Genre classification in Rust

Marko Lozajic, Anna Soboleva

Description

The project is implementing a single label genre classifier (on example of five genres). For feature extraction tf-df is used, information is reduced via SVD and the classification is made by Naive Bayes.

1. Corpus

Our corpus of approximately 500 texts split in train and test set in five genres(’Romance’,’Horror’,’De fiction’ and ’Erotica’). Considering that the amount of data is relatively small, we used a 70/30 distribution. Texts were collected from Project Gutenberg and Archive of our own in .html and .txt formats.

2. Preprocessing

Texts were read, tokenized with option to remove stopwords (as "stopwords" a set of nltk stopwords is used).

3. Classifier

First tf-idf is performed (all features from corpus are extracted), then two separate tf-idf matrices are made. Considering that it might take some time the possibility to save them is provided (and also a saved model is provided too.
Then SVD is performed. It is implemented twice: first using a LA crate, which appeared to be too slow. We ended up not using it in the final version, but left it in the code for future possible more Windows-friendly implementation.
Our final SVD implementation is made using a ndarray crate.
Then Naive Bayes classifier is used and evaluation of scores is made.

4. User experience

Program is mostly controlled via terminal with possible "autopilot" option provided. It works on Mac and Linux and will need a small changes to work on Windows.

5. Possible improvements

- Making a program more Windows friendly;
- Include data augmentation;
- Increase data set and amount of possible genres;
- Try out more classifier options;