

Zum Aufwärmen

✎ 0.1 Erstellen Sie das Jupyter-Notebook (JN) *00-WarmingUp* und laden Sie das referenzierte Movie-Dataset¹. Über welche Dimensionalität verfügt das Movie-Dataset bzw. das erstellte *DataFrame*?

✎ 0.2 Generieren Sie nachfolgende JN-Ausgabe. Anm.: Die vielen Fragezeichen können ignoriert werden.

	color	director_name	num_critic_for_reviews	duration	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_title
0	Color	James Cameron	723.0	178.0	936.0	7.9	1.78	
1	Color	Gore Verbinski	302.0	169.0	5000.0	7.1	2.35	
2	Color	Sam Mendes	602.0	148.0	393.0	6.8	2.35	
3	Color	Christopher Nolan	813.0	164.0	23000.0	8.5	2.35	
4	NaN	Doug Walker	NaN	NaN	12.0	7.1	NaN	

✎ 0.3 Selektiert man eine Spalte (z.B. *movie_title*) des erstellten *Movie-DataFrames*, ist das Ergebnis vom Typ *Series*. Treten Sie den Beweis an!

Mögliche JN-Ausgabe:

```
<class 'pandas.core.series.Series'>
```

✎ 0.4 Sicherlich wissen Sie noch, dass man *DataFrames* manipulieren kann. Fügen Sie die Spalte „has_seen“ mit dem Initialwert „0“ hinzu.

✎ 0.5 Wir sind a) an der Gesamtanzahl der *Director*- und *Actor_{1,2,3}*-Likes interessiert. Um nicht jedes Mal die Summe über alle bilden zu müssen, empfiehlt sich mit „director_actor_facebook_likes“ eine weitere Spalte, die die Summe der definierten Likes je Film aufweist. Bilden Sie diese.

¹ movie.csv

Mögliches Ergebnis:

budget	title_year	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes	has_seen	actor_director_facebook_likes
237000000.0	2009.0	936.0	7.9	1.78	33000	0	2791.0
300000000.0	2007.0	5000.0	7.1	2.35	0	0	46563.0
245000000.0	2015.0	393.0	6.8	2.35	85000	0	11554.0
250000000.0	2012.0	23000.0	8.5	2.35	164000	0	95000.0
NaN	NaN	12.0	7.1	NaN	0	0	NaN
263700000.0	2012.0	632.0	6.6	2.35	24000	0	2277.0
258000000.0	2007.0	11000.0	6.2	2.35	0	0	39000.0
260000000.0	2010.0	553.0	7.8	1.85	29000	0	1651.0
250000000.0	2015.0	21000.0	7.5	2.35	118000	0	66000.0

b) Wie man der exemplarischen Ergebnisdarstellung entnehmen kann, weist zumindest ein Eintrag den Wert „NaN“ auf. Was ist die Ursache?

Finden Sie c) heraus, um wie viele Einträge es tatsächlich handelt und eliminieren Sie d) das Problem, indem Sie die betreffenden Felder des Datasets auf „0“ setzen.

✎ 0.6 Es stellt sich die Frage, ob ein Film, dessen Produktion (vgl. Budget) 22 Millionen beträgt, kostspielig ist oder nicht. Wie könnte man das herausfinden?

✎ 0.7

Grundlage bildet das SMS-Spam-Dataset². Dieses weist 5574 Nachrichten auf, die entweder als Spam oder Ham (=kein Spam) Nachricht klassifiziert sind. Laden Sie a) den Datensatz (Plain text). Fügen Sie b) eine weitere Spalte hinzu, die die Gesamtlänge der jeweiligen Nachricht aufweist (siehe nachfolgende Abbildung).

	label	message	length
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	ham	U dun say so early hor... U c already then say...	49
4	ham	Nah I don't think he goes to usf, he lives aro...	61

² <http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>