SAKI SS 2021 Homework 1

Author: Andreas Thomas fo91juzi

Repository: https://github.com/chronark/saki

Summary

Classifying a small datase of banking records into 7 categories using scikit-learn's Multinomial Naive-Bayes Classifier.

The dataset contained many unusable features so I focused on only 3 of them: Buchungstext", "Verwendungszweck" and "Beguenstigter/Zahlungspflichtiger". I have done a little bit of cleaning by removing punctuation and transforming everything to lower case. This did not make a significant difference though. The relevant features were concatenated into one and tokenized using a CountVectorizer.

The other features were either very similar across all records or did not carry any meaning for this classification. I tried using the dates as well thinking that recurring transactions for salary or rent could be detected but it did not make a significant difference because the dataset is rather small.

Evaluation

Using the count vectorizer and a multinomial naive bayes algorithm I achieved ~ 88% accuracy. I used a multinomial classifier because multinomial is well suited and the input variables are discrete numbers, see: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html. I have chosen not to use a tfidf transformer after testing it [Cell 7]. It is generally not suited for this dataset because the type of used features is rather short and most tokens are unique [Cell 5].

As you can see in the confusion matrix [Cell 9] most incorrect classifications are in the categories "leisure" and "standardOfLiving".

Notebook