Text Analysis in R

Assignment

Helge Liebert

Winter School Université Fribourg

Assignment

Assignment: 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.
- Develop at least five different predictive modeling approaches and apply them to the classification problem.
- 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: Analysis

- Add other predictor variables from the data (e.g. country, funding duration, text from loan use statement) and compare the performance.
- Explore adding inputs from other data based upon inspection/domain-specific knowledge (e.g. climate data).
- Utilize word vector information in at least one model.

Assignment: Scope

- Feel free to restrict the domain of the model for homogeneity.
  - Example: Limiting the sample to a geographic region (e.g. Southeast Asia), country, time period or a specific sector.
- Depending on your computing resources, restrictions may also be required for tractability.
- Make sure the limitations you impose are sensible in the context of your specific prediction problem. Explain why you impose them.

Assignment: Report

- Write a report documenting your approach and results (pdf, html or notebook, markdown or notebook is encouraged).
- 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 3,000.

Assignment: Reproducibility

- Your analysis should be reproducible with minimum effort and documented with clear instructions how to do so.
  - Starting from the public Kiva csv file available online.
  - Do not include hardcoded absolute paths.
  - Create (sub-)directories from within R/Python.
- The code should also be easy to follow, so comment it well (especially if you are not using markdown or notebooks).
- I should be able to just run your submitted code file (possibly after adjusting the input directory).
- Submit code only, not data.

Assignment: Submission

- Submission:
  1. Report
  2. Code
- Submission deadline is March 31, 2022.
- Email: helge.liebert@econ.uzh.ch.
- https://filesender.switch.ch.