Maschinelle Sprachverarbeitung

Assignment 3: SPAM Classification

Ulf Leser

Assignment

- We give you a training set of real spam and real ham
- Implement a classifier that detects (only) spam mails
- Perform 10-fold CV to estimate performance
  - Take care to create folds with the same relative frequency of SPAM / HAM as in the entire corpus (“stratified”)
- We will evaluate on a different set of ham/spam mails
  - With similar class distribution

You are Free in Pretty Everything

- You can chose whatever features you want
  - Only text? Also header? Receiver and sender? Mail-path? MIME-types? Binary, TF, TF*IDF? ...
- You may use whatever feature selection method you like
  - Or none
- You may chose whatever classifier you think works best
  - Naïve Bayes, SVM, kNN, decision-trees, ...
- You may use whatever software you want (but not dedicated tools like SpamAssassin)
  - OpenNLP, NLTK, LingPipe, ...
  - We recommend to have a look at Weka
  - For SVM, we recommend dedicated libraries like libSVM

Submission by Mail to Ulf Leser

- Results due on 11.1.2016
- Submit one JAR file called groupX.jar
  - Two modes
    - java –jar groupX.jar learn spam_dir ham_dir model_name
      - Read all files from both directories and learn a model
      - Model must be written as model_name to the current directory
    - java –jar groupX.jar classify model_name directory_name result_file
      - Read model model_name from current dir
      - Classify all files in directory_name and write result to result_file
      - Write result to “result”-file as list “FILENAME\tSPAM/NOSPAM”
  - Include source code and results of 10-fold CV on training data

Competition

- Best accuracy wins
- Speed is irrelevant