1 NLP Project Preproposal

1.1 Research Problem

The research problem our group wants to investigate is detection of authors of texts. To do so we want to analyze texts from various authors and classify these texts according to their style of writing (vocabulary, linguistic style, grammar, etc.).

1.2 Approach & Evaluation

In order to obtain a data set we are planning to apply a web crawler (prototype from internet) to crawl blogs on the internet. The set is then split into training and test data sets. Before a classifier (e. g. Naive-Bayes, Neural Networks, etc.) is trained on the training data set and evaluated (Precision, Recall, etc.) using the test data set we need to engineer proper features that are useful for classifying the texts. The evaluation can be visualized with a confusion matrix and ROC curve plots.

1.3 Resources & Frameworks

Possible frameworks that are conceivable for solving the task:
DKPro, DKPro TC, Weka, DeepLearning4J, ...

Some references:
News Authorship Identification with Deep Learning
`https://cs224d.stanford.edu/reports/ZhouWang.pdf`
ILSP Focued Crawler
`http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki/Getting_Started`
Crawler4J:
`https://github.com/yasserg/crawler4j/`
Reuters corpus (instead of crawling)

1.4 Organization (Responsibilities & Timeline)

- Data provisioning (crawler, corpus, ...) ⇒ **Philipp Kapelle**
- Data preprocessing (format of data, feature engineering, ...) ⇒ **Fabian Otto**
- Analysis (train and choose classifiers, ...) ⇒ **Clemens Biehl, Daniel Wehner**
- Evaluation ⇒ **Clemens Biehl, Daniel Wehner**

Coarse Timeline:
Calendar Week 2: Provision of crawled data or corpus
Calendar Week 4: First trained machine learning models
Calendar Week 5: Comparison of trained models
Calendar Week 6 (08.02.): Final presentation