HW 3: Scrape Movie Reviews

Choose one of your favorite movies and find the URL of "Top Critics" at `rottentomatoes.com`

- e.g. for the latest movie 'Soul', the URL is: https://www.rottentomatoes.com/m/soul_2020/reviews?type=top_critics

- Q1. Write a function to scrape all **Top Critics on the first page**, including, **reviewer's name** (see (1) in Figure), **source** (see (2) in Figure), **content** (see (3) in Figure), **review date** (see (4) in Figure), and **score** (see (5) in Figure).

  - `Input` : "Top Critics" page URL
  - `Output` : save all reviews as a DataFrame of columns (`reviewer, source, content, date, score`). E.g., for the given URL, you can get 20 reviews.
  - If a field, e.g. score, is missing, use `None` to indicate it.

In [51]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd

def getReviews(page_url):

    reviews= None

    # Add your code here

    return reviews
```

In [52]:
```
url = 'https://www.rottentomatoes.com/m/soul_2020/reviews?type=top_critics'

reviews=getReviews(url)
reviews.head(5)
```

Out[52]:

	reviewer	source	content	date	score
0	Stephen Romei	The Australian	What follows is an absolute delight. It will m...	February 13, 2021	4/5
1	Kambole Campbell	Hyperallergic	It shortchanges itself at every turn by overex...	February 8, 2021	None
2	Nick Schager	The Daily Beast	It's got a bebopping spirit that's difficult t...	February 3, 2021	None
3	Udita Jhunjhunwala	Scroll.in	Doctor, Powers and Mike Jones have written a s...	January 26, 2021	None
4	Robert Daniels	Polygon	Even with the film's bid for a suspension of r...	January 25, 2021	None

- Q2 (Bonus). Modify the function you defined in Q1 to Scrape **Top Critics on all the pages**. Since a movie may have multiple pages, use the **next** button (see (6) in Figure) to navigate to the next page. Continue scraping all the pages until the `next button` is greyed out. `Please don't hardcode the page number in the URL because the number of pages varies by movie` .

Note:

- Test your function with a few movies to make your function is generic enough. A few URLs for testing:
  - https://www.rottentomatoes.com/m/soul_2020/reviews?type=top_critics
  - https://www.rottentomatoes.com/m/coco_2017/reviews?type=top_critics
- Follow the reference code structure below and save your script as .py file and submit to Canvas

In [49]:
```
# best practice to test your class
# if your script is exported as a module,
# the following part is ignored
# this is equivalent to main() in Java

if __name__ == "__main__":

    #url = 'https://www.rottentomatoes.com/m/soul_2020/reviews?type=top_critics'
    url = 'https://www.rottentomatoes.com/m/coco_2017/reviews?type=top_critics'

    reviews=getReviews(url)
    print(reviews.head(5))
```

```
           reviewer                     source  \
0    Robert Daniels              812filmreviews
1      Mark Kermode  Kermode & Mayo's Film Review
2  Mey Valdivia Rude                Autostraddle
3      Linda Marric                    HeyUGuys
4       Ryan Gilbey               New Statesman

                                        content              date score
0  In a country with an ever increasing Hispanic ...  September 5, 2018   3/4
1  I don't think there's any question that Coco i...    August 28, 2018  None
2  Several times I found myself sobbing without k...      April 20, 2018  None
3  A wonderful return to form for Pixar, who agai...    January 26, 2018   4/5
4  The film has a galloping rhythm, and the anima...    January 25, 2018  None
```

In [ ]:
```
```