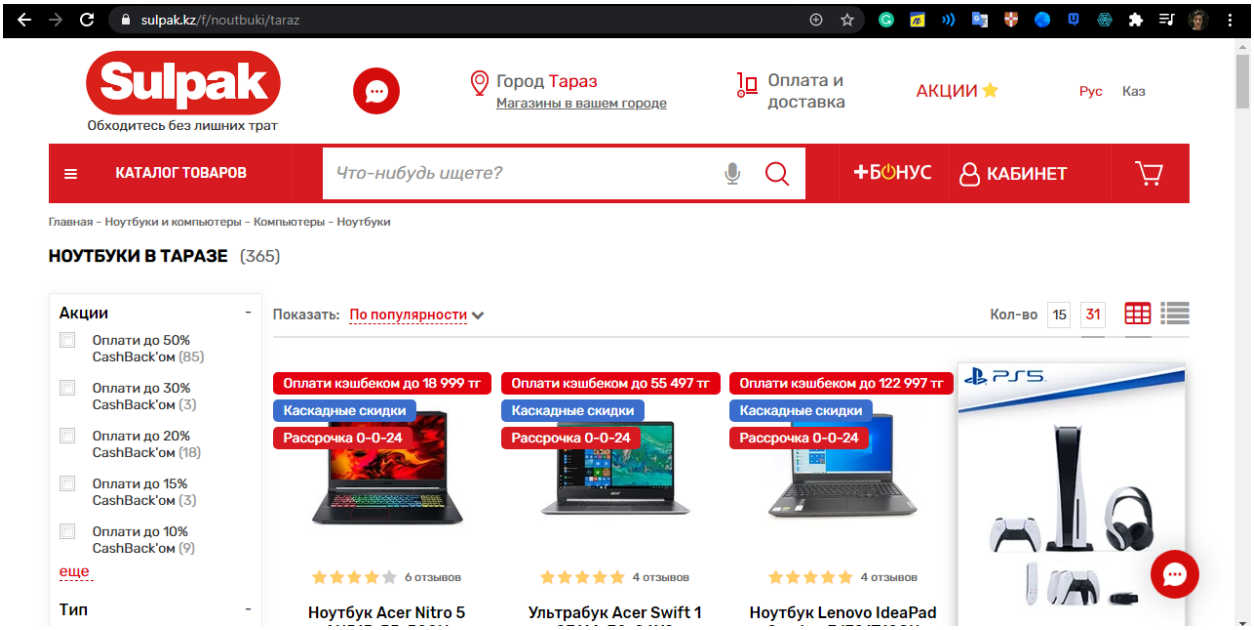
Ali Shalbayev

Home Work 3

Report

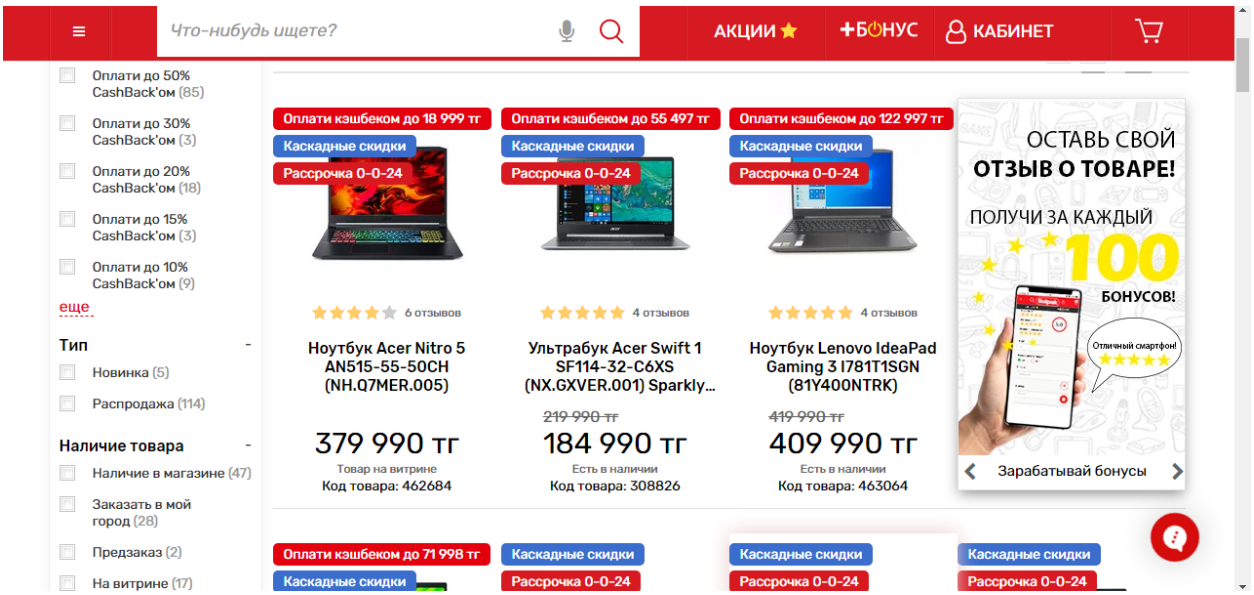
For the web scraping task, I chose Sulpak.kz as my target website. Particularly I decided to scrape the laptops page.



