Homework 8

Name: Catherine Squillante

PID: cat1997

a) IMBD.com, 2019 movies list

b) The list contained 66 different movies, each with numerous reviews. I scraped around 25 reviews from each movie. The site was structured so that each title on the main list page had a link that movie's page, and then that page had a link to customer reviews. I made a list of the urls to each movie's page, and then for each of those navigated to the reviews page. I then gathered all 25 of each attribute at once and stored them in lists with list comprehensions. I used the list indexes to divide each review into its own list of all of the attributes. The name stayed the same for all the reviews on each page so I had a variable outside of the list comprehensions keeping track of the name index so it would know when to go to the next name.

c) I minimized my impact by using time.sleep for 10 seconds in between page requests for the separate movies, and for one second between the first and second link for each movie

d) The imbd site was a lot easier than yelp since there were specific tags for most of the attributes, so there weren't too many challenges. One problem was that the helpful ratings were given in the format of 'x out of y people found this helpful' so I converted them to percentages in order to standadize them. The few with no ratings I didn't count by setting them to NaN. Also, some of the numbers had commas in them so I had to make a function to remove them and convert the strings to floats before the percentages could be calculated.

Get inital movie list page from requests

In [1]:

```
import requests
import time
url = 'https://www.imdb.com/list/ls041578836/?sort=date_added,desc&st_dt=&mode=deta
il&page=1'
header = {'user-agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKi
t/537.36 (KHTML, like Gecko) Chrome/77.0.3865.90 Safari/537.36'}
page = requests.get(url, headers=header)
page.reason
```

Out[1]:

```
'OK'
```

Find movie names and links to individual movie pages

In [2]:

```
import bs4
import numpy
og_soup = bs4.BeautifulSoup(page.text, 'html5lib') #initial page of movie name links
tags = [t.find('a') for t in og_soup.find_all('h3', {'class':'lister-item-header'})] #section w/ name and href
#extract name and href
pre_hrefs = [tag['href'] for tag in tags]
names = [t.text for t in tags]
rows = [] #list to hold all of the reviews' lists
```

Navigate to reviews page from movie page for each movie and extract review data

In [3]:

```
name_iter = 0 #var to keep track of name since it's outside main loop
for pre_href in pre_hrefs:
    #navigate to the reviews page from the title link
    time.sleep(10)
    page1 = requests.get('https://www.imdb.com' + pre_href, headers = header)

    #find link to user reviews
    soup1 = bs4.BeautifulSoup(page1.text, 'html5lib')
    href = soup1.find('div', {'class':'user-comments'}).find_all('a')[-1]['href'] #
actual href to review
    time.sleep(1)
    page2 = requests.get('https://www.imdb.com' + href, headers = header)

    #reviews page
    soup = bs4.BeautifulSoup(page2.text, 'html5lib')

    #get all attributes
    rating = [int(t.find('span').text) for t in soup.find_all('span', {'class':'rat
ing-other-user-rating'})]
    author = [t.contents[0].text for t in soup.find_all('span',{'class':'display-na
me-link'})]
    text = [t.contents[1].text for t in soup.find_all('div', {'class':'content'})]
    date = [t.text for t in soup.find_all('span',{'class':'review-date'})]
    helpful = [t.text.split('\n')[1].strip() for t in soup.find_all('div', {'class'
:'actions text-muted'})]

    #helpful is given in format '109 out of 129 found this helpful'-> find percent
    def to_float(h): #account for numbers with ',' when converting to float
        if ',' in h:
            splitt = h.split(',')
            new_h = splitt[0]+splitt[1]
            return float(new_h)
        else:
            return float(h)

    helpful_tups = [(to_float(h.split()[0]), to_float(h.split()[3])) for h in helpf
ul] #tuple holding
    helpful_percent = []
    for h in helpful_tups:
        if h[0] == 0 and h[1] == 0:
            helpful_percent.append(numpy.nan) #na if both are 0
        else:
            helpful_percent.append(h[0]/h[1])

    ##add all data to lists
    for i in range(len(rating)):
        curr_row = [names[name_iter], href, rating[i], author[i], text[i], date[i],
helpful_percent[i]]
        rows.append(curr_row)
    name_iter = name_iter+1
```

