Methods for unstructured data

Assignment

Helge Liebert

University of Basel

Assignment

Assignment

1. Web scraping assignment (20%)
2. Text analysis assignment and report (80%)

- Submission deadline is November 7, 2021.
- Email: helge.liebert@econ.uzh.ch.
- File transfer option: https://www.switch.ch/services/filesender/

Assignment, part I

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').

WHO Venomous Snakes Database

http://apps.who.int/bloodproducts/snakeantivenoms/database/SearchFrm.aspx

- For every country, collect *all* species of venomous snakes.

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

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

Assignment II: Input

1. Download the Kiva data: https://www.kiva.org/build/data-snapshots.
2. Data management: Clean and harmonize the text such that it is suitable for analysis.
   - Variable types, times, ...
   - Text contains many undesired elements: Notes, translation, HTML tags, foreign language elements, ...
   - Focus on loans in English.

Assignment: Analysis

- Choose one categorical outcome (e.g. a specific sector, activity or repayment interval) for classification, and one continuous outcome (e.g. duration until funded, loan amount) for regression.
- Develop at least four different predictive modeling approaches and apply them to both outcomes.
- Evaluate and compare their performance on a hold-out set.
- Approaches can both different models, or the same model with different inputs.
    - Example: Lasso, Naive Bayes, Logistic regression after PCA, random forest using (pre-trained) averaged document vectors, adaptive boosting using (pre-trained) averaged document vectors.

Assignment II: Analysis

- Add other predictor variables from the data (e.g. country, text from loan use statement).
- Feel free to add manual inputs based upon inspection/domain-specific knowledge (e.g. climate data).
- Utilize word vector information in at least one model.
- Feel free to restrict the domain of the model for homogeneity and computational tractability.
  - Example: Limiting the sample to a geographic region (e.g. Southeast Asia), country, time period or a specific sector.
  - Make sure the limitations you impose are sensible.

Assignment II: Report

- Write a report documenting your approach and results (pdf, html or notebook, markdown is encouraged).
- Your analysis should be reproducible with minimum effort and documented with clear instructions how to do so.
- Discuss the data and pre-processing steps you took.
- Explain the methods and your reasons for choosing them.
- Discuss why certain methods perform well and others do not.
- Individual assignment, max. 7,000 words (not counting tables/figures). Not less than half that.

Assignment II: Submission requirements

- Submission:
  1. Report
  2. Code
  3. Data
- Begin with the public data file available on the Kiva website.
- Or submit a smaller (intermediate) analysis data file. Downsample if is still too large to submit, proof-of-concept sufficient.
- Accounts for 80% of the final grade.