Intelligent Data Analysis

Exam: Spam (Project 5)

This project is part of the exam *Intelligent Data Analysis*. Each project assignment is to be resovled by a single student on his/her own. The student is supposed to present the solution as part of the oral exam. The student is required to present a printed version of the Python code together with diagrams, tables, etc. that summarize the results. The specific way of how the project is presented is up to the student's choice.

Problem setting

You have been hired by the IT department of a medium-sized company to train an email spam filter which should mark the incoming emails of all employees as spam or non-spam. The emails are parsed by a module and converted into the bag-of-words representation. A total of 57,173 different words (features) are distinguished. The aim of the filter is to identify a maximum number of spam emails, with a maximum of 0.2% of all legitimate emails being classified incorrectly. In addition, the company wants to make a statement about the effectiveness of the filter on future emails, i.e., what percentage of incoming spam emails will be identified in the future.

Aufgabe

From the employees' inboxes, 10,000 emails were extracted as training data (see emails.mat). Let X be the training data with the associated class labels Y ($+1$ stands for *spam*, -1 means *non-spam*). Identify a suitable learning technique for constructing a spam filter and implement it in Python.Train and evaluate the model. Make a statement about the expected quality of the filter and make sure that no more than 0.2% of all legitimate emails are filtered. For this purpose, plot a precision/recall curve and mark the model's position on the curve (for the selected threshold). Briefly motivate and document all the steps you have taken.