Genre Classification

Marko Ložajić
Anna Soboleva

Goal

Implement a single-label genre classifier, using various machine learning methods (KNN, Logistic Regression, Naïve Bayes, etc). Feature selection will be made with tf-idf and word2vec.

The **corpus** is composed of freely available books from Project Gutenberg (which were labeled as “Romance”, “Science Fiction”, “Detective”, “Fantasy” or “Horror”) and texts from archiveofourown.org (labeled as “Erotica” according to the site’s genre tags).

Perceived Difficulties

1) **Copyright.** Modern books mostly aren’t freely available, for example, our corpus for “Romance” is limited with 19 century and beginning of 20 century books (which makes genres different also on vocabulary level). It is difficult to collect large amounts of training/test data.
2) **Ambiguity of certain genres.** Especially considering we are using fanfiction for “Erotica” (there could be included fantasy/sci-fi or romance elements). Sometimes it is difficult to define a genre. (Is single-label method accurate then?)
3) **Find genre-specific words.** Simple inverse document frequency relies on words not appearing in some documents at all - consider less naïve IDF approach?

Third-Party Crates

For data preprocessing

1) **token:** for tokenizing text and splitting sentences;
2) **stopwords:** for cleaning up the input;

For machine learning algorithms

1) **rust-tfidf:** extract term-frequency/inverse document frequency;
2) **word2vec:** extract feature vectors;
3) **rusty-machine:** Logistic Regression, Naïve Bayes Classifier, (SVM);
4) **rustml:** KNN algorithm, (Neural Networks);