

Q (Spider): Beim Extrahieren der Description (bei imdb "storyline") fiel mir auf, dass am Ende des Textes noch ein Verfasser genannt wird à | "Written by ...". Sollen wir dieses Autoren-Tag extra ausschneiden oder können wir ihn drin lassen?

A: Das kann drin bleiben, siehe Beispiel auf der Website (https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws1617/ue_inforet/example_movie_avatar.json).

Q (Spider): Uns ist aufgefallen, dass in der movies.json Filme enthalten sind, die noch nicht in den Kinos sind und für die gewisse Eigenschaften (Filmlänge und Rating) noch nicht zur Verfügung stehen. Sollen wir diese Felder leer lassen (leerer String) oder gibt es default-Werte für diese Felder zwecks Sortierung?

A: leere Strings bzw. Arrays reichen. Die Klasse Movie.java kann mit leeren Strings umgehen. Es gibt jetzt auch den Testkorpus auf der Webseite. Daran könnt ihr euch auch orientieren.

Q (Queries): Soll die Ausgabe bei den Methoden jeweils eine ArrayList sein oder ist es unwichtig. Wenn ja, Titel oder das gesamte Movie-Objekt?

A: Bitte die Klasse IMDBQueries.java auf der Übungsseite nutzen. Da sind die Signaturen für jede Anfrage genau angegeben:

Q (Queries): Gibt es ein Zeitlimit für die Queries?

A: Da wir aber insgesamt ca. 25 Abgaben zu berücksichtigen haben, werden wir Queries, die länger als 10 Minuten brauchen (und damit im Wettbewerb auch keine Chance hätten), abbuchen müssen. Wir werden aber trotzdem prüfen, ob solche lang laufenden Queries korrekt implementiert sind.

Q: (Query 2): Hier heißt "US-American" Movies. Schließt das Co-Produktionen mit anderen Ländern ein? Also ein Film, der z.B. sowohl USA, IT, DE als Produktionsland hat? Oder ist hier USA only gemeint?

A: Das ist in der Tat nicht ganz eindeutig formuliert. Koproduktionen mit anderen Ländern sind jedenfalls mit eingeschlossen. Gemeint sind also alle Filme, bei denen USA in der Country-Liste auftaucht.

Q (Query 2): Also, der Verlust ist definiert als Budget - Gewinn. Das sollte doch umgekehrt sein, oder?

A: Das stimmt. Es ist aber am Ende nur eine Vorzeichenfrage. Hauptsache es wird am Ende so sortiert, dass die Filme mit größerem Verlust im Ranking weiter vorne stehen.

Q (Query 3): Wie soll die Häufigkeit der Vorkommen von "kill" und "love" nun gezählt werden?

a) Xx "love", Yx "kill" -> Vorkommnisse: $X+Y$

b) Xx "love", Yx "kill" -> Vorkommnisse: $\min(X,Y)$

A: gemeint ist $X+Y$.

Q (Query 3): Dazu zählen auch substrings wie z.B. bei "das fünfte Element": ...skilled... enthält ja kill und wäre ja somit schon ein Substring, oder?

A: Das stimmt. "kill" ist ein Substring von "skilled". Wir wollten an dieser Stelle keine Tokenization von euch verlangen. Also ist es in Ordnung, wenn die Query hier so implementiert wird, dass beim Wort "skilled" der Substring "kill" erkannt wird.

Q (Query 5): Spielt die Währung bei Collosal Failure eine Rolle? Also \$ oder € oder Yen oder ...?

A: Bei Query 5 geht es ja um US-amerikanische Filme (also alle Filme mit "USA" in der Country-Liste). Habt ihr denn den Fall, dass ein US-amerikanischer Film eine andere Währung als den Dollar hat? Für den Fall, dass euch solche Filme vorliegen, dürft ihr die Währung getrost ignorieren. Es gilt dann also: 1 Yen entspricht 1 Dollar entspricht 1 Euro.

Q (Query 8): Sollen nun der beste Filme des Jahres ausgegeben werden (also auch 10 Stk.), oder die besten Filme des Jahres mit einem Rating über 10k?

A: Bei der Beschreibung der "must see"-Query heißt es "Must See: List the best rated movie of each year starting from 1990 until (including) 2010 with at least 10,000 ratings. ..." (Singular). Es geht also nur um den einzelnen am besten bewerteten Film jedes Jahres (mit mindestens 10.000 Bewertungen) zwischen 1990 und 2010. Es sind also insgesamt 21 Filme - den besten (mit mindestens 10.000 Bewertungen) für 1990, 1991, ..., 2010.

Q (Query 10): Ihr habt bei einigen Queries die Ausgabe auf 10 beschränkt, laut Text auch bei den MagicCouples, jedoch fehlt die entsprechende Bedingung bei der Ausgabe.

A: Stimmt. Da haben wir in der main-Methode tatsächlich vergessen, zu prüfen, dass wirklich nur die Top-10 zurückgegeben werden. Hier gilt dann der Text aus der Aufgabenstellung, d.h. in unserer Referenzimplementierung werden nur die Top-10 zurückgegeben. Ansonsten wird die Ausgabe wohl recht umfangreich.