Capstone Project : Movie Critic Review Bias (Rotten Tomatoes)

Dimitri Kourouniotis , March 2017

Dataset Gathering:

Summary:

Generated ~ 600,000 movie reviews scores from ~1200 RT Movie Critics for analysis. I used Rvest to web-scrape the dataset.

Part 1 – Collect Movie Critics Names:

I created R script to read through the catalog of critics on Rotten Tomatoes and generate a list of critics' names. Using excel I corrected the list for non-English characters so I could construct working URLs. That initial list was ~2700 names. I manipulated the name string to remove other unwanted characters to make a list of ~2700 working URLs.

Running through these 2700 URLs created a dataset of the number of reviews each critic has published in RT. The list was cleaned up to remove critics with 0 entries and shortened to those with 26 or more reviews only. This final dataset of critics I worked with is ~1200.

Each critic's section has 50 reviews per page with several critics having many thousand reviews. With the names of the critics and the number of reviews for each I created a list of URLs of pages for each critic. This is a list of approximately 12,000 URLs.

Reference Script: "*Capstone1 critic names list.R*"

Part 2 – Collect Movie Reviews:

With the ~12,000 URLs I scraped up to 50 reviews per page for a total of ~ 600,000. This dataset contains the movie title, the reviewer's name, the reviewer's score, publication date and the first few lines of the review.

This script had to be run in batches as occasional webpage errors would occur. I merged the data sets from the different batches and removed duplicates. The script took several hours to run.

Some cleanup that can be done is to remove carriage returns and normalize the review score method since each critic writing for their own publication grades a movie with a different method whether it be out of 4 or 5 stars, or a letter grade for example.

Reference Script: "*Capstone2 reading 12000 review pages.R*"

Part 3 – Collecting Tomato Meter Scores:

The third part is to collate the movie titles and their Tomato Meter (tMeter) Score. This script is very similar to the 12000 URL script. For this I only collected the movie titles and tMeter scores on each page. There are additional advertising elements on each page for current movie promotions etc. There was a consistent extra 24 tMeter scores retrieved from the page not associated with the movies on the review page. To remove these I created 24 NA entries with each page that I then removed.

Some additional cleanup that I can do here is to remove (older) movies that a critic has reviewed in the past but that do not contain a tMeter score as averaged by RT.

Reference Script: "*CapStone3 Reading TMeter Score.R*"