One Pager for Task 1 of IPT NLP Hackathon

Isha Gupta, Moritz Reihs, Luca Entremont

Main source: https://github.com/lucaentremont/ipt-nlp-hackathon-service-classification

In order to classify the emails in a contextually good way, we decided to use a one-dimensional Convolutional Neural Network (using Keras in Python). We tried different architectures and settled for a four-layer version since it yielded the best results for us (we focused more on depth rather than width in order for the model to understand the data rather than memorise it).

Steps taken to Clean Data In order to have clean data (which is natually not possible with emails that sometimes have images attached, different grammars and generally ways of writing), we had to do some serious cleaning. The steps we undertook are as follows: We..

- removed punctuation (such as but not limited to the following : . - ! : ? ...)
- removed capitalisation of words
- removed stop-words (using a publicly available german stop-words-database) and removed words that simply occured too often in those emails (of course keeping the removal-rule in order to apply them to the evaluation data afterwards)
- Removed emails that were written in a different language than German (using the package "langdetect")
- Lemmatised the text using the package "spacy"
- Removed first as well as last names based on a large database we acquired online (together 300'000 names)
- removed html tags as well ass attached images and attachments generally
- **Did not spellcheck** the data, since our computers were too slow to check every word in a short enough amount of time

In order to achieve our best performance (which was 62% in the split test data (20% split))

For Task 1.2 (regarding the manuals), we believe we have quite successful code (70% accuracy) but the time did not suffice anymore to find the rest of the bugs (especially the exporting of the cases where no manual fit proved to be harder than anticipated). This will be explored outside the hackathon!