Movies dataset
OSEMN methodology, Step 1:
Obtaining movies data from various sources

Stepan Oskin

July 24, 2019

Abstract

This document describes the process of obtaining movies data from various sources.

1 Data on Entertainment

La Fabbrica della Realta, an Entertainment Marketing Consulting agency, provides a good overview of available data sources for entertainment industry:

https://www.lafabbricadellarealta.com/open-data-entertainment/

Of the data that allows the movie and TV industries to function, very little is open. From the list of sources that feature more than 1'000 items provided by La Fabbrica della Realta, some sources have a proper open data license, while most of them don't.

Below is the description of some of the sources listed by La Fabbrica della Realta which have been chosen for the purposes of this analysis.

2 IMDb Datasets

IMDb datasets provide a trove of troves of entertainment data that are refreshed daily and available for **personal and non-commercial use**. You are allowed to hold local copies of this data, and it is subject to IMBb terms and conditions.

2.1 IMDb Licence

Compliance can be verified by referring to:

- Non-Commercial Licensing

  https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/
  G5JTRESSHJBBHTGX?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=3aefe545-f8d3-4562-976a-e5e
  pf_rd_r=MBQYD9CR3MEXPZJ2PSQS&pf_rd_s=center-1&pf_rd_t=60601&
  pf_rd_i=interfaces&ref_=fea_mn_lk1

- copyright/licence

  http://www.imdb.com/Copyright?pf_rd_m=A2FGELUUNOQJNL&pf_
  rd_p=3aefe545-f8d3-4562-976a-e5eb47d1bb18&pf_rd_r=MBQYD9CR3MEXPZJ2PSQS&
  pf_rd_s=center-1&pf_rd_t=60601&pf_rd_i=interfaces&ref_=fea_mn_lk2

2.2 IMDb Data Location

The dataset files can be accessed and downloaded from:
https://datasets.imdbws.com/
The data is refreshed daily.

2.3 IMDb Dataset Details

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A `\N` is used to denote that a particular field is missing or null for that title/name.

The available datasets are as follows:

title.akas.tsv.gz - Contains the following information for titles:

- `titleId` (string)—a tconst, an alphanumeric unique identifier of the title
- `ordering` (integer)—a number to uniquely identify rows for a given `titleId`
- `title` (string)—the localized title
- `region` (string)—the region for this version of the title
- `language` (string)—the language of the title

- `types` (array)—Enumerated set of attributes for this alternative title. One or more of the following: ”alternative”, ”dvd”, ”festival”, ”tv”, ”video”, ”working”, ”original”, ”imdbDisplay”. New values may be added in the future without warning

- `attributes` (array)—Additional terms to describe this alternative title, not enumerated

- `isOriginalTitle` (boolean)—0: not original title; 1: original title

title.basics.tsv.gz—Contains the following information for titles:

- `tconst` (string)—alphanumeric unique identifier of the title

- `titleType` (string)—the type/format of the title (*e.g.,* movie, short, tvseries, tvepisode, video, *etc.*)

- `primaryTitle` (string)—the more popular title / the title used by the filmmakers on promotional materials at the point of release

- `originalTitle` (string)—original title, in the original language

- `isAdult` (boolean)—0: non-adult title; 1: adult title

- `startYear` (YYYY)—represents the release year of a title. In the case of TV Series, it is the series start year

- `endYear` (YYYY)—TV Series end year. `\N` for all other title types

- `runtimeMinutes`—primary runtime of the title, in minutes

- `genres` (string array)—includes up to three genres associated with the title

title.crew.tsv.gz—Contains the director and writer information for all the titles in IMDb. Fields include:

- `tconst` (string)—alphanumeric unique identifier of the title

- `directors` (array of nconsts)—director(s) of the given title

- `writers` (array of nconsts)—writer(s) of the given title

title.episode.tsv.gz—Contains the tv episode information. Fields include:

- `tconst` (string)—alphanumeric identifier of episode
- `parentTconst` (string)—alphanumeric identifier of the parent TV Series
- `seasonNumber` (integer)—season number the episode belongs to
- `episodeNumber` (integer)—episode number of the `tconst` in the TV series

title.principals.tsv.gz—Contains the principal cast/crew for titles

- `tconst` (string)—alphanumeric unique identifier of the title
- `ordering` (integer)—a number to uniquely identify rows for a given `titleId`
- `nconst` (string)—alphanumeric unique identifier of the name/person
- `category` (string)—the category of job that person was in
- `job` (string)—the specific job title if applicable, else `\N`
- `characters` (string)—the name of the character played if applicable, else `\N`

title.ratings.tsv.gz—Contains the IMDb rating and votes information for titles

- `tconst` (string)—alphanumeric unique identifier of the title
- `averageRating`—weighted average of all the individual user ratings
- `numVotes`—number of votes the title has received

name.basics.tsv.gz—Contains the following information for names:

- `nconst` (string)—alphanumeric unique identifier of the name/person
- `primaryName` (string)—name by which the person is most often credited

- `birthYear`—in YYYY format
- `deathYear`—in YYYY format if applicable, else `\N`
- `primaryProfession` (array of strings)—the top-3 professions of the person
- `knownForTitles` (array of tconsts)—titles the person is known for

3 Other data sources

Below are some other movie data sources that can be added at a later stage.

- **OMDb API**
  The Open Movie Database
  The OMDb API is a RESTful web service to obtain movie information, all content and images on the site are contributed and maintained by our users.
  https://www.omdbapi.com
- **MovieLens**
  GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.
  https://grouplens.org/datasets/movielens/
- **Movie Review Data**
  This page is a distribution site for movie-review data for use in sentiment-analysis experiments. Available are collections of movie-review documents labeled with respect to their overall sentiment polarity (positive or negative) or subjective rating (e.g., ”two and a half stars”) and sentences labeled with respect to their subjectivity status (subjective or objective) or polarity.
  http://www.cs.cornell.edu/people/pabo/movie-review-data/
- **UCI ML repository Movie Data Set**
  This data set contains a list of over 10000 films including many older, odd, and cult films. There is information on actors, casts, directors, producers, studios, etc.
  https://archive.ics.uci.edu/ml/datasets/Movie