UNIVERSITY OF LUCERNE

The ABC of Computational Text Analysis

#7 Working with (your own) Data

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

10 April 2025

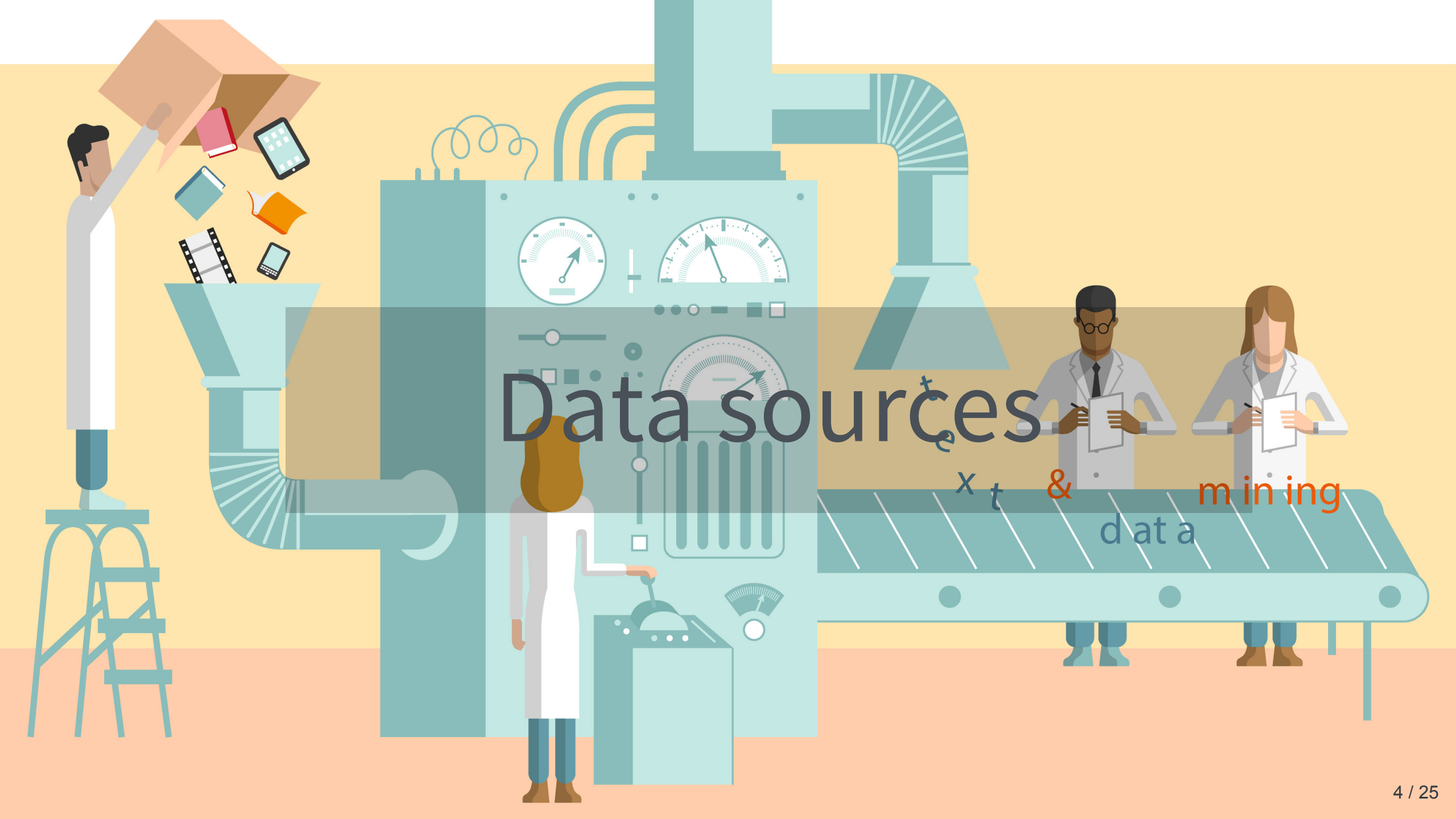
Recap last lecture

- introducing Python 🐍
- learning programming concepts & syntax

    data types, loops, indexing, functions…
- working with VS Code Editor

Outline

- learn about available data sources 🗞️
- analyze and justify your own data
  - convert `.pdf` into `.txt` 🔄
  - count words in Python 🧮


Data sources
& mining
data

What data sources are there?

