Problemset 3

Use the dataset **dw_articles.json** for this problemset.

Your task is to carry out an exploration of the corpus of Deutsche Welle (DW) articles, focusing both on text mining aspects and on netwok analysis. In particular:

1. Visualize some core patterns of the dataset: number of (1) articles per day, (2) per category, and (3) per region
2. Enrich the dataset with the following features derived from Part-of-Speech (POS) tagging and Named Entity Recognition (NER): (1) persons, (2) locations, and (3) nouns. Then create visualizations to show rankings for each of these features. How often are these features mentioned in the entire corpus? **Important Hint:** Note that the data set is large. Hence, test your approach on a small subset of the data first, to see whether it works. This may give you also a plausible estimate how long the processing is going to take for the entire dataset.
3. We want to understand better, which persons are connected closely with each other. For this purpose, create a network of **person co-occurrences** and visualize it appropriately. Fine-tune the visualization. Try to find a good balance between providing rich and interesting information and not overloading the plot. Summarise in your own words: what are interesting and surprising insights that can be gained?
4. Create a network of **person and location co-occurrences**. Visualize and finetune the visualization in a similar way as above to make it as informative as possible. Summarise in your own words: what are interesting and surprising insights that can be gained?