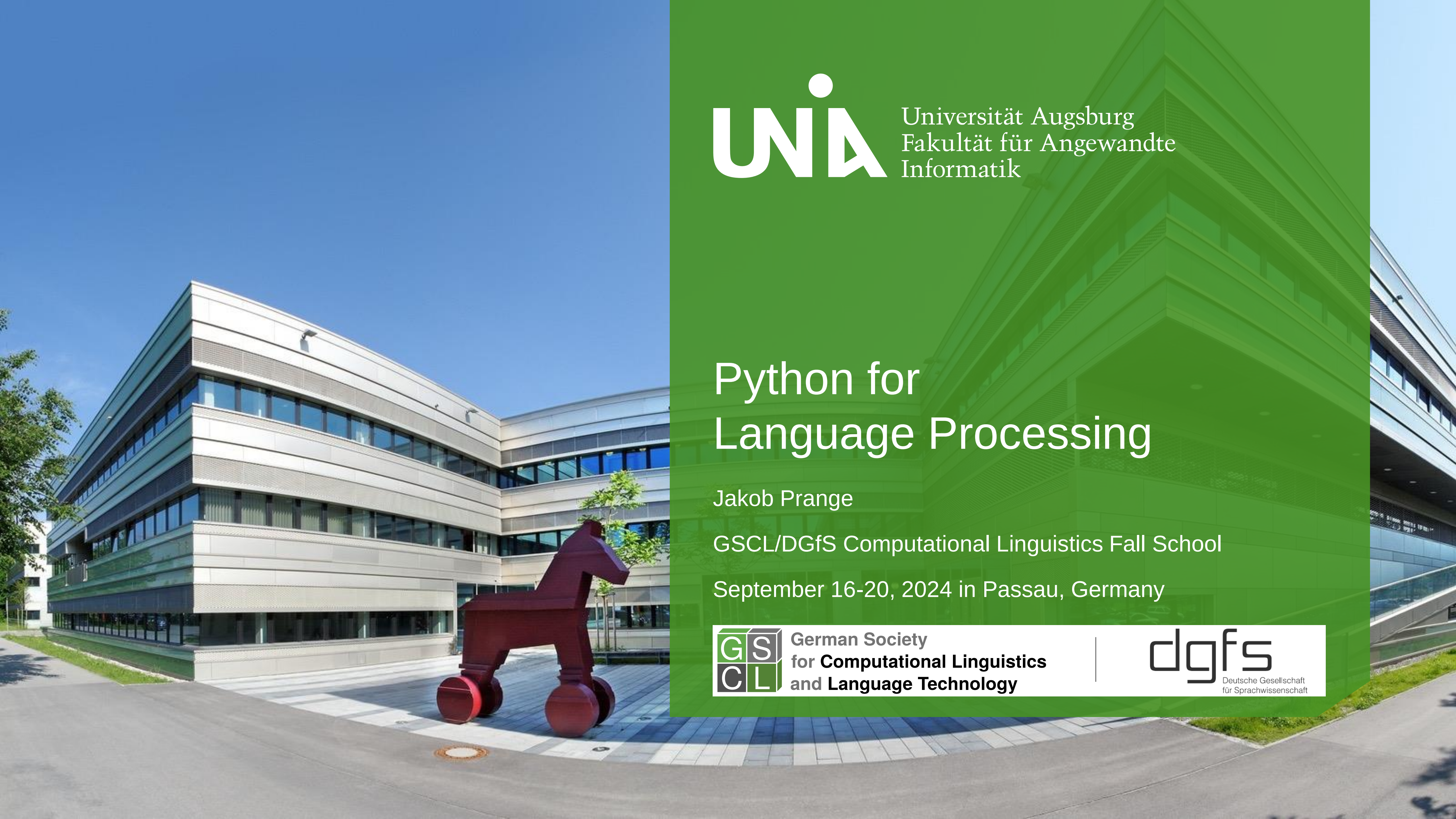


Universität Augsburg
Fakultät für Angewandte
Informatik

Python for Language Processing

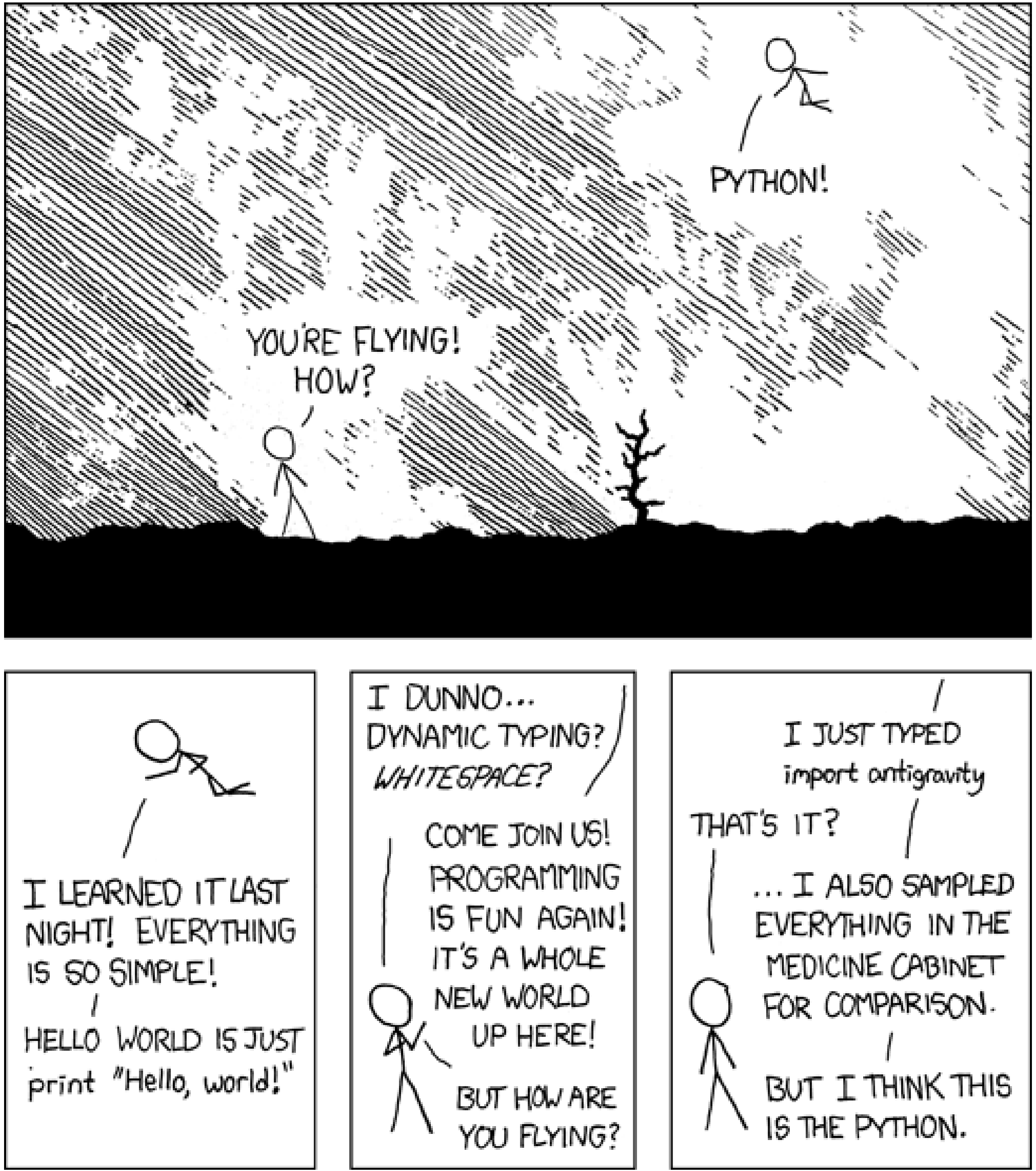
Jakob Prange

GSCL/DGfS Computational Linguistics Fall School

September 16-20, 2024 in Passau, Germany

GSCL German Society for Computational Linguistics and Language Technology

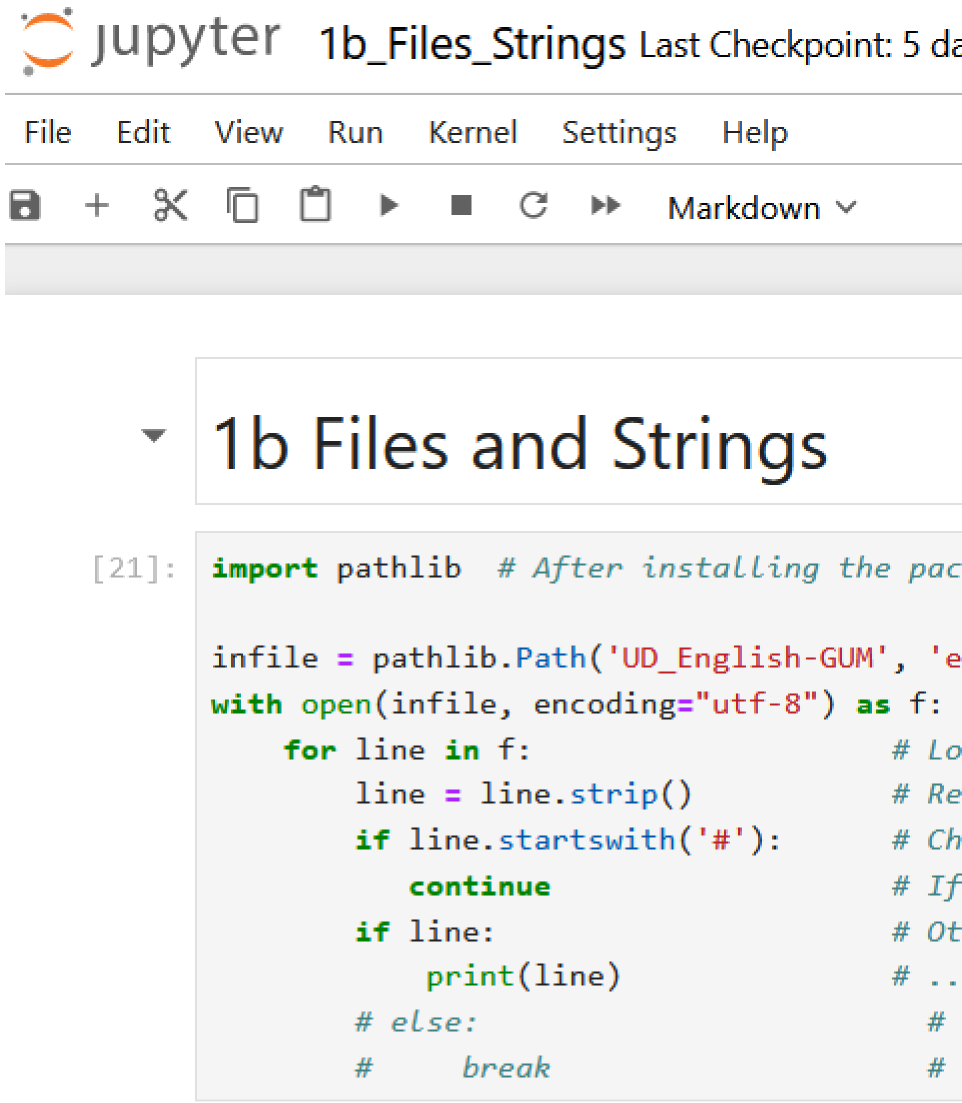
dgfs Deutsche Gesellschaft für Sprachwissenschaft