- broadly social
  - newspapers + magazines
  - websites + social media
  - reports by NGOs/GOs
- scientific articles
- economic
  - business plans/reports
  - contracts
  - patents

👉 basically, any textual documents…

How does the data look like?

Any text is data, yet some formats are more suitable

1. datasets like `.csv` or `.tsv` 🥰
2. plain text like `.txt` 🙂
3. text in other formats like `.pdf` 😬

Some great (historical) datasets

Datasets in `.csv` ready off-the-shelf

- 1 August speeches by Swiss Federal Councilors

  provided via course repo, converted from PDFs
- Human Rights Reports by various NGOs
- United Nations General Debate Corpus
- Corpus of Resolutions: UN Security Council
- Inaugural Speeches by US Presidents
- NewsWire

  2.7M US news articles, 1878-1977

😓 There are still not many.

Dedicated search engines for datasets

Use case: Search for existing datasets

- Harvard Dataverse

  open scientific data repository

- Google Dataset Search

  Google for datasets basically

👉 search for a topic followed by `corpus`, `text collection` or `text as data`

Search techniques 🔍

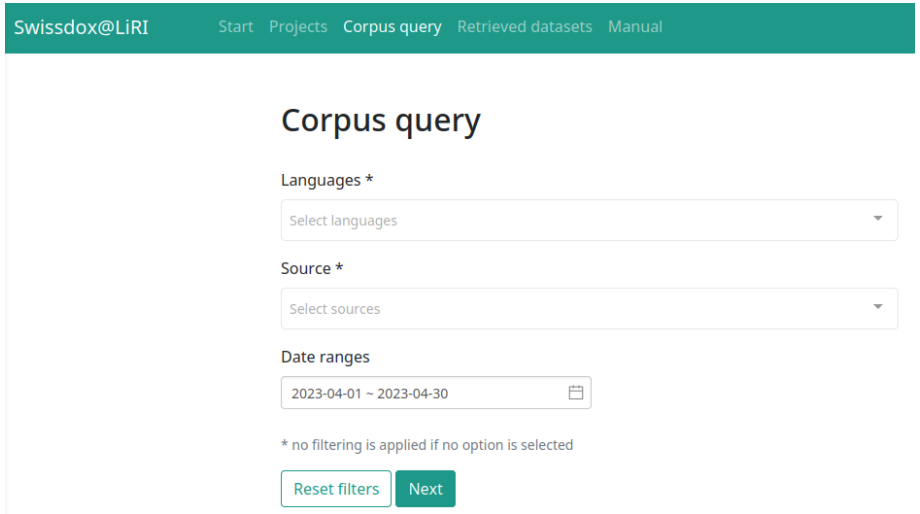
Make your (Google) web search more efficient by using dedicated operators.
Examples:

- `"computational social science"`
- `site:nytimes.com`
- `nature OR environment`

Swissdox: A game changer

Assemble a news dataset and download as `.tsv`

- over 250 Swiss newspapers
- historical and updated daily
- needs registration (free)



More publishers

- Nexis Uni
  - international newspaper, business + legal reports
  - licensed by the university
- Project Gutenberg and HathiTrust
  - massive collection of books
  - open, HathiTrust requires agreement
- Internet Archive
  - way-back-machine for web
  - find “lost” web content

