Pig Assignment

The goal for this assignment is to use Pig for data science purposes.

You are already familiar with u.data file located in ml-100k folder in your local file system. If you recall the u.data file contains 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab separated list of user id, item id, rating, timestamp. The time stamps are unix seconds since 1/1/1970 UTC.

There is another file in ml-100k folder called u.item. This file contains about the items (movies). This is a pipe separated list of movie id, movie title, release date, video release date, IMDb URL, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. The last 19 fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the u.data data set.

Write a Pig script called movie\_rating.pig to print out the complete list of all movies rated with the following information per line -- movie id, movie title, release date, IMDB URL, average rating for the movie.

**What to turn in?**

A sample run with output as a Word document. I have given you a template for assignment 8 solutions. Follow the guidelines, fill the document and submit.