##A Movie Rating Dataset Collected From Twitter

MovieTweetings is a dataset consisting of ratings on movies that were contained in well-structured tweets on Twitter. This dataset is the result of research conducted by [Simon Dooms] (<http://scholar.google.be/citations?user=owaD8qkAAAAJ>) (Ghent University, Belgium) and has been presented on the [CrowdRec 2013 workshop](http://crowdrec2013.noahlab.com.hk/) which is co-located with the [ACM RecSys 2013 conference](http://recsys.acm.org/recsys13/). Please cite the [corresponding paper](http://crowdrec2013.noahlab.com.hk/papers/crowdrec2013_Dooms.pdf) if you make use of this dataset. The presented slides can be found [on slideshare] (<http://www.slideshare.net/simondooms/movie-tweetings-a-movie-rating-dataset-collected-from-twitter>).

Follow us on Twitter ([@mvtweetings](https://twitter.com/mvtweetings)) for the latest news, info and fun facts about the dataset.

Bibtex: @conference{Dooms13crowdrec, author = {Dooms, Simon and De Pessemier, Toon and Martens, Luc}, title = {MovieTweetings: a Movie Rating Dataset Collected From Twitter}, booktitle = {Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013}, year = {2013} }

An excerpt of the abstract of the paper:

Public rating datasets, like MovieLens or Netflix, have long been popular and widely used in the recommender systems domain for experimentation and comparison. More and more however they are becoming outdated and fail to incorporate new and relevant items. In our work, we tap into the vast availability of social media and construct a new movie rating dataset 'MovieTweetings' based on public and well-structured tweets. With the addition of around 500 new ratings per day we believe this dataset can be very useful as an always up-to-date and natural rating dataset for movie recommenders.

The goal of this dataset is to provide the RecSys community with a live, natural and always up-to-date movie ratings dataset. The dataset will be updated as much as possible to incorporate rating data from the newest tweets available. The earliest rating contained in this dataset is from 28 Feb 2013, and I pledge to keep this system up and running for as long as I can. Note however that this dataset is automatically gathered and therefore depending on the continuation of the IMDb apps and Twitter API.

Don't hesitate to contact me for any comments, questions or proposals you might have.

##Ratings from Twitter

As said, this dataset consists of ratings extracted from tweets. To be able to extract the ratings, we query the Twitter API for well-structured tweets. We have found such tweets originating from the social rating widget available in IMDb apps. While rating movies, in these apps, a well-structured tweet is proposed of the form:

"I rated The Matrix 9/10[*http://www.imdb.com/title/tt0133093/*](http://www.imdb.com/title/tt0133093/)#IMDb"

On a daily basis the Twitter API is queried for the term **"I rated #IMDb"**. Through a series of regular expressions, relevant information such as user, movie and rating is extracted, and cross-referenced with the according IMDb page to provide also genre metadata. The numeric IMDb identifier was adopted as item id to facilitate additional metadata enrichment and guarantee movie uniqueness. For example, for the above tweet the item id would be **"0133093"** which allows to infer the corresponding IMDb page link (add [*http://www.imdb.com/title/tt*](http://www.imdb.com/title/tt)). The user id simply ranges from 1 to the number of users.

##The dataset

Since this dataset will be updated regularly we have structured the dataset in different folders /latest and /snapshots. The /latest folder will always contain the complete dataset as available at the time of the commit, while the /snapshots contain fixed portions of the dataset to allow experimentation and reproducibility of research. The 10K snapshot represents the ratings from the first 10,000 collected tweets, 20K the first 20,000, and so on.

The dataset files are modeled after the [MovieLens dataset] (<http://www.grouplens.org/node/73>) to make them as interchangeable as possible. There are three files: **users.dat**, **items.dat** and **ratings.dat**.

###users.dat

Contains the mapping of the users ids on their true Twitter id in the following format: userid::twitter\_id. For example:

1::177651718

We provide the Twitter id and not the Twitter @handle (username) because while the @handle can be changed, the id will always remain the same. Conversions from Twitter id to @handle can be done by means of an online tool like [Tweeterid] (<http://tweeterid.com/>) or simply through the Twitter API itself. The mapping provided here again facilitates additional metadata enrichment.

###items.dat

Contains the items (i.e., movies) that were rated in the tweets, together with their genre metadata in the following format: movie\_id::movie\_title (movie\_year)::genre|genre|genre. For example:

0110912::Pulp Fiction (1994)::Crime|Thriller

The file is UTF-8 encoded to deal with the many foreign movie titles contained in tweets.

###ratings.dat

In this file the extracted ratings are stored in the following format: user\_id::movie\_id::rating::rating\_timestamp. For example:

14927::0110912::9::1375657563

The ratings contained in the tweets are scaled from 0 to 10, as is the norm on the IMDb platform. To prevent information loss we have chosen to not down-scale this rating value, so all rating values of this dataset are contained in the interval [0,10].