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

https://xkcd.com/353/

Topics covered: Python programming

- Installing Python
- Setting up Jupyter Notebook
- Opening, reading, writing text files
- String manipulation
- Types, mutability, object identity in memory vs. value equality
- Lists, tuples, sets, dictionaries
- Counting word frequencies
- Branches & loops
- Regular expressions
- Functions



1b Files and Strings

[21]:

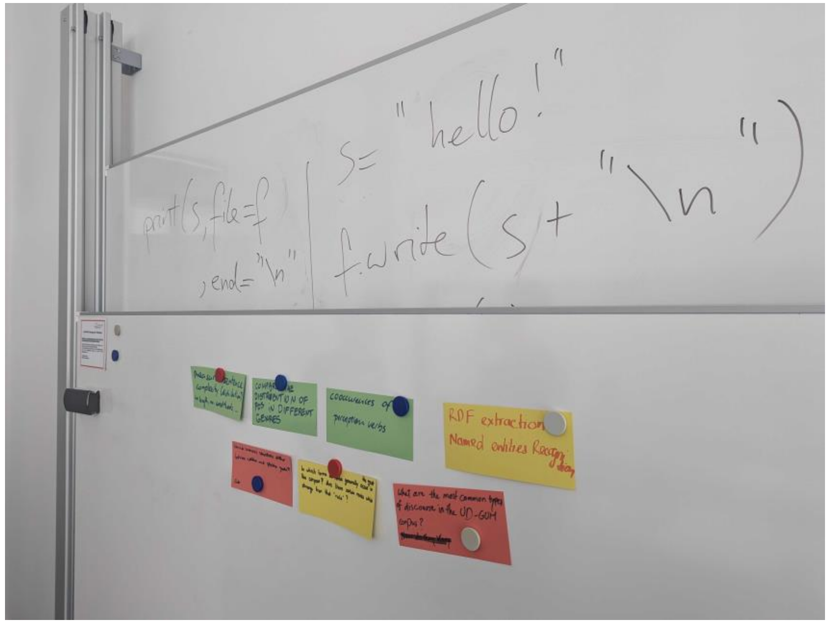
```
import pathlib  # After installing the pac

infile = pathlib.Path('UD_English-GUM', 'e
with open(infile, encoding="utf-8") as f:
    for line in f:                    # Lo
        line = line.strip()           # Re
        if line.startswith('#'):      # Ch
            continue                  # If
        if line:                      # Ot
            print(line)               # ..
        # else:                         #
        #     break                     #
```

