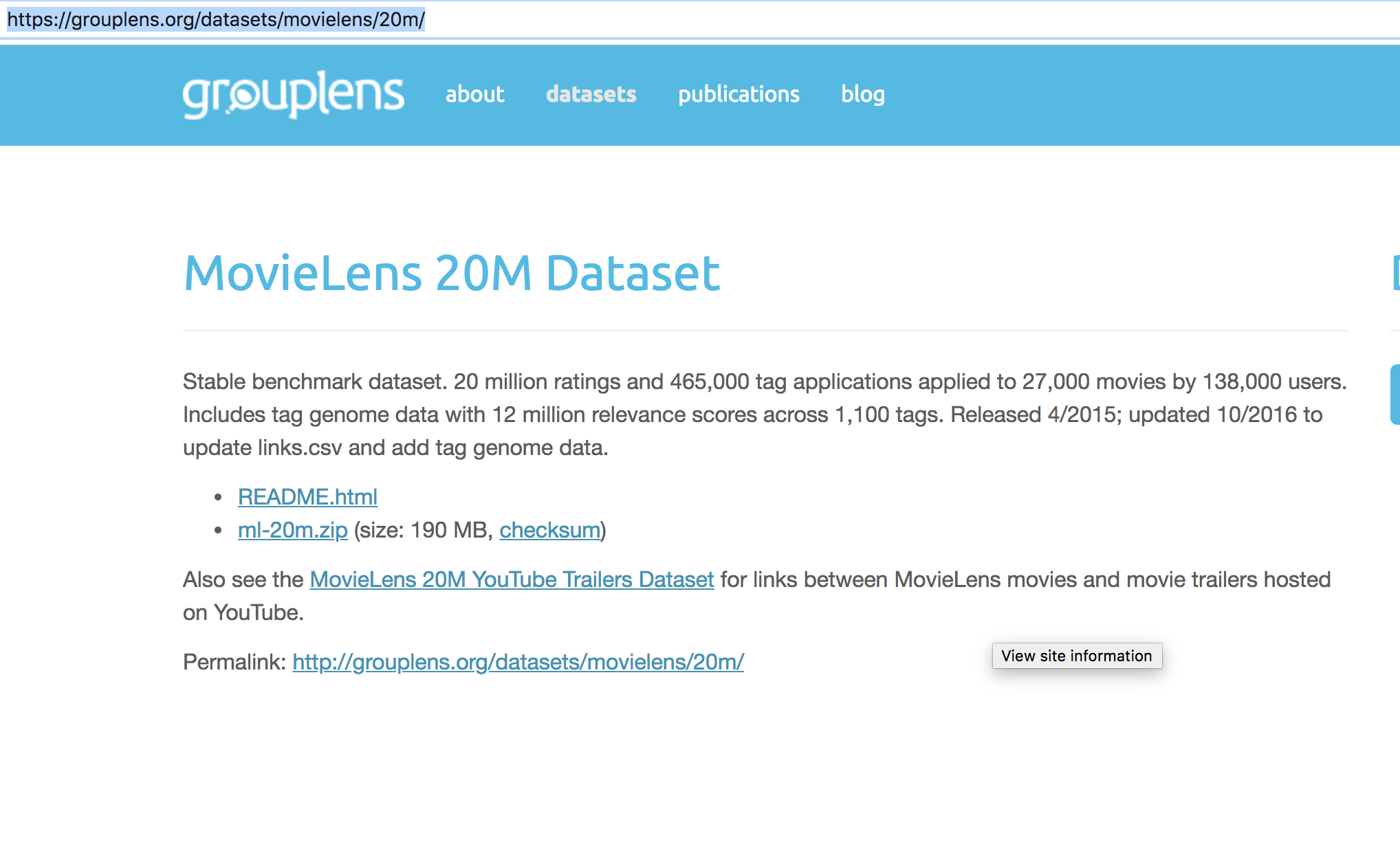
**Data Used in this exploration:**

The dataset is a zip file located at:

<https://grouplens.org/datasets/movielens/20m/>



Once downloaded and opened, the files look like below:



-Ratings contains the 20 million ratings given by 138,000 viewers to 27,000 movies. Ratings range from 0.5 to 5 by ½ steps. This is used to do the collaborative part.

-Genome-tags contains the name of 1,100 tags (afterlife, Africa, england, hotel, Robert dowey junior) – the tags – and their id.

-Genome-scores contains the movieid / tagid and the relevance of the tagid to the movie.

This is used to do the content part. The 1,100 tags are applied 465,000 times to 27,000 movies

-Movies contains the movieId and the title of the movies.

-Links link movies to the IMDB and the movie database (TMDB).

Tags – which tags were given to which movie by which viewer, when is not used here.

Files need to be put in the data folder. Running the 2 notebooks generate additional files and afterwards the data folder looks like:



-Cosine file contains the distance between movies.

-Svdpred contains the weights used to predict the ratings

-Contentreco contains the 100 movies selected during content filtering

-Hybridreco contains the subset of 100 movies which have not beed rated and ranked by predicted rate. What is returned in the application is the first 5 records.

-Indice is a table title-index

-Movietagged movie/tag

-Moviemapped: index – movieId –Title. Index goes from 0 to 10380. MovieId from 1 to 131170.