# Introduction

IMDB does not provide a free API to search their data. You may come across some IMDB API's, but they resort to screen scraping techniques ([here](http://www.deanclatworthy.com/imdb/), [here](http://imdbapi.poromenos.org/) and [here](http://code.google.com/p/imdb-api/)). The good news is that IMDB does provide their [data](http://www.imdb.com/interfaces) in flat text files released every Friday. The bad news is that these files are formatted just to be difficult to use. ([Love this question](http://www.reddit.com/r/programming/comments/6m33s/why_does_imdb_restrain_an_official_api_but/).) Thanks IMDB!

The files can be download from the following FTP servers:

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

The files are formatted in a general way:

* CRC Line with file name and date
* Header data, sometimes instructions
* Header line (MOVIES LIST, etc.)
* Separator line, usually dashes
* Optionally more data and instructions

The "real" data is between the "header line" and the "separator line.

The IMDB site uses an integer as a primary key. Of course, this information is missing in the text files. They use the actual “title” of the movie as the primary key throughout all of the files.

﻿﻿

For example:

The Godfather -> 68646

URL to their Godfather site: [http://www.imdb.com/title/tt**0068646**](http://www.imdb.com/title/tt0068646)

"Primary key" in the text files: “Godfather, The (1972)”

## Prerequisites

The following programs, utilities and scripts are writing in [Ruby](http://ruby-lang.org/), bash and other [UNIX](http://en.wikipedia.org/wiki/Unix) command line tools (e.g. [sed](http://en.wikipedia.org/wiki/Sed), [awk](http://en.wikipedia.org/wiki/Awk), [grep](http://en.wikipedia.org/wiki/Grep), etc.). I have run all of these successfully on OS X, Linux (Ubuntu) and I’m sure they will work just fine under [Cygwin](http://www.cygwin.com/) for Windows.

Required Ruby Gems:

require 'rubygems'

require 'optparse'

require 'ostruct'

require 'digest/sha1'

## Programs to Process the LAME IMDB Files

The 1st pass at reformatting the files gets them into one-record-per-line format that can be used for further processing.

* movies.list.rb
  + This program reformats the movies.list file into 7 columns (key,title,type,year,episode\_title,season,episode)
* list\_cleaner.rb
  + This program reformats the other files into 3 or more columns (key, title, data, data…)

The field outputs of the files are explained in each file section below.

## The Files

Below is a table with each file and type “type” of file. File names in red indicate these are people associated with a movie. In that case it implies a many-to-many relationship. File names in gray have not been considered for processing at this time.

|  |  |  |
| --- | --- | --- |
| **Type** | **File Name** | **Program Output Format (Tab Delmited)** |
| Multi-line | actors.list | key,title,person\_key,person\_name,character |
| Multi-line | actresses.list | key,title,person\_key,person\_name,character |
| n/a | aka-names.list | n/a |
| n/a | aka-titles.list | n/a |
| n/a | alternate-versions.list | n/a |
| Multi-line | biographies.list | ?????????????????? |
| n/a | business.list | n/a |
| n/a | certificates.list | n/a |
| Multi-line | cinematographers.list | key,title,person\_key,person\_name |
| n/a | color-info.list | n/a |
| n/a | complete-cast.list | n/a |
| n/a | complete-crew.list | n/a |
| Multi-line | composers.list | key,title,person\_key,person\_name |
| Multi-line | costume-designers.list | key,title,person\_key,person\_name |
| 1 row/record | countries.list | key,title,country |
| Multi-line | crazy-credits.list | key,title,edit1,…editn (**variable columns!**) |
| Multi-line | directors.list | key,title,person\_key,person\_name |
| n/a | distributors.list | n/a |
| Multi-line | editors.list | key,title,person\_key,person\_name |
| 1 row/record | genres.list | key,title,genre |
| n/a | german-aka-titles.list | n/a |
| Multi-line | goofs.list | key,title,goof1,…goofn (**variable columns!**) |
| n/a | iso-aka-titles.list | n/a |
| n/a | italian-aka-titles.list | n/a |
| 1 row/record | keywords.list | key,title,keyword |
| 1 row/record | language.list | key,title,language |
| n/a | laserdisc.list | n/a |
| n/a | literature.list | n/a |
| 1 row/record | locations.list | key,title,location |
| n/a | miscellaneous-companies.list | n/a |
| n/a | miscellaneous.list | n/a |
| n/a | movie-links.list | n/a |
| 1 row/record | movies.list | key,title,type,year,episode\_title,season,episode |
| Multi-line | mpaa-ratings-reasons.list | key,title,rating,reason |
| Multi-line | plot.list | key,title,plot |
| Multi-line | producers.list | key,plot,person\_key,person\_name |
| n/a | production-companies.list | n/a |
| Multi-line | production-designers.list | key,plot,person\_key,person\_name |
| n/a | quotes.list | n/a |
| n/a | ratings.list | n/a |
| 1 row/record | release-dates.list | key,title,country,date,parsed\_date |
| 1 row/record | running-times.list | key,title,running\_time |
| n/a | sound-mix.list | n/a |
| n/a | soundtracks.list | n/a |
| n/a | special-effects-companies.list | n/a |
| Multi-line | taglines.list | key,title,tag\_line1,…tag\_linen (**variable columns!**) |
| n/a | technical.list | n/a |
| n/a | trivia.list | n/a |
| Multi-line | writers.list | key,plot,person\_key,person\_name |

