Abfragen auf IMDB Datenbank: Musterlösungen

IMDB ([www.imdb.com](http://www.imdb.com)) stellt einen grossen Teil des Datenbestandes zur Verfügung in einem Textformat ([www.imdb.com/interfaces](http://www.imdb.com/interfaces)). Mit JMDB ([www.jmdb.de/](http://www.jmdb.de/)) ist ein Programm verfügbar, dass dieses Textformat liest und die Daten in eine MYSQL-Datenbank füllt. Und diese Datenbank verwenden wir für diese Übung.

Eine Grafik von jmdb.de zeigt auch das rasant steigende Datenvolumen von IMDB:



IMDB hat etwas über 30 Tabellen. Wir betrachten für diese Übung die folgenden Tabellen



Die Daten, die wir verwenden, sind auf dem Stand von November 2010: Ca. 1.7 Millionen Filme.

## Finden Sie den Film “The King’s Speech”, den Oscar Gewinner 2011 in der Kategorie Best Film: Welche ID hat der Film?

Es ist die ID 1580365.

### Lösung

In realen Daten sind die Daten nicht immer so sauber erfasst, wie man sich das wünschen würde. Eine exakte Suche nach 'The King’s Speech' funktioniert nicht, weil im Titel auch (2010) enthalten ist. Eine breite Suche nach '%King%Speech%' funktioniert nicht, weil dann auch der Film "Larry King Live" (1985) {Reaction to Ombama's Speech}" gefunden wird. Eine weitere Hürde ist, dass ' Teil des Filmtitel ist und wie bei regulären Ausdrücken als Sonderzeichen markiert werden muss mit \'.

SELECT \*

FROM movies

WHERE title LIKE '%The King\'s Speech%'

## Wer war der Director von “The King’s Speech?”

Verwenden Sie aus Effizienzgründen die movieid 1580365. So sparen Sie ein JOIN; die Abfrage wird etwa 100x schneller.

### Lösung schnell (mit nur einem JOIN) [0.04s]

SELECT \*

FROM directors

JOIN movies2directors USING(directorid)

WHERE movieid=1580365

### Lösung langsam (mit zweifachem JOIN) [14s]

SELECT \*

FROM directors

JOIN movies2directors USING(directorid)

JOIN movies USING(movieid)

WHERE title LIKE '%The King\'s Speech%'

## Wer waren die Schauspieler in “The King’s Speech”, und wie hiessen ihre Rollen (nach Namen aufsteigend sortiert)?

35 Schauspieler sind aufgeführt:

**name as\_character**

Ambridge, Dean (uncredited) [Royal Marine]

Andrews, Anthony (I) [Stanley Baldwin]

Applewhite, Dominic [Valentine Logue]

Barrows, Mark (II) [Cosmo Lang's Secretary]

Best, Eve [Wallis Simpson]

Bonham Carter, Helena [Queen Elizabeth]

…

### Lösung schnell (mit nur einem JOIN) [0.1s]

SELECT name, as\_character

FROM actors

JOIN movies2actors USING(actorid)

WHERE movieid=1580365

ORDER BY name ASC

### Lösung langsam (mit zweifachem JOIN) [3.5min]

SELECT name, as\_character

FROM actors

JOIN movies2actors USING(actorid)

JOIN movies USING(movieid)

WHERE title LIKE '%The King\'s Speech%'

ORDER BY name ASC

## Aus welchem Land sind wieviele Filme erfasst (absteigend sortiert nach Anzahl)? Geben Sie die Anzahl Filme und den prozentualen Anteil aus.

Die Länder mit den meisten Filmen sind:

**country anzahl prozent**

USA 358146 40.0141

UK 76510 8.5481

France 46870 5.2366

Germany 35438 3.9593

Japan 30737 3.4341

Canada 30640 3.4233

...

### Lösung

SELECT country, COUNT(country) AS anzahl, COUNT(country) / (SELECT COUNT(\*) FROM countries) \* 100 AS prozent

FROM countries

GROUP BY country

ORDER BY COUNT(country) DESC, country ASC

## Wie viele Filme sind aus der Schweiz, und wieviel Prozent aller Filme sind das?

Sie können die Zeilen mit WHERE oder die Gruppen mit HAVING filtern. Macht das für diese Abfrage einen Unterschied?

Die Schweiz hat immerhin einen Anteil von 4%:

**Country anzahl prozent**

Switzerland 3820 0.4268

### Lösung mit HAVING

SELECT country, COUNT(country) AS anzahl, COUNT(country) / (SELECT COUNT(\*) FROM countries) \* 100 AS prozent

FROM countries

GROUP BY country

HAVING country='Switzerland'

### Lösung mit WHERE

SELECT country, COUNT(country) AS anzahl, COUNT(country) / (SELECT COUNT(\*) FROM countries) \* 100 AS prozent

FROM countries

WHERE country='Switzerland'

GROUP BY country

## Zu welchem Genre gibt es in welchem Land wie viele Filme (absteigend nach Anzahl sortiert)?

**Country genre anzahl**

USA Short 116994

USA Comedy 69133

USA Drama 68896

USA Documentary 41742

USA Adult 38972

USA Animation 17180

France Short 16678

USA Action 15035

UK Short 14767

### Lösung

SELECT country, genre, COUNT(\*) AS anzahl

FROM countries

JOIN genres USING(movieid)

GROUP BY country, genre

ORDER BY COUNT(\*) DESC, country ASC, genre ASC

## Welches Genre ist in welchem Land prozentual wie häufig?

Absteigend sortiert nach Prozentsatz der Genres pro Land. Achtung: Länder mit nur einem Genre sollten ausgeschlossen werden, weil diese Länder die Auswertung sonst verfälschen.