Topics covered: Language processing

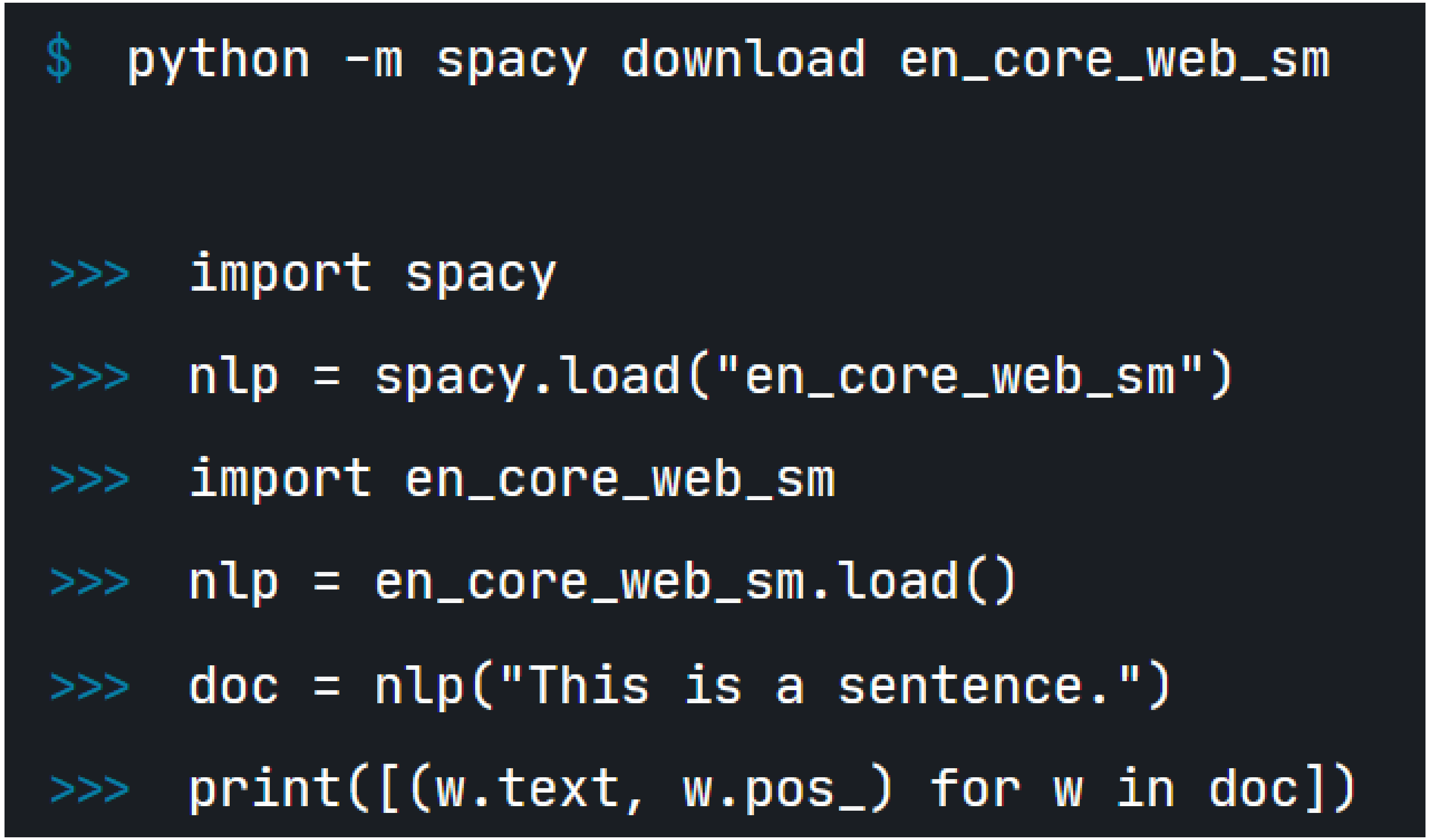
- GUM corpus, CoNLL-U format, UD annotations
- Tokens
- Coming up with computational linguistic research questions
- Designing a corpus analysis project
- Breaking down complex processes into functions
- Implementing everything in Python

- What is a word?
- How do I select data?
- How do I decide which preprocessing steps to do?


print(s, file=f
, end="\n"
s = "hello!"
f.write(s + "\n")
COMPARE THE DISTRIBUTION OF POS IN DIFFERENT GENRES
cooccurrences of perception verbs
RDF extraction
Named entities Recognition
What are the most common types of discourse in the UD-GUM corpus?

