NLP SoSe 2020 - Projektbeschreibung

Gruppe: Pascal Schlaak, Tim Weise

Wir möchten uns in diesem Projekt gerne mit Filmdaten der Plattform IMDb auseinandersetzen. Da wir den Arbeitsaufwand unserer einzelnen Ideen noch nicht einschätzen können haben wir folgende Ziele definiert, die wir soweit wie möglich ausarbeiten möchten.

Crawler

Für eine Datenakquise möchten wir einen eigenen Crawler entwickeln. Dieser soll es uns zum einen ermöglichen, Seiten auf IMDb zu identifizieren, die einzelne Film-Artikel enthalten. Zum anderen sollen alle Bewertungen der jeweiligen Filme identifiziert und in JSON überführt werden. Optional könnten diese Dokumenten basierten Daten in einer MongoDB verwaltet werden.

Strukturierung und Analyse

Die zuvor gewonnen Daten sollen im Anschluss für eine Analyse vorbereitet werden. Hierzu sollen Sätze im Text der Bewertung identifiziert und in Wörter aufgeteilt werden können. Es soll versucht werden, Fehler in den Sätzen zu erkennen und diese zu beheben. Mithilfe des POS-Tagging sollen die einzelnen Bestandteile der Sätze klassifiziert werden, um einen Mehrwert aus den Daten, wie z.B. Figuren und Schauplätze, zu generieren. Eine Analyse der Daten soll einen Überblick zu relevanten Themen (Zipf & Luhn) eines Films ermöglichen. Grundformen sollen ebenfalls erfasst werden (Stemming).

Transformation

Generell sollen verschiedenste Daten zu einem bestimmten Film, basierend auf dessen Bewertungen, in einem webbasierten User Interface dargestellt werden.

Primär sollen Informationen zu Figuren der Filme, z.B. Charakteristika und Beziehungen zu anderen Figuren, generiert werden. Da wir uns noch in die verschiedenen Möglichkeiten der Integration von ML einlesen müssen, haben wir noch keine konkrete Idee, welchen Algorithmus wir verwenden möchten und was wir damit an Mehrwert generieren möchten. Beispielsweise haben wir folgende Ideen, wobei wir die Umsetzbarkeit aktuell noch nicht genau einschätzen können.

Entweder können beispielsweise mithilfe eines Clustering-Algorithmus, inhaltlich ähnliche Filme identifiziert und basierend auf deren Übereinstimmung dargestellt werden.

Oder kann beispielsweise eine Generierung von neuen Kurzgeschichten der Handlung des ausgewählten Films angestrebt werden. Als Trainingsdaten sollen jedoch die zuvor gewonnenen handlungsrelevanten Daten der Filmbewertungen genutzt werden. Hierzu soll ein RNN mit Tensorflows High-Level API Keras entwickelt werden.