👉 check out other resources licensed by ZHB

Interesting sources as PDFs

Any organization of your interest 👍

- Party Programmes across Europe

  covers over 1000 parties from 1920 until today in over 50 countries

- Swiss voting booklets

  from 1977 until today

- Swissvotes

  collection of resources on Swiss public votings

- Curia Vista

  Swiss parliamentary debates

- Nestlé Annual Reports
- University of Zurich Annual Reports

Scraping PDFs from websites

Use case: Swiss voting booklets

- `wget` to download any files from the internet

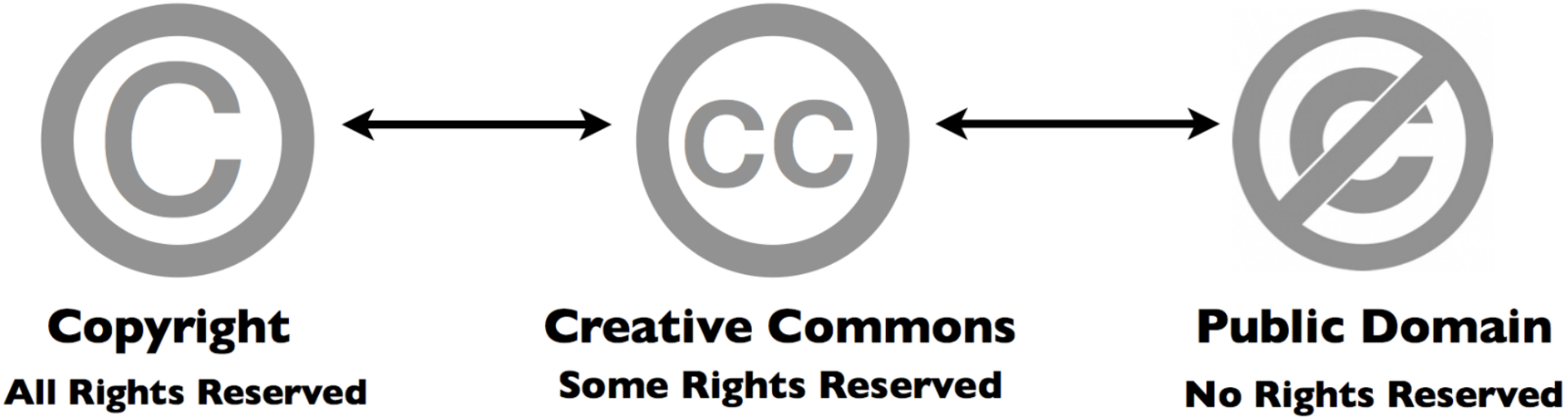
```
# get a single file
wget EXACT_URL

# get all linked pdf from a single webpage
wget --recursive --accept pdf -nH --cut-dirs=5 \
--ignore-case --wait 1 --level 1 --directory-prefix=data \
https://www.bk.admin.ch/bk/de/home/dokumentation/abstimmungsbuechlein.html

# --accept FORMAT_OF_YOUR_INTEREST
# --directory-prefix YOUR_OUTPUT_DIRECTORY
```

Data is no free lunch 😢

Data is property 🚫

… and has rights too

- copyrights may further limit access to high quality data
- check the rights before processing data



Imperfect data: A tail of bias

- noise in text

  non-content (e.g. table of content), inconsistent spelling

- archive holes

  lost or uncollected data

- selective corpus curation

  supposition that key-word(s) captures topic

- social bias

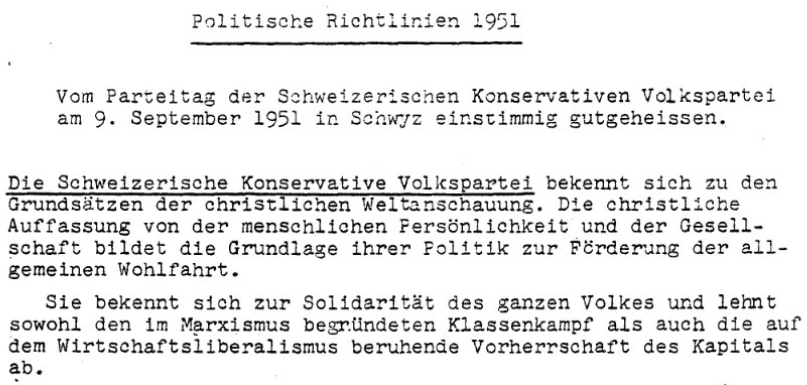
  view from somewhere, stereotypes

👉 think about the data and mitigate issues

. A world for humans …
… and a jungle of file formats.