Resources

https://spacy.io/models



```
$  python -m spacy download en_core_web_sm


>>>  import spacy
>>>  nlp = spacy.load("en_core_web_sm")
>>>  import en_core_web_sm
>>>  nlp = en_core_web_sm.load()
>>>  doc = nlp("This is a sentence.")
>>>  print([(w.text, w.pos_) for w in doc])
```

Resources

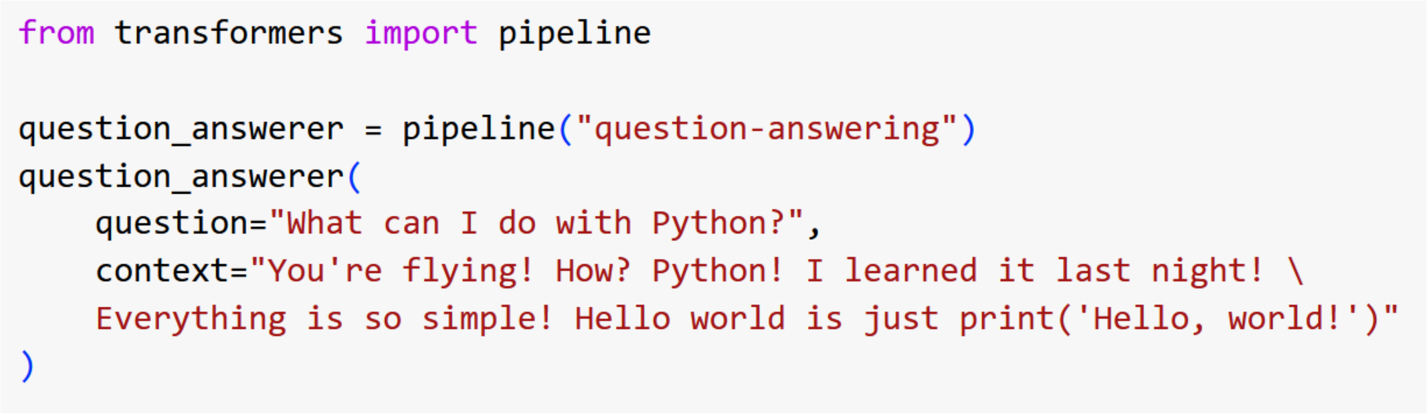
https://huggingface.co/

General overview demo:

https://colab.research.google.com/github/huggingface/notebooks/blob/master/course/en/chapter1/section3.ipynb

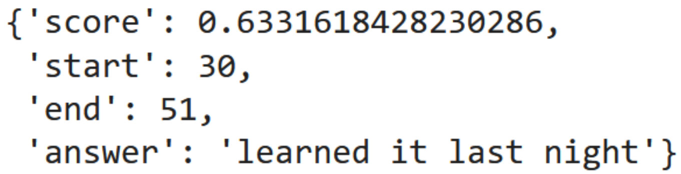
NLP course:

https://huggingface.co/learn/nlp-course/



```python
from transformers import pipeline

question_answerer = pipeline("question-answering")
question_answerer(
    question="What can I do with Python?",
    context="You're flying! How? Python! I learned it last night! \
    Everything is so simple! Hello world is just print('Hello, world!')"
)
```



```
{'score': 0.6331618428230286,
 'start': 30,
 'end': 51,
 'answer': 'learned it last night'}
```

Resources

Handling, filtering, combining tabular data

https://pandas.pydata.org/

4c_Data_Science.ipynb

Doing parallel math with vectors, matrices, high-precision floating point numbers

https://numpy.org/

Thank you!

Course tutors: Artur Romazanov & Jonas Barth

GSCL German Society for Computational Linguistics and Language Technology

dgfs Deutsche Gesellschaft für Sprachwissenschaft

President:
Prof. Dr. Annemarie Friedrich

Speaker and Fall School Organizer:
Prof. Dr. Annette Hautli-Janisz

HLT – Human Language Technology @ Augsburg

Universität Augsburg
Fakultät für Angewandte
Informatik

Prof. Dr. Annemarie Friedrich

Dr. Jakob Prange

We are hiring!