Hochschule RheinMain
Prof. Dr. Adrian Ulges

31.05.2021

Künstliche Intelligenz (SS 2021)

Aufgabenblatt 1

zu bearbeiten bis: 07.06.2021

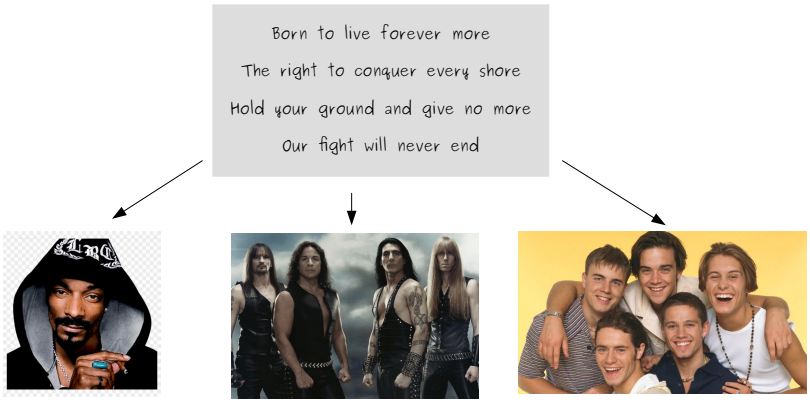
Bearbeiten Sie die Implementierungsaufgaben generell in Zweier-Teams!

Aufgabe 1.1 (Python und Numpy)
Wir werden im Rahmen des Praktikums ausgiebig mit Python und numpy (Pythons mächtiger Bibliothek für Numerik und lineare Algebra) entwickeln. Während die MI-ler zumindest mit Python vertraut sein sollten, haben viele AI-ler, WI-ler und ITS-ler bisher allenfalls kleine Programme geschrieben. Arbeiten Sie sich deshalb mit Hilfe des wunderbaren Python- und Numpy-Tutorial der Stanford University in Python und Numpy ein:

`http://cs231n.github.io/python-numpy-tutorial/`

Es reicht, das Tutorial bis zum Beginn des Themas "Broadcasting" zu bearbeiten.

Aufgabe 1.2 (Songifier)



Wir werden in den ersten Wochen der Veranstaltung einen **Song-Classifier** ("Songifier") implementieren, der Songtexte in Genres (*Hip-hop*, *Rock*, *Pop*, ...) einsortiert – ein extrem schwieriges Problem, mit dem man sich auch als menschlicher Experte schwer tut.
Als ersten Schritt hierzu schauen wir uns die zugehörigen Trainings- und Testdaten an: Sie finden im Read.MI die Dateien `train_small.csv`, `train_medium.csv` und `train_big.csv`, sowie eine Datei `test.csv`. Die Trainingsdateien enthalten 10.000 (train_small.csv) bzw. 50.000 (train_medium.csv) bzw. 290.183 (train_big.csv) Songs, jeweils mit Interpret, Titel, Text, sowie einem Genre. Die kleinen csv-Dateien enthalten Untermengen der großen Datei. Wir werden unseren Klassifikator auf jeweils einem Trainings-CSV trainieren. Eine Grundregel lautet "Iterate

Fast": Weil die Daten in `train_big.csv` recht groß und unhandlich sind, unternehmen Sie Ihre ersten "Gehversuche" mit der kleinen Teilmenge `train_small.csv`. Auf `test.csv` werden Sie später testen wie gut Ihr gelernetes Modell funktioniert.

- Laden Sie `train_small.csv`, z.B. mit dem Python-Modul `csv`. **Wichtig für Sie** sind vor allem die Spalten `Lyrics` (die späteren Features) und `Genre` (die spätere Klasse).
- Ermitteln Sie zunächst, wieviele und welche Genres/Klassen es gibt.
- Ermitteln Sie für jede Klasse jene 10 Terme, die in den meisten Songs der Klasse vorkommen. Geben Sie diese Terme aus. Warum können Sie anhand der Terme nicht die Klasse erraten?
- Ermitteln Sie zuletzt ein **Vokabular** von Termen, welches später die Basis für Ihren Klassifikator bilden wird. Tokenizen Sie Eingabetexte mit `nltk.word_tokenize()`, trennen Sie Kommata und Punkte ab, und lowercasen Sie die Worte. Ermitteln Sie dann (auf train_small.csv) alle Tokens die mindestens 10 mal und in mindestens 5 verschiedenen Dokumenten/Songs vorkommen. Wieviele Tokens enthält das resultierende Vokabular?