PDF: Digitized or digital?

Two flavors of `.pdf` documents



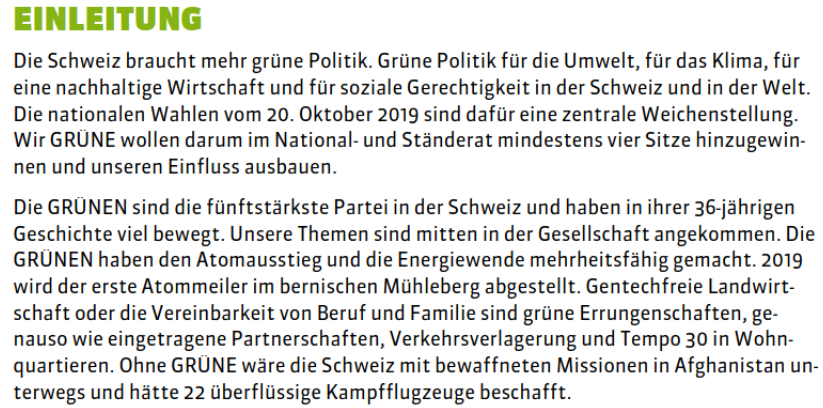
Politische Richtlinien 1951

Vom Parteitag der Schweizerischen Konservativen Volkspartei am 9. September 1951 in Schwyz einstimmig gutgeheissen.

Die Schweizerische Konservative Volkspartei bekennt sich zu den Grundsätzen der christlichen Weltanschauung. Die christliche Auffassung von der menschlichen Persönlichkeit und der Gesellschaft bildet die Grundlage ihrer Politik zur Förderung der allgemeinen Wohlfahrt.

Sie bekennt sich zur Solidarität des ganzen Volkes und lehnt sowohl den im Marxismus begründeten Klassenkampf als auch die auf dem Wirtschaftsliberalismus beruhende Vorherrschaft des Kapitals ab.

Digitalized PDF made from a scanned page



EINLEITUNG

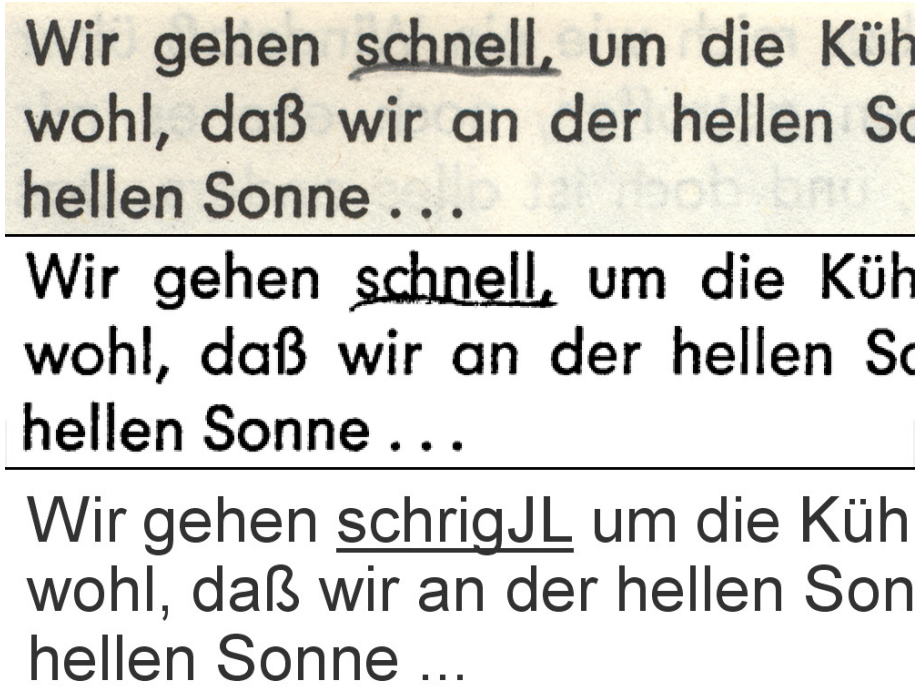
Die Schweiz braucht mehr grüne Politik. Grüne Politik für die Umwelt, für das Klima, für eine nachhaltige Wirtschaft und für soziale Gerechtigkeit in der Schweiz und in der Welt. Die nationalen Wahlen vom 20. Oktober 2019 sind dafür eine zentrale Weichenstellung. Wir GRÜNE wollen darum im National- und Ständerat mindestens vier Sitze hinzugewinnen und unseren Einfluss ausbauen.

