[GCC 4.2.1 (Apple Inc. build 5666) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> Print (3+3)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'Print' is not defined
>>>
```


Literate Programming mit Jupyter

- Browserbasiert
 - Einfache Darstellung
 - Einsteigerfreundlich
 - Kommentare und Code strukturiert



Jupyter 0_nlp_basics Last Checkpoint: vor 2 Stunden (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted | Python 3



Basic Linguistic Features

This notebook introduces basic linguistic features used in NLP. It contains exercises on:

- Part-of-Speech tagging
- Named Entity Recognition

We will use SpaCy for this task.

Part-of-Speech Tagging (PoS)

Documentation

- <https://spacy.io/api/annotation#pos-en>

```
In [4]: # Install spacy
#!pip install spacy

In [7]: #!python -m spacy download en_core_web_md

In [6]: from spacy import displacy
import en_core_web_md
import spacy

nlp = en_core_web_md.load()

In [8]: doc = nlp("This is a sentence.")
for token in doc:
    print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
          token.shape_, token.is_alpha, token.is_stop)
```

This this DET DT nsubj XXXX True True
 is be AUX VB2 ROOT xx True True
 a a DET DT det x True True

NLP-Tools für Python

- Programmierte und trainierte Werkzeuge
- Python als Schnittstelle zur Kommunikation
- Häufig verwendet:
 - Spacy
 - NLTK

Spacy

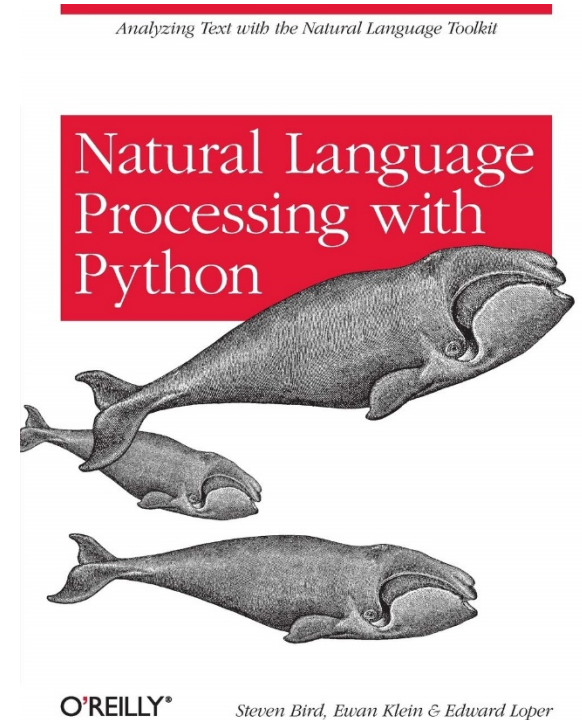
- Entwickelt von explosion AI in Deutschland
- Vorteile:
 - Schnell
 - Einsteigerfreundlich
- Nachteile:
 - spezialisiert



spaCy

NLTK

- Entwickelt durch Community
<https://github.com/nltk>
- Vorteile:
 - Verbreitet
 - Viele Zusatzpakete
- Nachteile:
 - Langsam
 - Sehr umfangreich



Tools: Zusammenfassung

- Viele Wege führen nach Rom
- Fokus auf das Problem und nicht auf die Tools
- Beck's Directive!

Beck's Directive

- Make it Work
- Make it Right
- Make it Fast