**country genre prozent**

Gibraltar Adult 0.9565

Central African Republic Documentary 0.8333

Liberia Documentary 0.8333

Burundi Short 0.7778

U.S. Virgin Islands Short 0.7714

Democratic Republic of the Congo Documentary 0.7500

Montenegro Drama 0.7500

Nigeria Drama 0.7358

Uganda Documentary 0.7143

Aruba Action 0.6923

Sudan Documentary 0.6923

Bermuda Biography 0.6667

Bermuda Documentary 0.6667

### Lösung - Achtung, langsame Abfrage [164s]

SELECT country, genre, COUNT(\*) / (SELECT COUNT(\*) FROM countries c1 WHERE c1.country = c2.country) AS prozent

FROM countries c2

JOIN genres USING(movieid)

GROUP BY country, genre

HAVING prozent < 1.0

ORDER BY prozent DESC, country ASC, genre ASC

## Welches sind die TOP 10 vergebenen MPA Ratings?

322 Rated R for language.

228 Rated R for violence and language.

144 Rated R for violence.

83 Rated R for language and some violence.

79 Rated R for some violence.

69 Rated R for some language.

66 Rated R for language and some sexual content.

61 Rated R for strong sexual content.

60 Rated R for violence and some language.

60 Rated R for sexual content and language.

### Lösung

SELECT COUNT(reasontext), reasontext

FROM mpaaratings

GROUP BY reasontext

ORDER BY COUNT(reasontext) DESC

LIMIT 10

## Welche Ratings sind häufiger: Ratings, deren Reasontext “violence” enthält, oder Ratings, deren Reasontext “sexual” enthält?

Können Sie diese Frage mit nur einer SQL Anfrage beantworten?

### Lösung

SELECT COUNT(reasontext)

FROM mpaaratings

WHERE reasontext LIKE '%violence%'

=> 5951 Treffer

SELECT COUNT(reasontext)

FROM mpaaratings

WHERE reasontext LIKE '%sexual%'

=> 5673 Treffer

Das lässt sich zu einer - wenn auch nicht intuitiven - SQL-Anweisung zusammenfassen:

SELECT

 (SELECT COUNT(reasontext) FROM mpaaratings WHERE reasontext LIKE '%violence%')

 >

 (SELECT COUNT(reasontext) FROM mpaaratings WHERE reasontext LIKE '%sexual%')

Die Antwort lautet dann “1” für “true”, das heisst, violence ist häufiger als sexual.

## Bei welchen Filmen war Mel Gibson Director?

Titel und Jahr sollen angezeigt werden, nach Jahr absteigend sortiert.

**title year**

Untitled Mel Gibson Project (2011) 2011

Apocalypto (2006) 2006

"Complete Savages" (2004) {Tutoring (#1.2)} 2004

"Complete Savages" (2004) {The Man Without a Ball (#1.11)} 2004

"Complete Savages" (2004) {Pilot (#1.1)} 2004

The Passion of the Christ (2004) 2004

Braveheart (1995) 1995

The Man Without a Face (1993) 1993

### Lösung

SELECT title, year

FROM movies2directors

JOIN directors USING (directorid)

JOIN movies USING (movieid)

WHERE name like '%Gibson, Mel%'

ORDER by year DESC;

Die Frage, in welchen Filmen Mel Gibson gleichzeitig mitgespielt hat, sprengt Datenbank-Kapazität, die Abfrage läuft und läuft...

## Welches sind die hundert am besten bewerteten Filme mit den meisten Stimmen?

**name rank votes**

"Ugly Betty" (2006) {Hello Goodbye (#4.20)} 9.9 91

"Kocka, kocka, kockica" (1991) 9.9 75

"Kamiondzije 2" (1983) 9.9 74

"The Dick Cavett Show" (1968) {Bette Davis} 9.9 63

Big Ass Fixation 4 (2009) (V) 9.9 60

Keine allzu bekannten Filme:-) Wenig Ratings, aber gute Ratings - loyale Fans:-)

### Lösung - Achtung, langsame Abfrage [Schule: 45s]

SELECT title, rank, votes

FROM movies

JOIN ratings USING (movieid)

ORDER BY rank DESC, votes DESC

LIMIT 100;

## Welches sind die hundert am besten bewerteten Filme mit den meisten Stimmen, die mindestens 100’000 Stimmen erhielten?

So kommen doch die eher bekannten Filme zum Vorschein:

**name rank votes**

The Shawshank Redemption (1994) 9.2 520993

The Godfather (1972) 9.2 412075

Inception (2010) 9.1 192231

The Godfather: Part II (1974) 9.0 246003

Il buono, il brutto, il cattivo. (1966) 9.0 160898

The Dark Knight (2008) 8.9 466285

Pulp Fiction (1994) 8.9 418792

Schindler's List (1993) 8.9 276983

One Flew Over the Cuckoo's Nest (1975) 8.9 215130

12 Angry Men (1957) 8.9 119564

The Lord of the Rings: The Fellowship of the Ring (2001) 8.8 388321

### Lösung

SELECT title, rank, votes

FROM movies

JOIN ratings USING (movieid)

WHERE votes > 100000

ORDER BY rank DESC, votes DESC

LIMIT 100;