Die GRÜNEN sind die fünftstärkste Partei in der Schweiz und haben in ihrer 36-jährigen Geschichte viel bewegt. Unsere Themen sind mitten in der Gesellschaft angekommen. Die GRÜNEN haben den Atomausstieg und die Energiewende mehrheitsfähig gemacht. 2019 wird der erste Atommeiler im bernischen Mühleberg abgestellt. Gentechfreie Landwirtschaft oder die Vereinbarkeit von Beruf und Familie sind grüne Errungenschaften, genauso wie eingetragene Partnerschaften, Verkehrsverlagerung und Tempo 30 in Wohnquartieren. Ohne GRÜNE wäre die Schweiz mit bewaffneten Missionen in Afghanistan unterwegs und hätte 22 überflüssige Kampfflugzeuge beschafft.

Native PDF converted from digital document (e.g., docx)

Optical Character Recognition (OCR)

- OCR ~ convert images/scans into text
  - may support handwriting + Fraktur texts
- conversion steps
  - convert to b/w image
  - run OCR model
  - correct spelling issues



Steps when performing OCR

Common conversions

Your text is of a particular type

⬇️

digital native documents

`.pdf`, `.docx`, `.html`

⬇️

extract text without formatting

⬇️

scans of (old) documents

`.pdf`, `.jpg`, `.png`

⬇️

Optical Character Recognition (OCR)

machine-readable formats: `.txt` or `.csv` ✅

What data are you interested in?

Think about your mini-project

- present in last lecture
- analyze any collection of text documents
    - compare historically
    - compare between actors
- form groups of 2-3 people
- requirements
    - apply quantitative measures on multiple documents
    - interpret and present results in class
    - share executable script

Illustration of text analysis generated by Image Creator from

Assignment #2 ✍️

- get/submit via OLAT
  - starting tomorrow
  - deadline: 19 April 2025, 23:59
- discuss issues on OLAT forum

Let’s start coding ✨

In-class: Exercises I

1. Make sure that your local copy of the Github repository KED2025 is up-to-date with `git pull`.
2. Use `wget` to download *cogito* and its predecessor *uniluAKTUELL* issues (PDF files) from the UniLu website. Start with downloading one issue first and then try to automatize the process to download all the listed issued using arguments for the `wget` command.
3. Convert the *cogito* and *uniluAKTUELL* PDF files into TXT files. You can use the code given for native, digital PDFs (no OCR needed). Try with a single issue first and then write a loop to batch process all of them.
4. What is the University of Lucerne talking about in its issues? Use the commands of the previous lectures to count the vocabulary.
5. Do the same as in 4.), yet analyze the vocabulary of *cogito* and *uniluAKTUELL* issues separately. Does the language and topics differ between the two magazines?


Questions?