NLP: Texte und Kodierungen

Christian Wartena

8. April 2021

1 Aufgabe

In dem Ordner *Texte* sind 10 Texte abgespeichert. Diese sind leider unterschiedlich kodiert und manche Buchstaben werden beim Öffnen falsch angezeigt.

Diese Dateien sollen alle in einem einheitlichen Format abgespeichert werden. Außerdem möchten wir von jedem Text wissen, mit welcher Kodierung er ursprünglich gespeichert wurde und es sollte möglich sein, jeden Text korrekt in einem Jupyter Notebook auszugeben.

2 Lernziele

Am Ende dieser Lerneinheit sollen Sie wissen:

1. wie Buchstaben und Texte gespeichert werden können;
2. was ein Codepage ist;
3. welche wichtige Codepages für Deutsche Texte relevant sind;
4. was Unicode und UTF8 sind.

Außerdem sollen Sie in der Lage sein:

1. Texte mit unterschiedlichen Kodierungen in Python einzulesen;
2. Die korrekte Kodierung eines Textes zu ermitteln;
3. Texte in Python in einer gewünschten Kodierung abzuspeichern.

3 Materialien

Um die Aufgabe zu lösen, müssen Sie sich vermutlich zuerst theoretisch mit dem Thema beschäftigen und einige Vorübungen dazu machen. Auf Moodle stehen hierfür folgende Materialien zur Verfügung:

- Ein Foliensatz;
- Ein Screencast zu diesem Foliensatz;
- Eine Übungsblatt;
- Musterlösungen zum Übungsblatt;
- Ein Jupyter-Notebook mit grundlegenden Funktionen zum Lesen und Speichern von Texten.