Sprachverarbeitung: Übung

SoSe 24

Janis Pagel
Department for Digital Humanities, University of Cologne

2024-06-10

For this exercise, you need to both submit manual calculations as well as Python code. Please submit two files in Ilias, one a PDF with your calculations and one a file containing your Python code (either Jupyter Notebook or Python script). You can also combine both files into a zip-archive and submit only the archive. You can either solve the calculations by hand on a sheet of paper, scan it and submit as a PDF file or use the capabilities to write mathematical equations of tools like MS Word / LibreOffice / LaTeX, etc. to write down your calculations digitally.

Exercise 1.

For this part of the exercise, you need to do manual calculations and submit your solution in a PDF file.

Given are the following nine sentences in different languages, Dutch (nl), English (en) and French (fr):

Sentence	Class
een man rookt	nl
er loopt een dier	nl
de man eet pizza	nl
the man snorted	en
never received the product	en
never received my package	en
apprendre de manière ludique	fr
ça ne marche absolument pas	fr
pas de version en français	fr

Table 1: Language identification dataset

Given this training data, create a Naïve Bayes model by hand by calculating all probabilities for a certain feature to occur/not occur in a sentence given a certain class (language). Apply "Add-One" smoothing to all probabilities. Only use the following seven features: "een", "man", "de", "never", "pas", "version" and "en".

Using the calculated probabilites, determine the class of the following three sentences by calculating the probability of a certain class given the presence or absence of the features in the sentences:

Sentence
een man en een vrouw praten
never ordered this version
je n'ai pas les mots

Exercise 2.

For this part of the exercise, you need to write Python code.

The dataset in table 1 is part of a larger dataset from `https://huggingface.co/datasets/papluca/language-identification`. From the course website, download the files `https://lehre.idh.uni-koeln.de/site/assets/files/5151/languageidentification_train.tsv` and `https://lehre.idh.uni-koeln.de/site/assets/files/5151/languageidentificatio` `test.tsv`, which contain a train and test split for this dataset, with each column representing a token feature. The column containing the language classes is called "labels". Train a sklearn Multinominal Naïve Bayes model on the train set and let it predict on the test set. Use sklearn's `classification_report` function to obtain several evaluation metrics for your prediction. Also create a confusion matrix of the test classes and the classes predicted by the model using sklearn's confusion_matrix function. Afterwards, plot the confusion matrix using seaborn's heatmap function.