# IMDB Files

## MOVIES.LIST - The Main File

Unfortunately, the movies.list file also contains TV shows and TV episode data. On a good note, the data is formatted with 1 record per line.  The records are TAB separated and follow this format:

IMDB Title<TAB>....(any number of tabs)....<TAB>Year(optional tabs and other data)

How do you know if it is a movie vs. TV?  All TV show names are in double quotes:

"Cheers" (1982) 1982-1993

"Cheers" (1982) {'I' on Sports (#6.2)} 1987

"Cheers" (1982) {2 Good 2 Be 4 Real (#4.7)} 1985

You will notice that the 1st line in the section of "Cheers" does not contain any episode data. Episode data is placed between curly braces. If the curly braces are missing in the title, this indicates that this is the TV show entry where the year will usually be a year range. If curly braces are present, the line will have episode information.

So, parsing this file we will end up with 3 types of "things":

1. Movies/Videos
2. TV shows
3. TV episodes

You will also see entries that have (TV) and (V) at the end of the title. These seem to indicate that is either a "made for TV movie" (TV) or a movie released to "video" (V). There does not seem to be any rhyme or reason how many tabs are between the title and year/year-year fields.

The movie.list.rb utility will take as input the movies.list file and reformat it with the following 7 fields:

{key}\t{title}\{type}\t{year}\t{episode\_title}\t{season}\t{episode}\n

Where:

1. {key} is the SHA1 sum of the 1st field in the file, the title (This is the “key” used to reference the other files which use the title as the key.)
2. {title} is the title of the entry, with TV episode information removed, if any
3. {type} will be: 1 = Movie/Video, 2 = TV Show, 3 = TV Episode
4. {year} will be either the year or the year range if it is a TV Show (type = 2)
5. {episode\_title} is the title of the TV show epsiode
6. {season} is the season of the episode 1..x
7. {episode} is the episode number of the episode 1..x

For example:

./movies.list.rb -c -f imdb/movies.list 1> data/movies.list.txt

It can process the file in less than 2 minutes.

Here is the "Army of Darkness" line from movies.list and then from movies.txt:

grep "Army of Darkness (1992)" ../../../../DATA/imdb/movies.list | cat -vet

Army of Darkness (1992)^I^I^I^I^I1992$

grep "Army of Darkness (1992)" data/movies.list.txt | cat -vet

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25^IArmy of Darkness (1992)^I1^I1992^I^I0^I0$

If you want to extract just the movies and videos from the file, select on the type column ($3):

awk 'BEGIN{FS="\t"}{ if ($3 == "1") { print $0 } }' < data/movies.list.txt

If you wan just the TV shows:

awk 'BEGIN{FS="\t"}{ if ($3 == "2") { print $0 } }' < data/movies.list.txt

To validate the file is being parsed properly you can do some sanity checks on the movies.txt file (this should print 1, 2, 3 on separate lines):

awk 'BEGIN{FS="\t"}{ print $3 }' < data/movies.list.txt | sort -u

You can see all of the years with this:

awk 'BEGIN{FS="\t"}{ print $4 }' < data/movies.list.txt | sort -u

You can do the same for the seasons and episode numbers ($6 and $7).

## ACTORS.LIST - "THE ACTORS LIST"

## ACTRESSES.LIST - "THE ACTRESSES LIST"

## AKA-NAMES.LIST - "AKA NAMES LIST" – N/A

## AKA-TITLES.LIST - "AKA TITLES LIST" – N/A

## ALTERNATE-VERSIONS.LIST - "ALTERNATE VERSIONS LIST" – N/A

## BIOGRAPHIES.LIST - "BIOGRAPHY LIST"

## BUSINESS.LIST - "BUSINESS LIST" – N/A

## CERTIFICATES.LIST - "CERTIFICATES LIST" – N/A

## CINEMATOGRAPHERS.LIST - "THE CINEMATOGRAPHERS LIST"

## COLOR-INFO.LIST - "COLOR INFO LIST" – N/A

## COMPLETE-CAST.LIST - "CAST COVERAGE TRACKING LIST" – N/A

## COMPLETE-CREW.LIST - "CREW COVERAGE TRACKING LIST" – N/A

## COMPOSERS.LIST - "THE COMPOSERS LIST"

## COSTUME-DESIGNERS.LIST - "THE COSTUME DESIGNERS LIST"

## COUNTRIES.LIST - "COUNTRIES LIST"

The countries.list file is formatted similarly to the movies.list file with one line per entry.  However, there may be duplicate entries for each movie!  For example:

grep "La sconosciuta (2006)" ../../../../../DATA/imdb/countries.list

La sconosciuta (2006) Italy

La sconosciuta (2006) France

I have no idea why. So, if you are joining the data be aware of this issue.

