Project Description for the ATiML project on Gutenberg Corpus

Team members: Amit Kumar, Mark Trebeljahr, Jesper Dannath, Richhiey Thomas, Siva Matta

Task

Our project has the goal to predict the genres of documents from a dataset of about 1000 books from 19th century english fiction. We want to use machine learning methods to detect with a high accuracy, which of 10 given genres a specific document has.

Method

Software

We will use Python as programming Language to achieve our goal. Precisely we plan to use the Scikit Learn Library for most of our Data preprocessing and model Training.

Preprocessing

The Dataset contains about 440 MB of documents from various length. So, we will have to perform some kind of vectorization to make it accessible for training a classifier. Therefor we will probably use N-Gramms of the text as features. We will also extract some other relevant feature. Afterwards we will select the most promising features for model training using a feature selection method (probably filtering by some test statistic). Then we will also perform a split of the data to a training an evaluation and a test set.

Model

We have not yet decided on a specific machine learning model or training Algorithm. We will try to fit different alternatives like Naive Bayes or logistic regression and determine in a model selection process which one to use in the end.

Evaluation

Our model Performance measure will be Accuracy, so technically speaking our goal will be to find the model that maximizes the accuracy of on the validation dataset and report our final performance on the test set.

Presentation

We will prepare a report in form of a jupyter notebook and split the parts of the project between all team members. We will also prepare charts and graphics for our whole project process such as feature distributions (histograms), learing curves, loss curves, Accuracy of different models etc.