NLP: Klassifikation

9. September 2022

1 Aufgabe: Klassifikation

- Suchen Sie zwei ähnliche Kategorien aus Wikipedia, mit jeweils etwa 50 bis 100 Artikel
- Speichern Sie alle Artikel aus diesen Kategorien
- Teilen Sie die Daten in Test- und Trainingsdaten auf und trainieren Sie einen Classifier
- Evaluieren Sie den Classifier
- Suchen Sie einige Texte, die falsch klassifiziert wurden und finden Sie die Wörter, die für die falsche Klassifikation verantwortlich waren, also typische Wörter aus der jeweils anderen Klasse.
- Extra: Bei der logistischen Regression bekommt jedes Wort für jede Klasse ein Gewicht. Schauen SIe in der Dokumentation, wie SIe diese Gewichte extrahieren können, und suchen Sie für beide Klassen die 10 wichtigsten Wörter.

2 Lernziele

Am Ende dieser Lerneinheit sollen Sie in der Lage sein:

1. einen Textclassifier mit ScikitLearn zu trainieren und anzuwenden.
2. den Classifier zu eveluieren und zu beurteilen