IMDb Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.

Data Location

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

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