**Information Retrieval Project 1 Report**

**Naive Bayes**

Preprocessing: Stopword Removal, Porter Stemmer

The implementation of the algorithm follows the lecture slides (Text Categorization, Slides 21-22) very closely including Laplace smoothing. We use a one-vs-all approach for the multilabel-classification, i.e. we train probability vectors for each label. For the classification we use a global threshold for all labels.

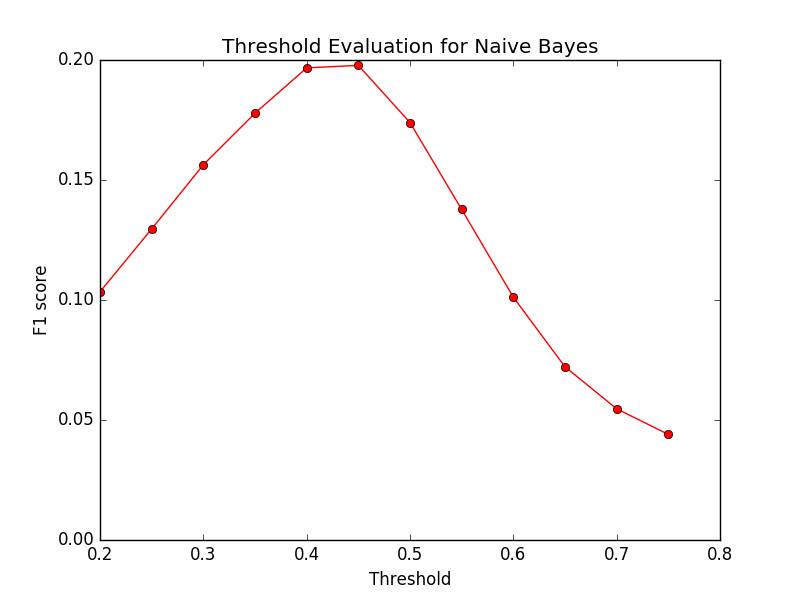
We were not able to run the algorithm on the full training set, due to memory problems.

Fig 1: evaluating best threshold for Naive Bayes

We evaluated the performance of Naive Bayes at different thresholds. Within a training set of 1000 news documents and 500 validation documents, it yields the best results at threshold 0.45.

**Logistic Regression**

Preprocessing: Stopword Removal, Porter Stemmer

Feature Vector: term frequencies

The approach is again similar to what is on the slides (Text Categorization, Slides 26-30). We reuse code from the tinyIR library to update the gradient and perform (sparse) vector operations. We did the stochastic optimization and choose our samples uniformly random.

We again classify in a one-vs-all manner with a global threshold.

*Scores?*

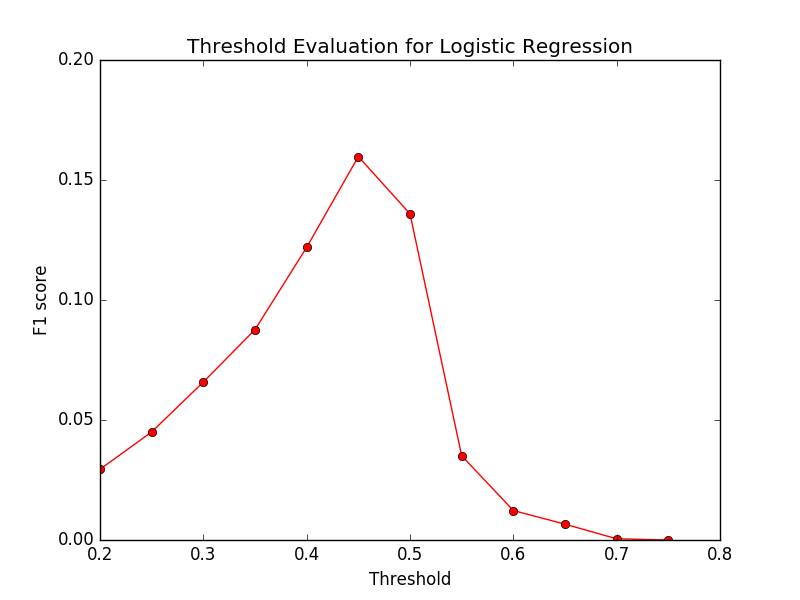
*Runtime information?*

Fig 2: evaluating best threshold for Logistic Regression

We also evaluated the performance of Logistic Regression at different thresholds. In the same set up (1000 training, 500 validation docs), it yields the best results also at threshold 0.45.

**Support Vector Machine**

Preprocessing: Stopword Removal, Porter Stemmer

Feature Vector: term frequencies

For the SVM classifier, we implemented linear classifiers without kernel functions. We find our model vectors with the Pegasos algorithm mentioned in the slides. The classification then is done one-vs-all.

*Scores*

We run our SVM on 1000 training files and evaluated on 1000 validation files. The resulting average F1 score was 0.237.

This took 100 minutes and used about 2GB of memory on an Asus Laptop with an Intel® Core™ i7-4700HQ CPU @ 2.40GHz × 8