0_nlp_intro_schlaak_weise

August 23, 2020

1 Projekt - Natural Language Processing (SoSe 20)

Dieses Projekt befasst sich anhand eines realitätsnahen Anwendungsfalls mit der Analyse natürlicher Sprache. Hierzu wurden verschiedene Anforderungen an Themen definiert, welche die Vorlesungsinhalte des Fachs "Natural Language Processing" wiederspiegeln.

Alle hier hervorgehobenen Hyperlinks bieten eine Navigation zwischen den einzelnen Notebooks. Am Ende jedes Notebooks findet sich jeweils ein Link, welcher eine Navigation zum nächsten Notebook, sowie zurück zur Übersicht, ermöglicht.

1.0.1 Authoren

- Pascal Schlaak (58738)
- Tim Weise (58716)

Bei weiteren Fragen können Sie uns unter den hier referenzierten E-Mailadressen kontaktieren.

1.0.2 Setup

Zu einem erfolgreichen Ausführen dieser Notebooks wird die Nutzung von `Python 3.8.2` empfohlen. Ebenso sollte eine virtuelle Umgebung erstellt werden, in welcher die in der `requirements.txt` aufgelisteten Module installiert werden sollten. Im folgenden Codeausschnitt wird dargestellt, wie ein solches Setup realisiert werden kann.

```
cd /dieses/projekt/verzeichnis
python3 -m venv "venv-name"
source ".venv-name"/bin/activate
pip install -r ./requirements.txt
```

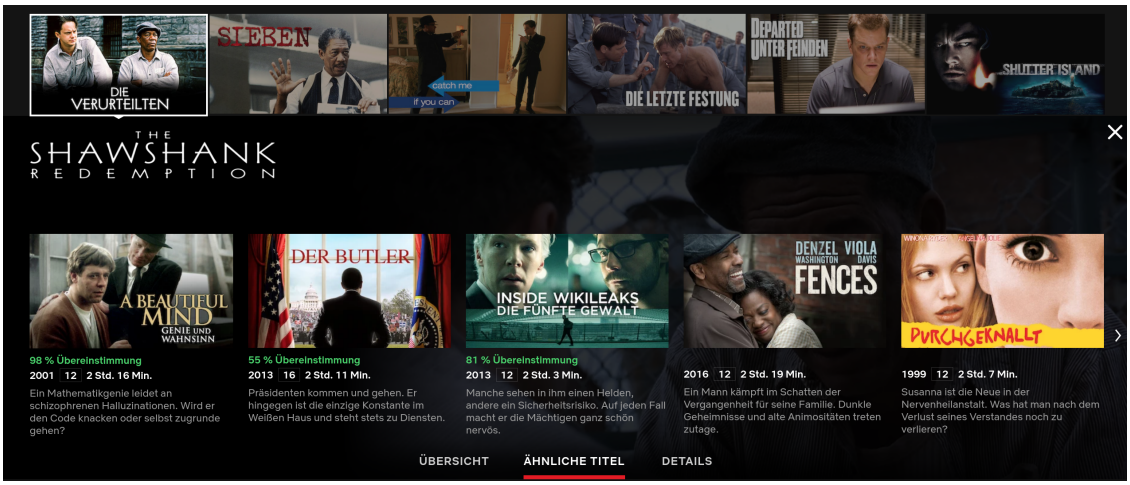
Ressourcen

- Python 3.8.2
- Module siehe requirements.txt

Achtung: Die Laufzeitdauer mancher Codezellen ist sehr hoch. Sämtliche Outputs wurden deshalb bereits von uns generiert. Wir empfehlen daher lediglich das Durchlesen der Notebooks. Insbesondere bei der Anwendung des Clustering-Algorithmus kann es zu Abweichungen oder einer anderen Reihenfolge der Cluster kommen.

1.1 Problemstellung

Die Projektidee ist aus den Empfehlungslisten für ähnliche Titel bei der Filmwiedergabe auf Streaming-Plattformen bekannt: Firmen wie beispielsweise Netflix nutzen bereits solche Funktionalität in ihrem System (siehe Abbildung), wobei verschiedenste Ansätze für die Generierung dieser Vorschläge existieren.



Welche Datenbasis für die Vorschläge verwendet wird ist unklar. Es kommen mehrere Quellen in Frage:

- Bewertungen des Nutzers
- Streaming-Verlauf des Nutzers
- Filme des gleichen Genres
- Inhaltlich ähnliche Filme
- Filmneuheiten
- etc.

Dieses Projekt beschränkt sich auf Handlungsinhalte aus Zusammenfassungen der Filmbeschreibung in natürlicher Sprache und in Textform, um die Empfehlung ähnlicher Filme zu generieren und zu untersuchen, wie gut sich diese Datenbasis für ein Clustering eignet.

1.2 Gliederung

Mithilfe eines Web Crawlers wird eine Datenbasis verschiedener Filme generiert. Im Folgenden erfolgt ein erstes Sichten der Daten, um die Korrektheit der Daten sicherstellen zu können. Anschließend werden alle Daten dahingehend vorverarbeitet, dass eine spätere Analyse möglich ist. Zuletzt soll eine Analyse der Filmdaten durchgeführt werden. Hierbei werden alle Einträge mithilfe eines k-Means-Clustering-Algorithmus einem Genre zugeordnet. Ebenfalls wird eine Ähnlichkeitsanalyse durchgeführt, womit eine Empfehlung ähnlicher Titel ermöglicht werden soll.

Wir empfehlen das Durcharbeiten der folgenden Notebooks in chronologischer Reihenfolge.

1. Datencrawling
2. Datenvalidierung
3. Verarbeitung aller Dateneinträge
4. Clusteranalyse und Vorhersage Empfehlung

[ ]: