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

06 April 2024

Recap last Lecture

- describe text as pattern using RegEx
- extract + replace textual parts 🧩

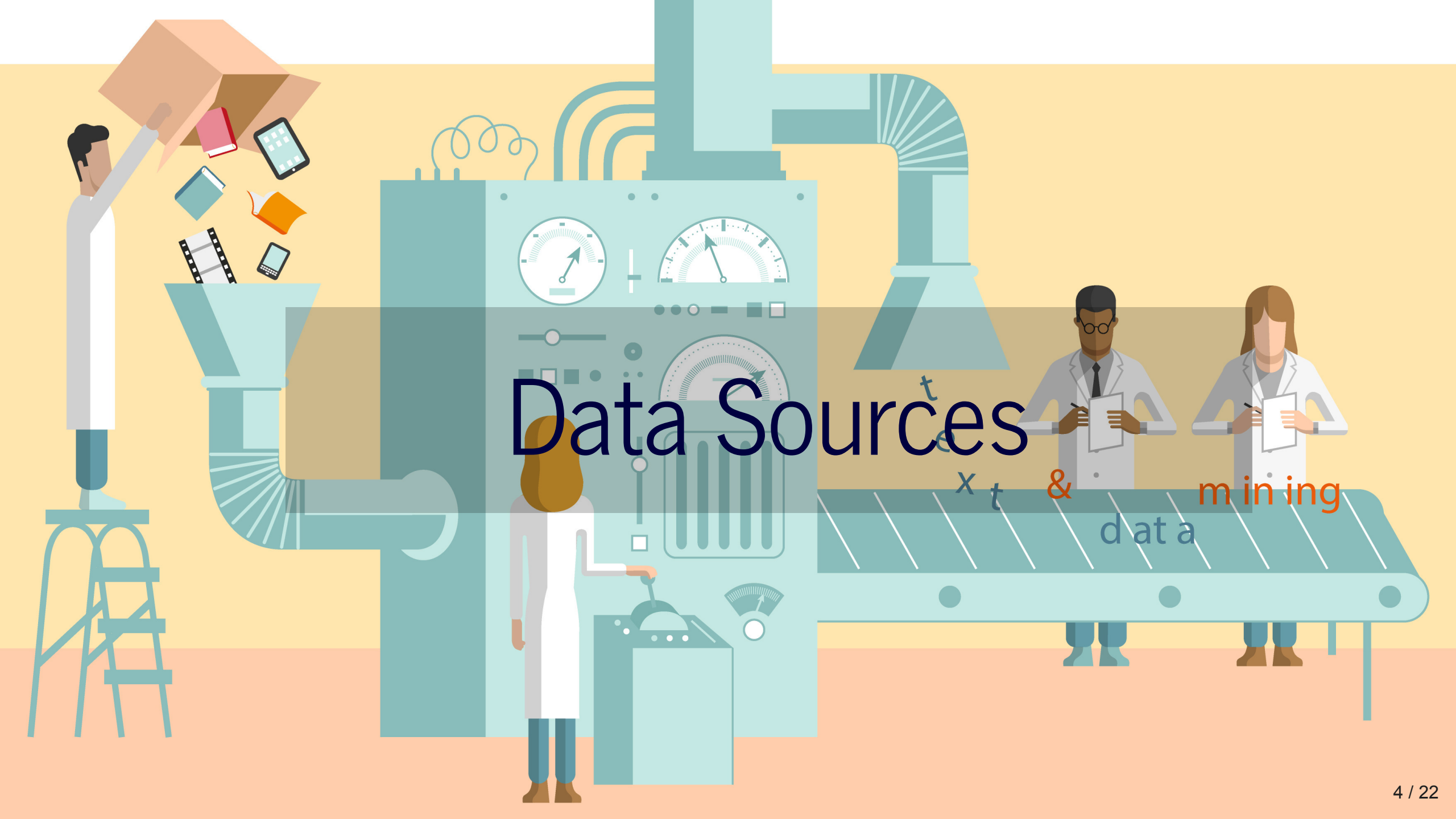
  literal: `text a b c`

  meta: `\w \s [^abc] *`

  power of `.*`
- questions regarding assignment 2

Outline

- learn about available data sources
- Build and justify (!) your own dataset
  - any text ✅
  - “any” format ✅
  - from anywhere ✅


Data Sources
t
e
x
t
&
m in ing
d at a

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…

Some great (historical) Corpora

Ready off-the-shelf, machine-readable

- 1 August speeches by Swiss Federal Councilors

  provided via course repo
- Human Rights Reports by various NGOs
- United Nations General Debate Corpus
- Inaugural Speeches by US Presidents

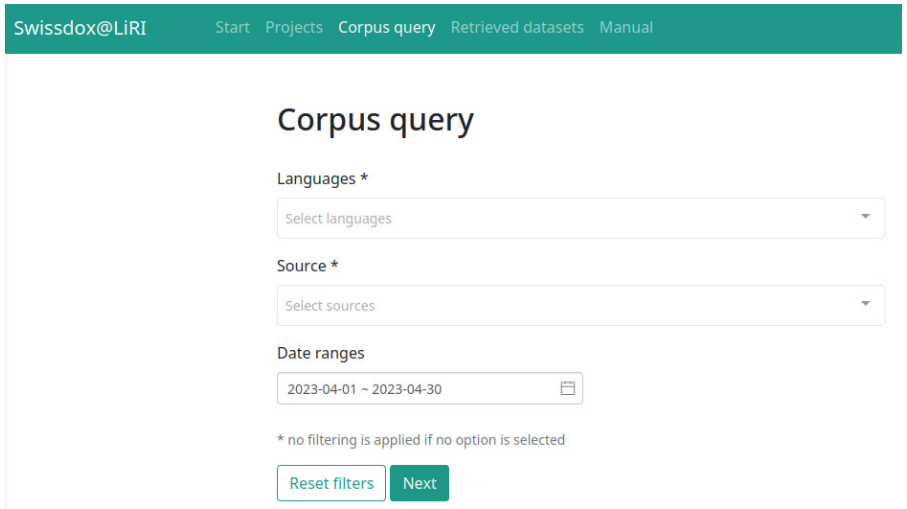
...

😓 There are still not many.

Swissdox: A game changer

Assemble corpus and export as `.csv`

- over 250 Swiss newspapers
- historical and updated daily
- needs registration



Scraping PDF from Websites

Use case: Swiss voting booklet

- `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/abstimmungsbuechle

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

::: notes

- bis hierher: Wie Daten in txt Format bringen

More Interesting Publishers

- Project Gutenberg and HathiTrust
  - massive collection of books
  - open, HathiTrust requires agreement
- Nexis Uni
  - international newspaper, business + legal reports
  - licensed by the university
- Constellate
  - scientific articles of JSTOR across disciplines
  - provides an easy dataset builder

👉 check out other resources licensed by ZHB

::: notes - Nexis ähnlich Swissdox, aber international - Zugang tlw. über ezproxy - Constellate - kurze Demo von Constellate - brandneue Platform - einfache Zusammenstellung von JSTOR Artikeln - sehr gute Metadaten - auch gut für

Search 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`

::: notes - fortschrittliche Wissenschaft veröffentlicht nicht nur Papers, sondern auch Daten und Code - computergestützte Textanalyse ist aber immer noch Nische - Suchmaschinen für Datensätze - allerlei Datensätze, primär aus Wissenschaft :::

Search Techniques 🔍

Make your web search more efficient by using dedicated tags. Examples:

