



DIS08 – Data Modeling

07 – Basic Data Processing with Python

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: WS 2021

Technology
Arts Sciences
TH Köln

Agenda

Last weeks

- The shell
- GIT / GitHub
- Regular expressions
- Open data
- Data cleaning in a nutshell

This week

- We will start with Python!
- Our final goal is to **program our own web scraper** to gather data from the web for further analyses.

www.ksta.de/koeln/zahlenchaos-darum-war-der-inzidenzwert-in-koeln-nicht

Anzeigen | Abo | wir helfen | Shop | Service | Leserreisen | FORUMBLAU | Specials | Jobbörse | Immobilien | Traueranzeigen

Kölner Stadt-Anzeiger

Köln Sport Schule Restaurants Immobilien Podcast Newsletter E-Paper KStA Blog Arena

Abo 🔍 👤 ☰

+++EILMELDUNG+++ Nach Henning Krautmacher Zwei weitere Mitglieder der Höchner mit Coronavirus infiziert

Kölner Stadt-Anzeiger ▶ Köln ▶ Falsche Zahlen vom RKI und LZG: Darum ist der Inzidenzwert für Köln aktuell nicht verlässlich

Zahlenchaos Darum war der Inzidenzwert in Köln nicht verlässlich



Schedule for lectures WS 2021/22

15.10.21	Introduction, Markdown		first week (no tutorial)
22.10.21	Unix Shell		Markdown and style guides
29.10.21	Versioning, Git, GitHub		Shell tutorial
05.11.21	Regular Expressions		Hands-on Git, GitHub
12.11.21	CSV, JSON, XML		Regex tutorial
19.11.21	OpenData, Tidy Data Principles		open.cologne open data
26.11.21	Project week (no lecture)		Submit assignment 1 (no tutorial)
03.12.21	Python: Data structures		Hands-on Python
10.12.21	Python: Files, folders and more		Hands-on Python
17.12.21	Python: Pandas		Hands-on Python
24.12.21	Christmas (no lecture)		Christmas (no tutorial)
31.12.21	Christmas (no lecture)		Christmas (no tutorial)
07.01.22	Python: Web Scraping		Hands-on Python
14.01.22	Q&A - Summary		Hands-on Python
04.02.22			Submit assignment 2

We will work with Jupyter Notebooks...

The screenshot shows a Jupyter Notebook interface in a web browser. The browser address bar shows `localhost:8888/notebooks/dis08/07-python-data/dis08-python-data.ipynb#`. The notebook title is `dis08-python-data` with a status of `Last Checkpoint: vor 20 Minuten (autosaved)`. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar with icons for saving, running, and other actions. The notebook content consists of a text cell and a code cell.

The text cell contains the following text:

string. The `str()`, `int()`, and `float()` functions will evaluate to the string, integer, or floating-point number form of the value they are passed.

In a nutshell, you learnt about

- math operators,
- data type (integer, float, strings, ...),
- string concatenation and replication,
- variables,
- running code within the Spyder IDE / Jupyter Notebooks,
- basic functions (`print()`, `input()`, `len()`, `str()`, `int()`, and `float()`).

Small exercise on data types, math, and string concatenations

The code cell shows the following code:

```
In [1]: 1 # Where is the bug in the following code snippet? How can you correct it?
        2 currentAge = 39
        3 print('You will be ' + currentAge + 1 + ' in one year.')
```

The code cell shows a `TypeError` traceback:

```
-----
TypeError                                Traceback (most recent call last)
<ipython-input-1-cbbb803f49df> in <module>
      1 # Where is the bug in the following code snippet? How can you correct it?
      2 currentAge = 39
----> 3 print('You will be ' + currentAge + 1 + ' in one year.')
```

The error message is: `TypeError: can only concatenate str (not "int") to str`

For today we will focus on...

Repeating some basic Python stuff that you already heard from Prof. Strahringer or Dr. Schaible/Dr. Carevic (very quick and dirty repetition)

- Python basics
- Flow Control
- Functions
- Lists
- Dictionaries and structuring data
- Manipulation Strings
- Regex in Python
- Reading and writing files