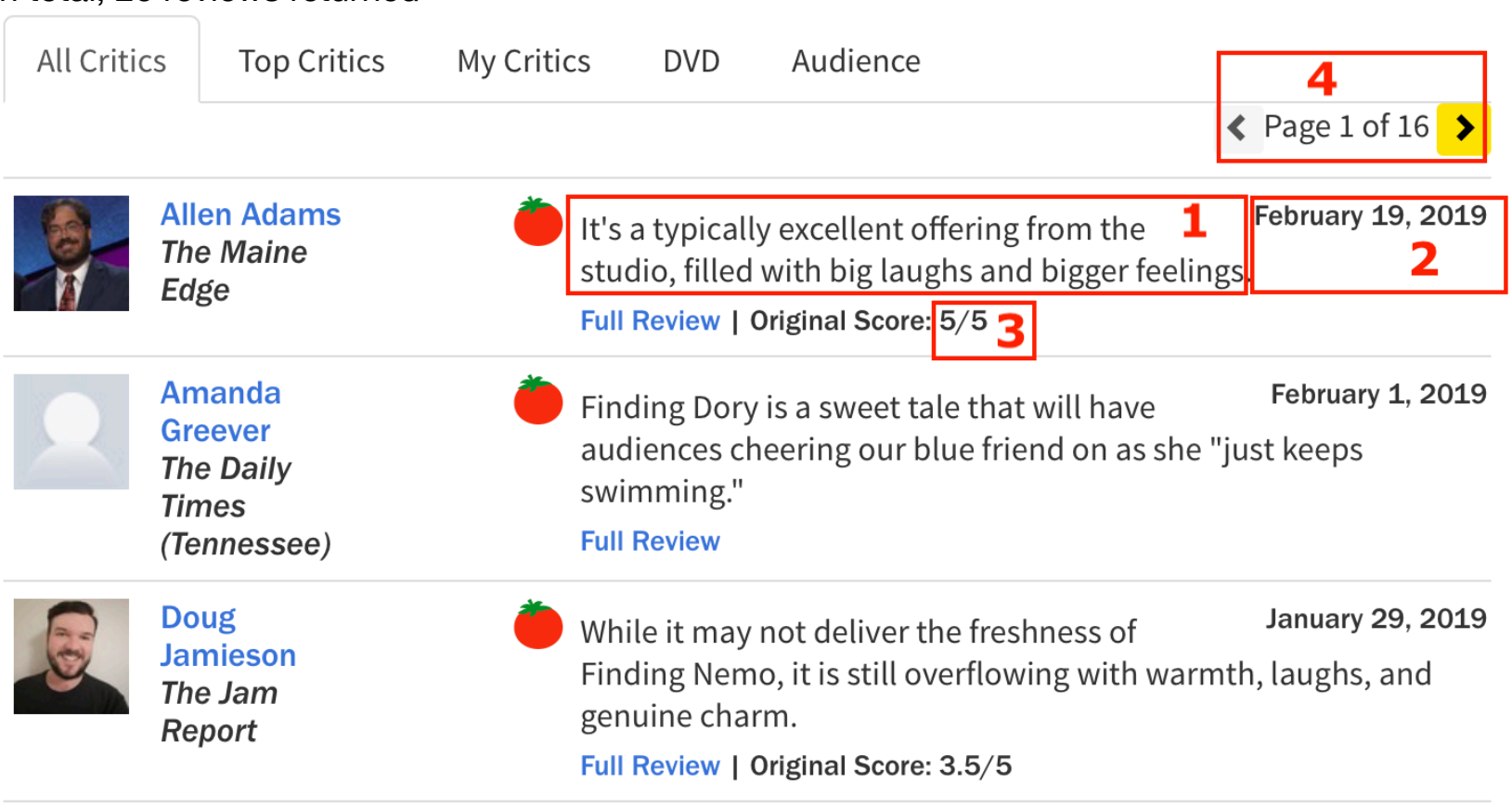
Assignment 3: Web Scraping

Q1. Scrape Movie Reviews

- Choose one of your favorite movies and find this id of this movie at rottentomatoes.com
- Write a function getData(*movie_id*) to scrape reviews, including review date (see (2) in Figure), review description (see (1) in Figure), and score (see (3) in Figure) from the current page.
  - Input: movie id in rottentomatoes
  - Output: a list of 20 tuples, i.e. [("February 19, 2019", "It's a typically excellent offering from the..." , "5/5"), ...]
- Test your function with a few movies to make your function is generic enough