Convert list data into frame

In [4]:

```
import pandas
df = pandas.DataFrame(rows, columns = ['Title', 'Url', 'Rating', 'Author', 'Review', 'Date', 'Helpful %'])
export_csv = df.to_csv (r'/Users/catsquillante/Documents/Documents/Data Visualization/Homework08.csv', index = None, header=True)
```

In [5]:

```
df
```

Out[5]:

	Title	Url	Rating	Author	Review	Date	I
0	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	9	davidgiorgione	That was INTENSE. Decided to watch this after ...	27 December 2019	0.8
1	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	8	marcus-191-212062	All reviews here are correct. The good ones an...	31 January 2020	0.8
2	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	10	kjproulx	It has been a fascinating ride watching the nu...	15 September 2019	0.
3	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	10	stefanruby	Saw this at TIFF. From front to back, this mov...	16 September 2019	0.
4	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	8	gortx	UNCUT GEMS (2019). First things first. This is...	23 December 2019	0.
5	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	9	Quinoa1984	They should hand out high-grade blood pressure...	28 December 2019	0.
6	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	iullah	I am pretty sure this would be an Oscar nod fo...	21 December 2019	0.
7	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	4	lydiataylor-52314	The plot largely consists of Adam Sadler yelli...	9 February 2020	0.
8	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	TheTopDawgCritic	How is this even a movie? For starters, who's ...	17 December 2019	0.5
9	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	3	jcodo	Wtf is this? Why the high ratings?\nIt's 135 m...	28 December 2019	0.5
10	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	2	nwsurfrider	I went into this film with fairly high hopes f...	26 December 2019	0.5

	Title	Url	Rating	Author	Review	Date	I
11	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	3	lauradys-93661	I feel compelled to write a note due the bewil...	20 January 2020	0.
12	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	4	farshad-persianguy	The only interesting thing about this movie wa...	3 January 2020	0.
13	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	3	clarkchris-48748	Uncut Gems was every bit as thrilling and chao...	21 January 2020	0.
14	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	helenroylesjones	This film is essentially 2 hours of people sho...	28 March 2020	0.
15	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	teresarkt	I have only walked out of one other movie in m...	21 January 2020	0.
16	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	2	gino-agostinelli	I couldn't even make it all the way through. T...	18 March 2020	0.
17	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	WHalawa	Yelling is not acting or my kid is Al Pacino, ...	23 December 2019	0.
18	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	steveevans-35154	Two hours of people just shouting. It's absolu...	18 January 2020	0.
19	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	1	michpetty	Wish I had left immediately but I was with som...	18 January 2020	0.
20	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	4	ciro_ciampi	I go with the honest & truthful reviews posted...	1 February 2020	0.
21	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	9	theironman8	The music was terrible. The constant yelling a...	28 December 2019	0.
22	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	5	RinoBortone91	A frantic episode in the life of Howard Ratner...	25 December 2019	0.
23	Uncut Gems	/title/tt5727208/reviews?ref_=tt_urv	2	Kreig303	I wanted to like this movie. I really did. But...	14 December 2019	0.

