Homework Assignment #2

(Data Wrangling Principles)

Congratulations are in order! Your initial project as a Data Scientist for Movie Intelligence at Netflicks.com was well-received by your boss. He's quite excited by the possibilities that Data Science techniques have to offer, but he's been hearing scary stories at cocktail parties about the "data wrangling problem". Your next assignment is to become "dangerous" in terms of your knowledge of data wrangling by reading the assigned mini-book (hopefully you're done with that already) and doing some back-of-the-envelope analyses of some of the problems and potential solutions for some popular movie data sets. You will continue to use the IMDB data in this next assignment - and in addition you will make use of another publically available data set about movies called ***MovieLens***.

Get Ready...

Finish reading ***Principles of Data Wrangling: Practical Techniques for Data Preparation*** if you have not already done so.

Get (Data) Set...

Your next step is to grab a copy of the relevant ***MovieLens*** data. Go to the MovieLens web site (https://grouplens.org/datasets/movielens/) and download a copy of the *Small* version of the *MovieLens Latest Datasets*. You're just looking to understand the data and the kind of wrangling that it may need, for now, so the small version of the data will be enough for this assignment.

Go...!

It's time to get back to work. First read the *README* file that's next to the data's download link and then unzip the downloaded data. Your boss wants you to examine the ***MovieLens*** data and, considering both this new data and the IMDB data that you've already begun using, to answer the following questions. You can use whatever tools you like to explore the new data, but you shouldn't need anything more than your favorite text editor - in fact, you should use a text editor even if you also use other tools, as it will be important to examine the raw data in order to properly assess the data's "issues" and answer all of the questions. (You may also want to look at the IMDB data that way - perhaps just looking at the *.head.tsv samples to refresh your memory about that data.) Note: When answering the following questions, consider each of the files as being its own dataset in the book's terminology.

1. Your work begins at the Raw Data Stage. It's time to familiarize yourself with the data by following the book's prescription for doing so.

a. Start by understanding the *structure* of your newly acquired ***MovieLens*** data. To do so, for each of the four data files, briefly answer the Basic Questions to Assess Structure on pp. 16-17 of the book.
b. Next, improve your understanding of the semantics and *granularity* of the data by answering the Basic Questions to Assess Data Granularity on p. 18 of the book for each of the data files.
c. Since this data was neatly assembled for you, skip the Accuracy and Temporality analyses - though do review those questions so you know what you're skipping. Instead, do a *scope* analysis by answering the Basic Questions to Assess Data Scope on p. 22-23 for each data file. Though you're just using a sample here, your scope answers should be about the *Full* version of the dataset. (Refer back to the site where you got the data for more information about the full version.)

2. Now that you've hopefully wrapped your head around the new data, it's time to think about the problem of wrangling it. Fast forward to the transformation and refinement challenges and opportunities posed by this data. Carefully examine the four data files (and the IMDB files where asked to do so) again and answer the following questions:
   a. There is at least one piece of information that's embedded in a place where it doesn't really belong - i.e., it's a classic *value extraction* use case! Identify this piece of information and explain how you might go about extracting its values. (MovieLens)
   b. Another common data transformation involves extracting ("lifting") the content of one field into several fields for further analysis. Do you see any fields in either the ***MovieLens*** data files or the IMDB data files for which this might be a useful step? Explain. (MovieLens, IMDB)
   c. Aggregation is an example of an inter-record data transformation step that can sometimes be useful when preparing data for further analysis. Using the ratings file as an example, give an example of how one might choose to aggregate the ratings data - and include a snippet of what the resulting (transformed) data file would look like. (MovieLens)
   d. The movies file, as provided, would likely be the most problematic ***MovieLens*** file to deal with. How might you transform dataset to make it more amenable to processing by table-oriented tools (e.g., by Excel or PostgreSQL)? Explain, and again include a snippet showing what the transformed data file would look like. (MovieLens)
   e. Your boss watches the new TV show called the "Wisdom of the Crowd" and he wants to exploit that wisdom to do movie analytics within Netflicks.com. Both of your data collections include reviews for movies - what transformation steps or operations would be needed in order to combine them? Explain briefly, again including a snippet showing what the resulting combined dataset would contain. (MovieLens, IMDB)

What To Turn In

Use your favorite word processing application (Word, Google Docs, …) to document your analysis and use EEE to turn in a PDF file of your writeup. (Please be sure to put your name somewhere on the first page of your writeup. :-))