NPFL 054 – Homework #2

Outline of the final report

I. The authorship recognition task and the data

- Describe briefly your task, and the structure of the development data.
  - You can use texts from the assignment presentation.
  - You should also use your own words, how you understand your job and what is important in your view.
  - What is your intuition about this kind of machine learning task?

II. Sentence length model

- Analyze the sentence length distributions for different authors, and make a model to predict the author using this kind of information extracted from your training data. Estimate its generalization error.

III. N-gram model

- Describe how you prepared your n-gram feature set(s), and how you selected different subsets for your experiments. That includes the data preprocessing, and frequency thresholds you used for filtering of the n-gram features.
  - Analyze the distribution of the n-gram features in the given collection of books, and in passages.
- Describe how you developed and tuned your SVM model(s)
- Describe your development evaluation, and your estimation of the generalization error.
- Can you recognize which features or feature subsets are most effective?
- Describe your error analysis. What are the most frequent types of errors, and what are their possible reasons?

IV. Final predictor

- Use all development data you have to train your “best” predictor.
  - Describe all details of the model and explain your reasons.
  - Explicitly estimate its generalization error – i.e. your expectation about its performance on the “blind” test data set.

- Be ready to use your final model to predict authors of the “blind” test passages!
  - The format of the test data set will be exactly the same as the format of your development data, but the author attribute will be missing.
  - Describe how to run your predictor. It should take the test data as an input, and produce a vector of predictions. Then the teacher will evaluate your output.

- Make your own conclusions about the task and your experimental work.

General remarks

– Be clear and organize your experimental results into nice tables/graphs.

– Especially, explain clearly

- *which data sets you used in different experiments, and*
- *what feature set(s) and what hyperparameter settings you used*