Description of the IMDb and MovieLens datasets

Rasmus Pagh and Francesco Silvestri

October 18, 2013

Abstract

This note describes the available datasets containing information extracted from the IMDb and MovieLens websites. These datasets will be used in exercises that form the base for the final projects.

1 IMDb

The Internet Movie Database (IMDb) is an online database of information related to films, television programs, and video games. This includes actors, production crew, and fictional characters featured in these. There are some publicly available datasets containing a subset of the online database, which can be downloaded and processed offline.

1.1 D1: IMDb dataset provided by the University of Washington

We provide at `http://itu.dk/people/pagh/imdb-r.zip` (password provided to course students) a cleaned copy of the IMDb dataset provided by the University of Washington (originally from `http://www.webstepbook.com/supplements-2ed.shtml`). The dataset consists of one file containing seven tables: `actors`, `directors`, `directors_genres`, `movies`, `movies_directors`, `movies_genres`, `roles`. Each table begins with the line "`LOCK TABLES table_name WRITE;`", where `table_name` is one of the above names. Then follows a list of lines, each one containing one entry of the table with attributes separated by a comma. For example:

```
LOCK TABLES `actors` WRITE;
  213673,'Anthony','Hopkins','M',118
  358968,'Al','Pacino','M',65
  770247,'Julia','Roberts','F',59
  803350,'Meryl','Streep','F',90
```

The attributes of each table are:

- `actors`: actor id, first name, last name, gender (F or M), number of films;
  E.g.: `213673,'Anthony','Hopkins','M',118`

- `directors`: director id, first name, last name;
  E.g.: `71703,'Ridley','Scott'`
- `directors_genres`: director id, genre (e.g., Drama, Adventure,...), a float value in $[0, 1]$ (unknown meaning); a director can have many genres;
  E.g.: `71703,'Thriller',0.263158`
- `movies`: movie id, title, year, rank (NULL if not defined), running time in minutes (fake random value);
  E.g.: `138463,'Hannibal',2001,6.2, 137`
- `movies_directors`: movie id, director id;
  E.g.: `71703,138463`
- `movies_genres`: movie id, genre (e.g., Drama, Adventure,...); a movie can have many genres;
  E.g.: `138463,'Thriller'`
- `roles`: actor id; movie id, string describing the role;
  E.g. `213673,138463,'Dr. Hannibal Lecter'`

An SQL file that can be used for populating a database management system can be found at `http://www.webstepbook.com/supplements-2ed/databases/imdb.zip`. A smaller file, useful for testing, is also available here `http://www.webstepbook.com/supplements-2ed/databases/imdb_small.sql`. (Movie running times are not given in these datasets.) You can import these databases into MySQL using a terminal command such as: `mysql -u root -p < filename.sql`. However, note that most problems considered in the class can not be efficiently solved with a DBMS, so you will eventually have to parse the data in Java (or a similar programming language).

1.2 D2: Raw IMDb dataset

The IMDb web site offers (see `http://www.imdb.com/interfaces`) a dataset containing a lot of information on cast, crew, titles, technical details, and biographies. The dataset is organized into a set of compressed text files, where each file contains information on a specific aspect. A description is provided in the file tools/movie-database-faq within the dataset, and at the top of each single file. The raw IMDb dataset may not be easy to process, so we suggest to use the cleaned data set above dataset whenever possible.

For instance, the file `actors.list.gz` (and *actress.list.gz*) contains a list of actor names, and each name is followed by a set of lines, each one containing the title of a movies where the actor appeared and other details (year, awards, character name, position in the credits).

```
Hopkins, Anthony (I)  360 (2011)  [John]  <22>
                      84 Charing Cross Road (1987)  [Frank P. Doel]  <2>
                      A Bridge Too Far (1977)  [Lieutenant Colonel Frost]  <24>
                      A Century of Cinema (1994)  [Himself]
                      A Change of Seasons (1980)  [Adam Evans]  <2>
```

All files can be downloaded through anonymous ftp:

- `ftp://ftp.fu-berlin.de/pub/misc/movies/database/` (Germany)
- `ftp://ftp.funet.fi/pub/mirrors/ftp.imdb.com/pub/` (Finland)
- `ftp://ftp.sunet.se/pub/tv+movies/imdb/` (Sweden)

2 D3: MovieLens

The MovieLens web site collects user recommendations on movies. Three datasets of different sizes (100K, 1M, 10M ratings) are available at `http://grouplens.org/datasets/movielens/`.

The 10M and 1M datasets have the following structure (the 100K dataset slightly differs). Data are stored in three files, containing an entry per line; attributes of an entry are separated by two colons (::). Please refer to the `README.txt` file associated with each dataset for more information.

- `movies.dat`. The file contains information on movies and each entry contains three attributes: movie id, movie title, list of genres separated by |.
  E.g.: `1::Toy Story (1995)::Adventure|Animation|Children|Comedy|Fantasy`
- `ratings.dat`. Each line of this file represents one rating of one movie by one user, and has the following format: user id, movie id, 5-star rating (with half-star increments), timestamp. Every user included in the data set has rated at least 20 movies.
  E.g.: `33::1::4::849544161`
- `tags.dat`. Each line of this file represents one tag applied to one movie by one user, and has the following format: user id, movie id, tag, timestamp. Each tag is typically a single word or short phrase.
  E.g.: `2456::1::animation::1163101260`

Movie titles should match those found in IMDb, including year of release: however, they are entered manually, so errors and inconsistencies may exist.