Methods for unstructured data

Assignment

University of Basel, FS 2022

Helge Liebert

Assignment

Assignment

1. Web scraping assignment (20%)
2. Prediction assignment (30%)
3. Research proposal (50%)

- Email: helge.liebert@econ.uzh.ch.
- File transfer option: https://www.switch.ch/services/filesender/

Assignment I: Web scaper

Choose ***one*** of the following.

Open Data Tanzania

https://tanzania.opendataforafrica.org

- For each region in Tanzania, get the basic facts (top of the page: capital, languages, area, ...), population and mortality statistics (tables under links 'Population projections, total' and 'Deaths').

Books to sell

https://books.toscrape.com/index.html

- For every category, collect all books with title, price, availability, stock and product description.

Swiss Tax Calculator

https://swisstaxcalculator.estv.admin.ch/#/taxburden/income-wealth-tax

- Collect municipal income tax burden statistics for the following parameters: All payers, atheist, geographical comparison (all municipalities), all years, income 50'000 CHF.

Assignment I: Web scraper

- Comment your code or submit a short description alongside.
- A proof-of-concept restricted to the first couple of regions/books/... is fine.
- Should be reproducible for me with minimal effort (e.g. just `Rscript your-file.r` or `python3 your-file.py`).
  - Do not include hardcoded absolute paths.
  - Create (sub-)directories from within R/Python.
- Submit code only, not data.
- Individual assignment, accounts for 20% of the final grade.

Assignment II: Prediction competition

- Download the data from this link.
- Input is the Kiva data for a set of Latin American countries.
- Develop a classifier to predict whether the loan is for an enterprise in either the food or the agricultural sector.
- Label is already recorded in the data as *target*.
- The data has not been processed in any way.
- Anything contained in the data besides the sector name is fair game to be used as modeling input.
- You are free to transform the inputs in any way you deem reasonable.
- You have to pick a single, final model to submit (can be an ensemble).

Assignment II: Prediction competition

- Comment your code to explain what you are doing and why you made your choices.
- Submit your code and a serialized version of the classifier (e.g. using `saveRDS()` in R or `joblib.dump()` in Python).
- After submission, I will upload a private test set.
- You obtain predictions from your estimated model for the new data (applying any transformations you made to the data, but no retraining!).
- You send me the performance of your model on the test data (accuracy, confusion matrix).
- Small prize for the person who develops the best performing model.
- Individual assignment, accounts for 30% of the final grade.

Assignment III: Research proposal

- Write a research proposal utilizing the text analysis methods from the course.
- Should outline a research question, why it is relevant, a data source and a method/methods to be used.
- Discuss how text data could be used to analyze a specific research question. Your proposal should demonstrate an appropriate choice of techniques for the given problem.
- Free to choose any corpus/data you find interesting.
- Ideally, the question is related to your field of research.
- Can include preliminary analyses and descriptive statistics.
- Max. 10 pages (not counting tables/figures).
- Individual assignment, accounts for 50% of the final grade.

Deadlines

- Submission deadline assignment I and II: 28.11.2022.
- Submission deadline assignment III: 19.12.2022.
- Email: helge.liebert@econ.uzh.ch.
- File transfer option: https://www.switch.ch/services/filesender/