Energy Informatics

https://proglang.informatik.uni-freiburg.de/teaching/energy-informatics/2018ws/

Exercise Sheet 6 – Pandas and Bundestag data

2018-11-27

Pandas

Pandas is a library to manipulate arrays of data called "dataframes". Pandas provide many operations similar to SQL queries. The documentation is available on https://pandas.pydata.org/ and can be installed with `pip install xlrd pandas`.

```
a = pd.read_excel("movies.xls") # Load a spreadsheet
a.head() # See the top of the spreadsheet
a.describe() # Simple stats
a.income # or a['income'] : Array containing the column "income"
a[a.income > 10000] # Rows where the income is greater than 10000
a.groupby("Company").sum().income # Sum of income by company
```

In Exercise sheet 2, we examined the Bundestag data from http://www.bundestag.de/parlament/plenum/abstimmung/liste using a spreadsheet software. We now examine it again, but using python and pandas we can easily analyse multiple voting sessions.

Exercise 1 (Analysis of individual sessions)

- Download the data from multiple (5+) sessions and load them all in python using a list of dataframes.
- Check that the tables are well-formed (at most one vote per row, ...).
- Computes simple per-session metrics: proportion of yes/no, participation, agreement among factions...

Exercise 2 (Analysis on multiple sessions)

- Draw the graph (using mathplotlib) of agreements among factions in function of times.
- Who is the most present parliaments member? The most absent? Is there a faction that is, generally, more absent?
- Who is the most independent parliamentary (who votes along party-lines the least often)?
- Find the vote that was the most divisive, included *inside each faction*.