Exercise Week 6

Sprachverarbeitung (VL + Ü)

Nils Reiter, nils.reiter@uni-koeln.de

May 9, 2023 (Summer term 2023)

The file `/teaching/summer-2023/sprachverarbeitung/data/titanic.csv` contains passenger data from the crossing of the Titanic, which sank in the North Atlantic on 15 April 1912.[1] The file contains information on the passengers, together with details of who survived the sinking and who did not.

The following tasks require a mixture between reading/thinking and applying command line tools. All tasks can be solved purely manual, but this is cumbersome and takes a lot of time.

1. Find out what features are given for the passengers.
2. Assign a data type to each feature.
3. For category features: Find out which categories there are, and how common each is.
4. For numeric features: Find out what range of values they assume (minimum and maximum).

Hints

- One column is particularly tricky, because it contains relatively free text, including commas and quotation marks. Maybe first clean up this column by removing commas etc. or remove the column entirely. Both can be done with regular expressions as we have used them before.
- Accessing individual columns could be done with a similar trick as when we defined the context for the concordance in words instead of characters: Define some content of the column and the character that separates the columns as a group, and ask this group to be repeated.

[1] `https://en.wikipedia.org/wiki/Titanic`