PREDICTING THE NEXT

BEST PICTURE

Nicholas Amoscato ◆ Julie De Lorenzo ◆ Chun Ping Ng

1. Export MovieLens Database

from movielens.umn.edu

```
MovieId  Rating   Average  ImdbId   Title
318      3.5      4.5      0111161  Shawshank Redemption, The (1994)
7502     3.0      4.3      0185906  Band of Brothers (HBO) (2001)
50       3.0      4.3      0114814  Usual Suspects, The (1995)
```

2. Extract List of IMDb IDs

and initialize a JavaScript array

```
tt0111161
tt0185906
tt0114814
tt0108052
```

```
var movie_ids = new Array();
movie_ids[0] = "tt0111161";
movie_ids[1] = "tt0185906";
movie_ids[2] = "tt0114814";
```

3. Get Raw Feature Values

from imdbapi.org


```
{"runtime": ["142 min"], "rating": 9.3, "genres": ["Crime", "Drama"], "rated":
"R", "language": ["English"], "business": {"opening_weekend": [{"money":
"$727,327", "year": 1994, "remarks": ["33 Screens"], "country": "USA", "day": 25,
"month": 9}], "gross": [{"country": "USA", "year": 2012, "money": "$28,341,469",
"day": 5, "month": 8}, {"country": "UK", "year": 1995, "money": "\u00a32,344,349",
"day": 18, "month": 5}, {"country": "UK", "year": 1995, "money":
"\u00a31,732,123", "day": 16, "month": 4}, {"money": "$58,500,000", "country":
"Worldwide"}, {"money": "$555,480", "country": "Belgium"}, {"money": "ESP
637,291,985", "country": "Spain"}], "filming_dates": [[{"month": 6, "day": 16,
"year": 1993}, {"month": 9, "day": 10, "year": 1993}]], "budget": [{"remarks":
["estimated"], "money": "$25,000,000"}], "admissions": [{"money": "82,890",
"country": "Belgium"}, {"country": "France", "year": 1995, "money": "163,594",
"day": 28, "month": 3}, {"country": "Germany", "year": 1995, "money": "410,811",
"day": 31, "month": 12}, {"money": "1,245,604", "country": "Spain"}],
"copyright holder": [["copyright 1994 Castle Rock Entertainment"]]}, "title": "The
```


4. Format JSON Response

with features separated by tab; multiple feature values separated by comma

```
0        tt0111161        9.3      864423   Tim Robbins,Morgan
Freeman,Bob Gunton,William Sadler,Clancy Brown      Frank Darabont
         Stephen King,Frank Darabont        Crime,Drama       English
         USA      R        1994     10       14       142
```

5. Fix Mistakes

- Remove `<x> more credit[s]` from actors, directors and writers.
- Replace language value of `None` with `null`.
- Replace writers value of `See more` with `null`.
- Remove duplicate actors, directors and writers.
- Replace `release_day` null value with two derived `null` feature values.

6. Process Formatted Data

and output two data sets

Clean Data only contains examples with no missing values

```
0   tt0047478   8.8   139697   Takashi Shimura,Kamatari Fujiwara,Yukiko ...
1   tt0043014   8.6   81413    William Holden,Erich von Stroheim,Fred Clark, ...
2   tt0114814   8.7   406394   Gabriel Byrne,Benicio Del Toro,Kevin Spacey, ...
3   tt0108052   8.9   448339   Ralph Fiennes,Caroline Goodall,Liam Neeson, ...
```

Noisy Data contains manipulated examples based on the following:

- If example has over five missing values, **scrap it**.
- If example is missing release year, actors, writers or directors, **scrap it**.
- If example is missing rating or rating count, replace null with the standard **average** value of the data set.
- If example is missing genres, replace null with mode **Drama**.
- If example is missing languages, replace null with mode **English**.
- If example is missing country, replace null with mode **USA**.
- If example is missing MPAA rating, replace null with **unrated**.
- If example is missing release month, replace null with mode **10**.
- If example is missing runtime, replace with **average** value of data set.

```
0   tt0047478   8.8   139697   Takashi Shimura,Kamatari Fujiwara,Yukiko ...
1   tt0043014   8.6   81413    William Holden,Erich von Stroheim,Fred Clark, ...
2   tt0114814   8.7   406394   Gabriel Byrne,Benicio Del Toro,Kevin Spacey, ...
3   tt0108052   8.9   448339   Ralph Fiennes,Caroline Goodall,Liam Neeson, ...
```