	Title	Url	Rating	Author	Review	Date	
24	1917	/title/tt8579674/reviews?ref_=tt_urv	10	dr-peter-coldwell	Last night COL Ferry and I (COL Coldwell, both...	13 December 2019	0.8
25	1917	/title/tt8579674/reviews?ref_=tt_urv	10	tgrafflin	I sat in a packed yet silent theater this morn...	5 January 2020	0.7
26	1917	/title/tt8579674/reviews?ref_=tt_urv	10	gkdidaxi	This film is overwhelming. I have nothing furt...	4 January 2020	0.7
27	1917	/title/tt8579674/reviews?ref_=tt_urv	10	allentyson-89230	Don't listen to the critics saying this movie ...	10 January 2020	0.7
28	1917	/title/tt8579674/reviews?ref_=tt_urv	10	diegosays	1917 is a poem.\nIs the most deep, impressive ...	7 January 2020	0.7
29	1917	/title/tt8579674/reviews?ref_=tt_urv	10	frederic-22	Guaranteed Oscar. A technical and visual trium...	13 December 2019	0.7
...	...	...	...	...	...	...	
2382	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	10	ethan-33027	It's amazing!! Go watch it now!Honestly, it's ...	22 December 2018	0.5
2383	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	10	rannynm	What would you do if you could be a superhero?...	12 December 2018	0.5
2384	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	9	vincenttciccarello	Amazing movie. Great animation, the best Spide...	20 December 2018	0.5
2385	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	10	reesepaul	I was very lucky as my son and I just got to s...	6 December 2018	0.5
2386	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	7	educallejero	The movie is visually spectacular (with the ex...	25 February 2019	0.6

	Title	Url	Rating	Author	Review	Date	
2387	Spider-Man: Into the Spider-Verse	/title/tt4633694/reviews?ref_=tt_urv	10	chriscarlisle25	I knew that this movie would be good from the ...	15 December 2018	0.
2388	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	7	Bertaut	The Favourite, the seventh feature from Greek ...	20 January 2019	0.
2389	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	8	roblesar99	Sumptuous and stunning. With THE FAVOURITE, di...	22 September 2018	0.
2390	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	9	FrenchEddieFelson	In early 18th century, the friendship between ...	20 February 2019	0.
2391	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	chrisbrady1	I read the rave reviews before the film opened...	26 December 2018	0.
2392	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	4	ccrisss	I was excited to watch this but found myself a...	7 January 2019	0.
2393	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	2	andrewestrella	Let me preface my review with me saying that I...	14 December 2018	0.
2394	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	2	babybenz-41229	I'm sorry, but I don't agree with the high rat...	20 December 2018	0.
2395	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	lnedwards	Bizarre. Not funny. Waste of a plotline with 3...	16 December 2018	0.
2396	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	8	howard-85268	I was invited to see this film by my wife and ...	2 January 2019	0.
2397	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	andrewroy-04316	While Lanthimos continues his eccentric and ab...	23 February 2019	0.
2398	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	10	rsmithsd	I literately signed up with IMDB to share was ...	16 December 2018	0.

	Title	Url	Rating	Author	Review	Date	I
2399	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	10	pjdesmedt-99161	Whenever a movie splits an audience into 10 an...	10 January 2019	0.5
2400	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	2	andrebacci-88902	Enthralling from the very beginning and bursti...	6 November 2018	0.5
2401	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	2	thomrobbin-1	I am astounded that this film has received so ...	29 December 2018	0.5
2402	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	8	princessrb	Nothing to spoil here. Two hours of agony (wit...	2 December 2018	0.5
2403	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	3	sillysillymarym	Interesting thing about this movie, is that th...	30 November 2018	0.5
2404	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	brianjohnson-20043	I was expecting better. The film has some bone...	10 January 2019	0.5
2405	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	andypratt-39170	Where did they find (or hire) so many people t...	14 December 2018	0.5
2406	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	9	rnjc3	Not just me and my girl but several people voc...	9 December 2018	0.5
2407	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	8	KarenAM	The Favourite is, so far, the best movie I've ...	3 October 2018	0.5
2408	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	Gresh854	The Favourite was not what I expected. This is...	14 October 2018	0.5
2409	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	9	maribs1971	Do not waste your time watching this movie! As...	29 December 2018	0.5
2410	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	shawthingproductions	This is not my usual type of film but the trai...	1 January 2019	0.5
2411	The Favourite	/title/tt5083738/reviews?ref_=tt_urv	1	enterprise77	The trailer portrays this movie as a comedy. N...	26 December 2018	0.5

2412 rows × 7 columns

```
In [ ]:
```