20-00-0546-iv Foundations of Language Technology

Homework 10
Classify text

21. January 2021

In case your submission consists of several files, compress these to a zip-file. Indicate clearly which submission corresponds to which question. Include comments in your program code to make it easier readable. It is very important that you submit your solution as a Jupyter Notebook file (`.ipnyb`). The deadline for the homework is **Thursday, 28.01.21 23:59 CET**.

10.1 Homework

Homework 10.1 *(10 points) For the sake of science, let's assume that we are on the other side as a spammer.*

(a) What could you do to avoid your spam being detected with respect to the previously implemented features?

- *Formulate at least two counter-strategies. (2 points)*

*Keep in mind that as a spammer all you need is the short attention of the reader for the main message. It does not really matter what the rest of the email looks like. Also the spam message only needs to be readable, not unaltered. Please take care, that you only change the **spam messages** (the messages from the corpus in the training and the test data that belong to the* SPAM *class), not the valid emails, as (hopefully) the spammer has no way of changing the valid emails on the user's computer*

(b) Implement your ideas by adding a method that changes the content of spam emails before they are feed into the classifier for training.

- *write a function to read the mails from the $emails.zip$. You may use the $LazyCorpusLoader$ (1 point)*
- *implement at least two of your ideas from (a) (4 points)*
- *Use the classifier that you implemented in the Exercise 10. (1 point)*
- *Report the precision, recall and f-score. (1 point)*
- *How much can you decrease the classifier's performance? Discuss results (1 point)*