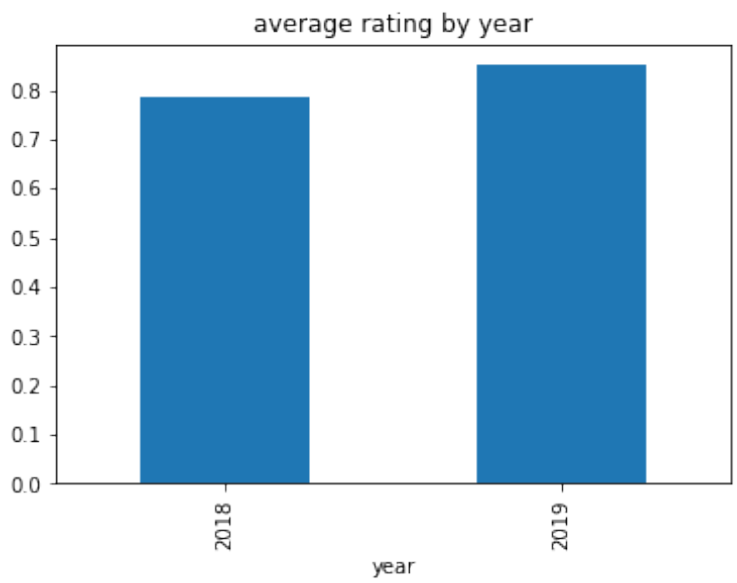
Example:

- https://www.rottentomatoes.com/m/finding_dory/reviews/ (https://www.rottentomatoes.com/m/finding_dory/reviews/)
- in total, 20 reviews returned



Q2. Plot data

- Create a function plot_data which
  - takes the list of tuples from Q1 as an input
  - converts the ratings to numbers. For example, 3.5/5 is converted to 0.7. For all reviews without a rating or with an alphabetic rating (e.g. A), set its rating to None
    - Hint: you can use try/except block to handle ratings which cannot be converted floats. See https://stackoverflow.com/questions/379906/how-do-i-parse-a-string-to-a-float-or-int-in-python (https://stackoverflow.com/questions/379906/how-do-i-parse-a-string-to-a-float-or-int-in-python) for reference.
  - calculates the average rating by the year of the review date
  - plots a bar chart for the average rating of each year. The plot may look similar to the figure below.



Q3 (Bonus) Expand your solution to Q1 to scrape all the views for a movie.

- Write a function getFullData(*movie_id*) to scrape reviews in all the pages. For the example shown in Figure of Q1, reviews are organized into 16 pages (See (4) of the figure). Scrape reviews from all the 16 pages. Save the result similar to Q1.
- Note, you **should not hardcode** the number of pages, because the number of pages varies by movies. Instead, you should dynamically determine if the next review page exists or not.

In [5]:

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import matplotlib.pyplot as plt


 # Q1
def getData(movie_id):


    data=[]  # variable to hold all book data


    # your code here

    return data

#Q2
def plot_data(data):

    # fill your code here


# Q3
def getFullData(movie_id):
    data=[]

    # fill your code here

    return data



if __name__ == "__main__":

    # Test Q1
    data=getData("finding_dory")
    print(data)

    # Test Q2
    plot_data(data)

    # Test Q3
    data=getFullData("finding_dory")
    print(len(data), data[-1])
    plot_data(data)
```