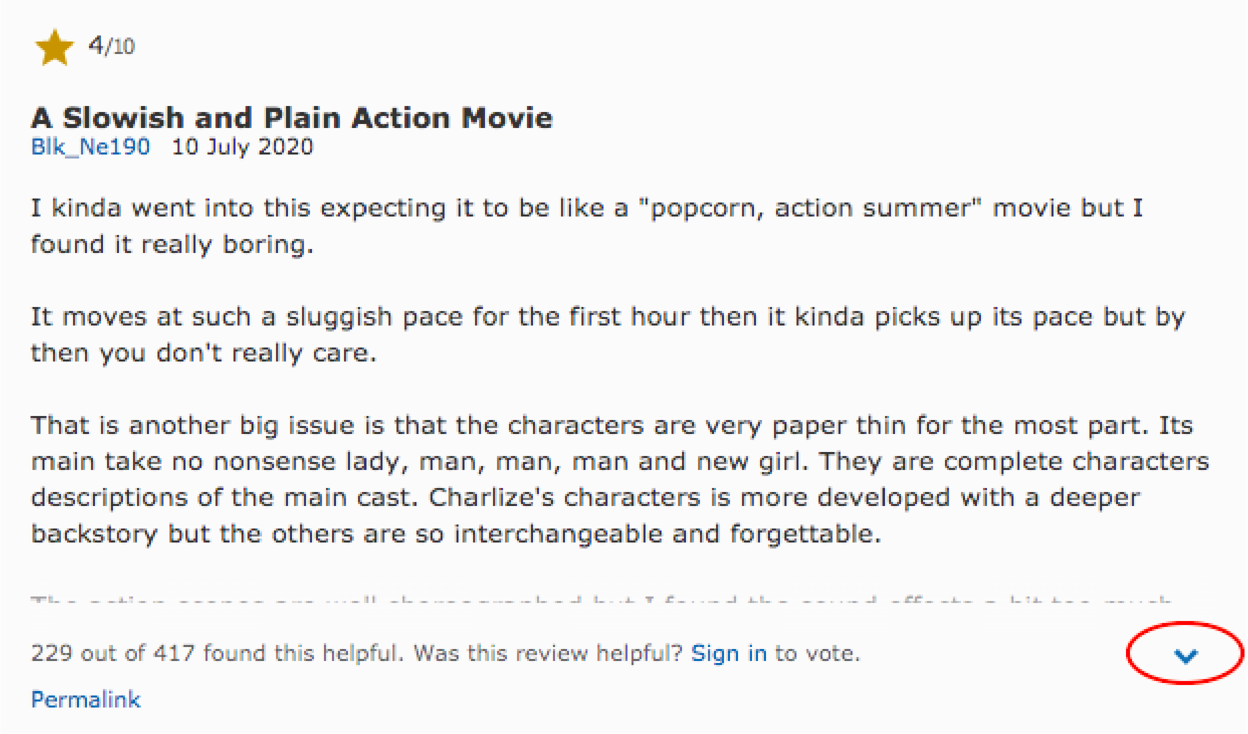
1. The starting point could be this webpage (Links to an external site.).

- The URLs of user review pages have the same pattern and can be construct programmatically.

2. Collect user reviews and ratings (note that some reviews may have no ratings) for 100 movies, 100 reviews per movie (with spoilers excluded; note that on each review page, there is a checkbox labeled "Hide Spoilers").

- Use R Selenium (covered in the last lab) to render full content before scraping a user review page:
  - tick the checkbox to hide spoilers
  - load all reviews for each movie by repeatedly clicking the “load more” button at the bottom of a review page. The number of clicks can be determined by the number of reviews displayed above the "Hide Spoilers" checkbox.
  - load the text content for each long review by clicking the down arrow button.



★ 4/10

A Slowish and Plain Action Movie

Blk_Ne190 10 July 2020

I kinda went into this expecting it to be like a "popcorn, action summer" movie but I found it really boring.

It moves at such a sluggish pace for the first hour then it kinda picks up its pace but by then you don't really care.

That is another big issue is that the characters are very paper thin for the most part. Its main take no nonsense lady, man, man, man and new girl. They are complete characters descriptions of the main cast. Charlize's characters is more developed with a deeper backstory but the others are so interchangeable and forgettable.

