READ ME

Zielsetzung:

Das Ziel des Data Exploration Projektes war es ein Recommender System zu erstellen. Diese dienen dazu, einem User anhand dessen Verhaltens einen Vorschlag für zum Beispiel Filme oder ähnlichem zu erstellen. Dies geschieht, indem man das Verhalten mit dem von anderen Usern vergleicht und deren nachfolgenden Entscheidungen betrachtet.

Hierzu fand das Projekt-Team einen Kaggeldatensatz zu dem Thema Filmen und baute anhand diesem ein eigenes Recommender System.

Gruppenteilnehmer:

Marc Franke – 9408418

Niklas Luczak - 5523187

Quellcodeausführung:

Alle initial genutzten Datensets sind unter „https://www.kaggle.com/rounakbanik/the-movies-dataset?select=ratings.csv“ zu finden. Aus dieser Datensetsammlung wurden jedoch nur die Datensätze ratings.csv und movies_metadata.csv genutzt. Darüber hinaus sind die drei Pythonpackages Pandas, SK-Learn und random notwendig zum Ausführen des Skriptes.

Der Code des Data Exploration Projektes besteht nun eigentlich aus zwei Teilen beziehungsweise Files. Der erste ist hierbei die Datenaufbereitung (data_preperation.py). Dieser importiert die zwei Kaggle Datensätze aus dem Filmdatenset ratings.csv und movies_metadata.csv. Diese bearbeitet er dann wie in der Ausarbeitung und Präsentation beschrieben und kreiert eine neue csv-Datei.

Diese wird nun im zweiten Teilen in der Algorithmus-Datei (algorithm.py) zunächst einmal importiert. Anhand dieses entsteht nun das beschriebene Clustering und die Clusterzuordnung durch die Supervised Learning Algorithmen. Schlussendlich sind in diesem dann auch die Ergebnisse des Projektes in Form von den Algorithmen-Modellen festgehalten. Darüber hinaus ist in diesem auch ein Beispiel für eine Recommandation-Erstellung dokumentiert.

Die Aufteilung wurde aus Performancegründen durchgeführt. So musste für die Erstellung und das Optimieren der Algorithmen nicht jedes Mal die gesamte Datenaufbereitung ablaufen, was einen besseren und schnelleren Workflow zuließ.

Zur Vereinfachung der Abgabe des Projektes wurden nun diese zwei Pythondatein einfach zu einer zusammengeschlossen (complete_project.py). In diesem sind nun die entsprechenden Speicher- und Ladebefehle auskommentiert und die Namen in dem Code entsprechend angepasst.

Zum Ausführen der Dateien müssen jedoch jeweils die entsprechenden Datensets heruntergeladen werden und die initialen Importbefehle dieser angepasst werden.

Sollten die Files ebenfalls einzeln nacheinander ausgeführt werden, muss auch noch der Exportbefehl des Data-Preperation-Files und der Importbefehl des Algorithm-Files angepasst werden. Hierzu stehen jeweils zwei Befehle zur Verfügung. Einer dieser ist bereits auskommentiert da er die csv ratings_wide.csv erstellt beziehungsweise lädt. Dies ist der Datensatz, welcher alle 260.000 Zeilen enthält. Dieser konnte jedoch im zweiten Teil von dem Projektteam aufgrund Performanceprobleme nicht im vollen Ausmaß genutzt werden. Aus diesem Grund hat man eine verkürzte Variante mit lediglich 50.000 Zeilen erstellt. Diese wird als short.csv exportiert und importiert.