Problemset 1

Your task is to scrape articles from Deutsche Welle (DW) along with additional metadata about the articles.

1 Exercise 1

Start with the following article from Deutsche Welle

- Use the Python Package BeautifulSoup to extract the following information from this website and store it in a Python dictionary:
  - author
  - date
  - title
  - summary (text printed in bold letters at the beginning)
  - main text
  - related topics (Minorities, Women's rights, . . .)
- Analyze the article: what additional information (metadata or other) related to this article is available? Make a list of such items. Then pick one of these items, extract it also from this website, and update your dictionary with this item.
- Print the dictionary.

2 Exercise 2

Now visit the following search page for English articles.

- Your task is to scrape all articles (only media type "Article") from 01.03.2024 extracting the same type of information as in exercise 1. Store the information in a Pandas DataFrame such that a single row represents one article, and that the columns represent the extracted features (author, date, . . .). Hint: First, you need to extract the urls to the single articles published on that day. Secondly, you need to open, parse, and extract information from each of these articles.
- Scrape politly by including delays between requests and by scraping not more data than you actually need.
- Print the shape of the DataFrame and display its first five rows. Display it in such a way that all columns are visible.