The action scenes are well choreographed but I found the sound effects a bit too much

229 out of 417 found this helpful. Was this review helpful? Sign in to vote.

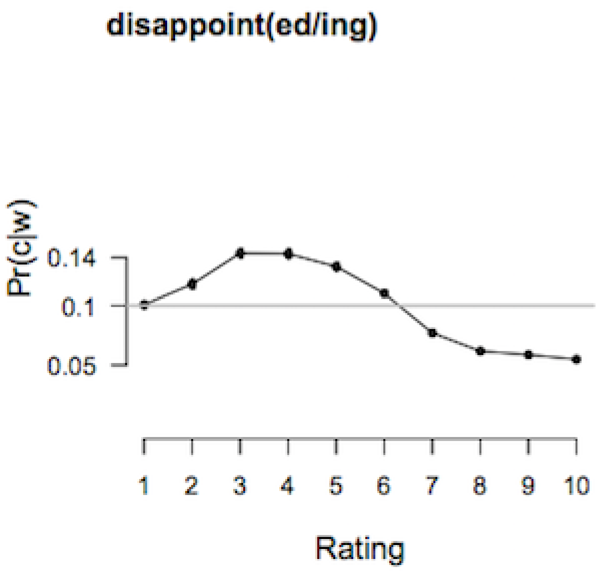
Permalink

3. Sample 2500 positive reviews and 2500 negative reviews (for example, a movie receiving a rating higher than 6 is considered positive or getting thumbs-up, while one with a rating less than 5 is negative or getting thumbs-down) to form a movie review corpus.

- Note that the most noteworthy fact for review corpora collected from the Web is that the majority of reviews are highly positive, with 6-star to 10-star reviews. If you cannot find a sufficient number of negative reviews from your 10,000 reviews, report the imbalance of the distribution.

4. Use the `BING` lexicon to find sentiment-associated words and predict sentiment for each review. Evaluate outcomes by contrasting them with actual ratings.

5. Find 5 most commonly-used positive and negative words in the corpus. Create a multipanel plot. In each panel, plot a graphic similar to the following one, which describes the extent to which a word affects the rating of the review where it appears. Specifically, w represents the word, while c represents categories defined in terms of ratings. For instance, the point at rating 3 means that a review containing "disappoint" has probability 0.14 of receiving rating 3.



Criteria for evaluating your submission

Your submission should be an R markdown file that wraps up your write-up and code, and a csv file containing the dataset you use. It will be evaluated based on the following criteria:

- Whether the interpretation is clear and in-depth. The write-up in the R markdown file should clearly explain the steps you take to perform the analyses, and accurately interpret the results, with appropriate caveats for what the technique can and cannot do. Visualize the results of your analyses whenever possible.
- Whether the code is reproducible (including the code for Web scraping), with necessarily detailed annotation.
- Whether the dataset created from Web scraping and used as the input to your analyses is present.

If your group faces a severe free rider problem, and if it cannot be addressed by your private efforts, please inform us by dropping me or Charles an email. Once it is verified, grade penalty will be exercised accordingly.

Rubrics

- Correctness: Deductions resulting from mistakes will be made at the discretion of the grader.
- **Knitting**: **-5%** deduction if the Rmd file you submit does not knit correctly (i.e., if there are errors and no HTML file is produced when the grader attempts to knit your Rmd file.) If your Rmd file fails to knit, you will be contacted by the grader and will be given 24 hours to resubmit your work. You will need to trace the source of the error(s) and correct it.
- **Style**: Coding style is important. You will receive a deduction of up to 10% if you do not adhere to good coding style. Your code is considered to have a good coding style if:
  -
    - good, consistent coding style
    - appropriate use of variables
    - appropriate use of functions
    - good commenting
    - good choice of variable names
    - appropriate use of inline code chunks

What constitutes plagiarism in a coding class?

The course collaboration policy allows you to discuss the problems with other students, but requires that you complete the work on your own. Every line of text and line of code that you submit must be written by you personally. You may not refer to another student's code, or a "common set of code" while writing your own code. You may, of course, copy/modify lines of code that you saw in lecture or lab.

You may find a discussion from the Computer Science and Engineering Department at the University of Washington (Links to an external site.) helpful in understanding the bounds of the collaboration policy