Text Analysis in R

Assignment

Helge Liebert

Winter School Université Fribourg 2021

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, 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, or using topic model input, 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, text from loan use statement).
- Feel free to add manual inputs based upon inspection/domain-specific knowledge (e.g. climate data).
- Try to utilize word vector information in at least one model.
- 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.

Assignment: Report

- Write a report documenting your approach and results.
- Briefly 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. 6,000 words/12 pages (not counting tables/figures).

Assignment: Submission requirements

- Submission:
  1. Report
  2. Code
  3. [Data]
- Downsample data if too large to submit. Proof-of-concept sufficient.
- Submission deadline is March 31, 2021.
- Email: helge.liebert@econ.uzh.ch.