The list\_cleaner.rb program will cleanup this list nicely:

./list\_cleaner.rb -c -f ../../../../DATA/imdb/countries.list > data/countries.list.txt

You can validate the countries by looking at this list:

awk 'BEGIN{FS="\t"}{ print $3 }' < data/countries.list.txt  | sort -u

Let's find out from what country "Army of Darkness" originates.  We will use the SHA1 has to search the files:

grep b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 data/movies.list.txt

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) 1 1992 0 0

grep b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 data/countries.list.txt

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) USA

I see USA!

## CRAZY-CREDITS.LIST - "CRAZY CREDITS"

## DIRECTORS.LIST - "THE DIRECTORS LIST"

## DISTRIBUTORS.LIST - "DISTRIBUTORS LIST" – N/A

## EDITORS.LIST - "THE EDITORS LIST"

## GENRES.LIST - "THE GENRES LIST"

## GERMAN-AKA-TITLES.LIST - "AKA TITLES LIST GERMAN" – N/A

## GOOFS.LIST - "GOOFS LIST"

## ISO-AKA-TITLES.LIST - "AKA TITLES LIST ISO" – N/A

## ITALIAN-AKA-TITLES.LIST - "AKA TITLES LIST ITALIAN" – N/A

## KEYWORDS.LIST - "THE KEYWORDS LIST"

## LANGUAGE.LIST - "LANGUAGE LIST"

## LASERDISC.LIST - "LASERDISC LIST" – N/A

## LITERATURE.LIST - "LITERATURE LIST" – N/A

## LOCATIONS.LIST - "LOCATIONS LIST"

## MISCELLANEOUS-COMPANIES.LIST - "MISCELLANEOUS COMPANY LIST" – N/A

## MISCELLANEOUS.LIST - "THE MISCELLANEOUS FILMOGRAPHY LIST" – N/A

## MOVIE-LINKS.LIST - "MOVIE LINKS LIST" – N/A

## MOVIES.LIST - "MOVIES LIST" – See 1st section

## MPAA-RATINGS-REASONS.LIST - "MPAA RATINGS REASONS LIST"

## PLOT.LIST - "PLOT SUMMARIES LIST"

## PRODUCERS.LIST - "THE PRODUCERS LIST"

## PRODUCTION-COMPANIES.LIST - "PRODUCTION COMPANIES LIST" – N/A

## PRODUCTION-DESIGNERS.LIST - "THE PRODUCTION DESIGNERS LIST"

## QUOTES.LIST - "QUOTES LIST" – N/A

## RATINGS.LIST - "MOVIE RATINGS REPORT" – N/A

## RELEASE-DATES.LIST - "RELEASE DATES LIST"

The release-dates.list file is formatted similarly to the movies.list file with one line per entry.  However, there may be duplicate entries for each movie!  For example:

grep "Army of Darkness (1992)" ../../../../DATA/imdb/release-dates.list

Army of Darkness (1992) Spain:9 October 1992 (Sitges Film Festival) (director's cut)

Army of Darkness (1992) UK:November 1992 (London Film Festival)

Army of Darkness (1992) Taiwan:21 November 1992

Army of Darkness (1992) France:January 1993 (Avoriaz Film Festival)

Army of Darkness (1992) Portugal:February 1993 (Fantasporto Film Festival)

It appears that every time the move is released it will get an entry in this file. The Army of Darkness has 47 entries!

The list\_cleaner.rb program will cleanup this list nicely:

./list\_cleaner.rb -c -f ../../../../DATA/imdb/release-dates.list > data/release-dates.txt

Looking for Army of Darkness with the SHA1 key renders:

grep b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 data/release-dates.list.txt

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Argentina:1 November 2003

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Australia:22 April 1993

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Austria:14 May 1993

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Belgium:12 March 1993

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Belgium:5 January 1994

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Brazil:December 1993

b8d3d50b3a16d00c7d8caeb8d22005e6c8e9db25 Army of Darkness (1992) Canada:1 October 2002

## RUNNING-TIMES.LIST - "RUNNING TIMES LIST"

## SOUND-MIX.LIST - "SOUND-MIX LIST" – N/A

## SOUNDTRACKS.LIST - "SOUNDTRACKS LIST" – N/A

## SPECIAL-EFFECTS-COMPANIES.LIST - "SFXCO COMPANIES LIST" – N/A

## TAGLINES.LIST - "TAG LINES LIST"

## TECHNICAL.LIST - "TECHNICAL LIST" – N/A

## TRIVIA.LIST - "FILM TRIVIA" – N/A

## WRITERS.LIST - "THE WRITERS LIST"

# Make a Real Database