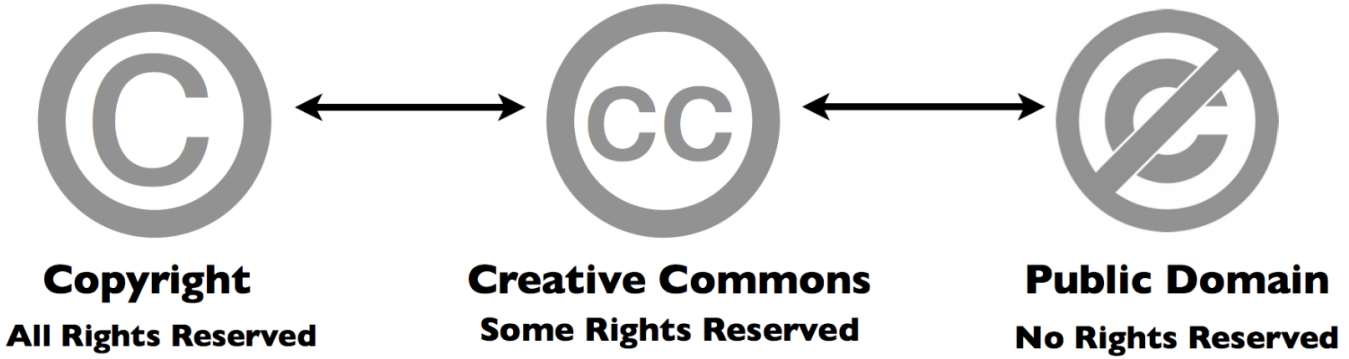
- `"computational social science"`
- `site:nytimes.com`
- `nature OR environment`

::: notes - Quotes für Wörter die zusammen gehören - Summe von Wörter haben evtl andere Bedeutung - Boolean Search - OR / AND :::

Data is Property 🚫

... and has rights too

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



::: notes - Zugang zu Daten nicht immer einfach - open data unterschiedlich unterstützt - Datenbereitstellung oftmals Teil von Geschäftsmodell - dann restriktiv - oftmals ist Verwendung nicht geregelt - nutzt Graubereich :::

Interesting Resources

- Party Programmes across Europe

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

- Swissvotes

  collection of resources on Swiss public votings

- Swiss voting booklets

  from 1978 until today

- 1 August speeches by Swiss Federal Councillors
- Nestlé Annual Reports
- University of Zurich Annual Reports
- … any organization of your interest 👍

What data are you interested in?

Think about the topic of your mini-project

Illustration of text analysis generated by Image Creator from Microsoft Bing

Preparing your own Data

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

Common Conversions

news, press releases, reports from organizations

⬇️

digital native documents

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

⬇️

convert to `.txt`

⬇️

scans of (old) documents

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

⬇️

Optical Character Recognition (OCR)

machine-readable ✅

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**


Questions?

In-class: Exercises I

1. Make sure that your local copy of the Github repository KED2024 is up-to-date with `git pull`. Check out the data samples in `materials/data` and the scripts to extract their text in `materials/code`.
2. Decide on one use-case that interests you most. Install the missing tool with the commands given on the respective slides (e.g., `pandoc`, `imagemagick`, `poppler`).
3. **Apply the commands to reproduce on the given data. Test them on your own data. Check the resources. Ask questions. Think about your mini-project.**

In-class: Exercises II

1. 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.
2. Convert the *cogito* and *uniluAKTUELL* PDF files into TXT files using `pdftotext`. Try with a single issue first and then write a loop to batch process all of them.
3. What is the University of Lucerne talking about in its issues? Use the commands of the previous lectures to count the vocabulary.
4. Do the same as in 3.), yet analyze the vocabulary of *cogito* and *uniluAKTUELL* issues separately. Does the language and topics differ between the two magazines?

In-class: Exercises III

1. Use `wget` to download a book from Project Gutenberg and count some things (e.g., good/bad, joy/sad).
2. `wget` is a powerful tool. Have a look at its arguments and search for more examples in tutorials on the web.Have a nice
   Easter break! {data-background-image=../images/easter-eggs.jpg}