




Step 2: Integration urban-octo-spoon

Isabella Lisica
Eyad Hamed
Nureldien Gebril

Movies, Ratings & Revenues: Link Datasets

- Es werden die Datensätze zweier Datenbanken verbunden, um Zusammen die Information über die Bewertungen und Einnahmen eines Filmes zu erhalten.
- Im weiteren verlauf könnte dies zum Beispiel interessant sein, für eine Analyse, ob die Einnahme im Zusammenhang mit den Bewertungen stehen
- Problem: 

id	date	title	revenue	theaters	distributor
8b19ad43-3a7e-b14b-49e9-1f7a0eb1568e	2004-09-20	Sky Captain and the World of Tomorrow	925482	3170.0	Paramount Pictures
481fc700-fcdd-1919-c53c-09fcd423a596	2004-09-20	Resident Evil: Apocalypse	643680	3284.0	Screen Gems



Filme können ähnliche Namen haben
und neue Verfilmungen sein

movielid	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy



userId	movielid	rating	timestamp
1	296	5.0	1147880044
1	306	3.5	1147868817

Data Integration

Isabella Lisica

Eyad Hamed

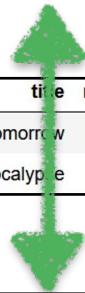
Nureldien Gebril

Movies, Ratings & Revenues: Lösungsansatz

- Ähnlichkeitsmaß für Titel der 1.DB und der 2. DB
- Für die 1. DB verbinden wir title und date und vergleichen mit title aus 2. DB (enthält Datum)

Toy Story 1995-11-22 = Toy Story (1995) ?

id	date	title	revenue	theaters	distributor
8b19ad43-3a7e-b14b-49e9-1f7a0eb1568e	2004-09-20	Sky Captain and the World of Tomorrow	925482	3170.0	Paramount Pictures
481fc700-fcdd-1919-c53c-09fd423a596	2004-09-20	Resident Evil: Apocalypse	643680	3284.0	Screen Gems



Filme können ähnliche Namen haben
und neue Verfilmungen sein

Bsp.:

```
In [14]: from difflib import SequenceMatcher  
  
def similar(a, b):  
    return SequenceMatcher(None, a, b).ratio()  
  
similar("Toy Story (1995)", "Toy Story 1997-11-22")
```

Out[14]: 0.7222222222222222

movielid	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy

userId	movielid	rating	timestamp
1	296	5.0	1147880044
1	306	3.5	1147868817

Data Integration

Isabella Lisica

Eyad Hamed

Nureldien Gebril