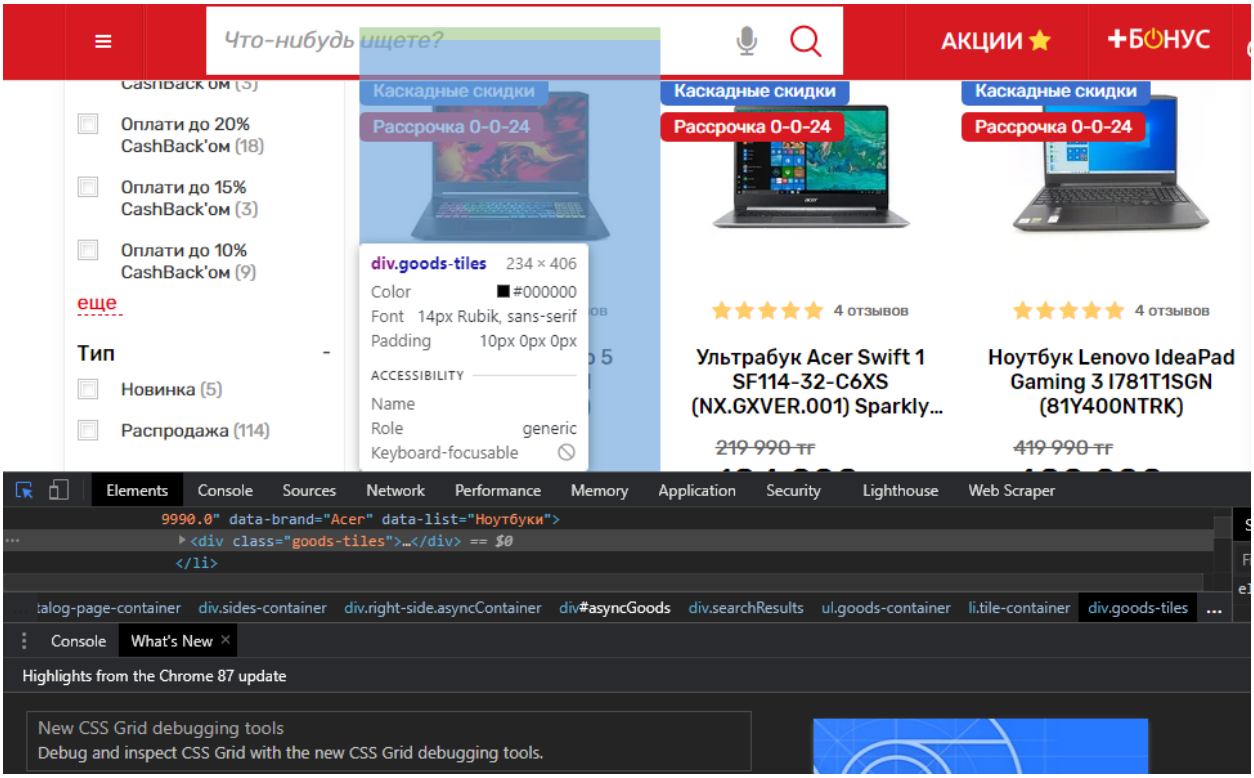
Here I am importing the libraries for the beginning.

```
In [1]: import requests
        from bs4 import BeautifulSoup

In [2]: mainurl = 'https://www.sulpak.kz'
        headers = {'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88 S
```

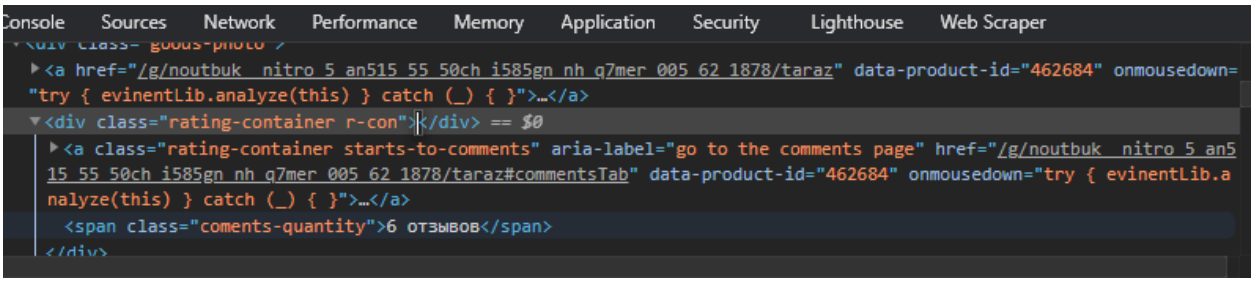
My first task was to gather the links for each of these laptops.



By hotkey ctrl+shift+i I identified the parent tag



The next task was to determine where links are located.



Then I wrote the loop for collecting these links since there were 3 links per each product, I was forced to get rid of them.

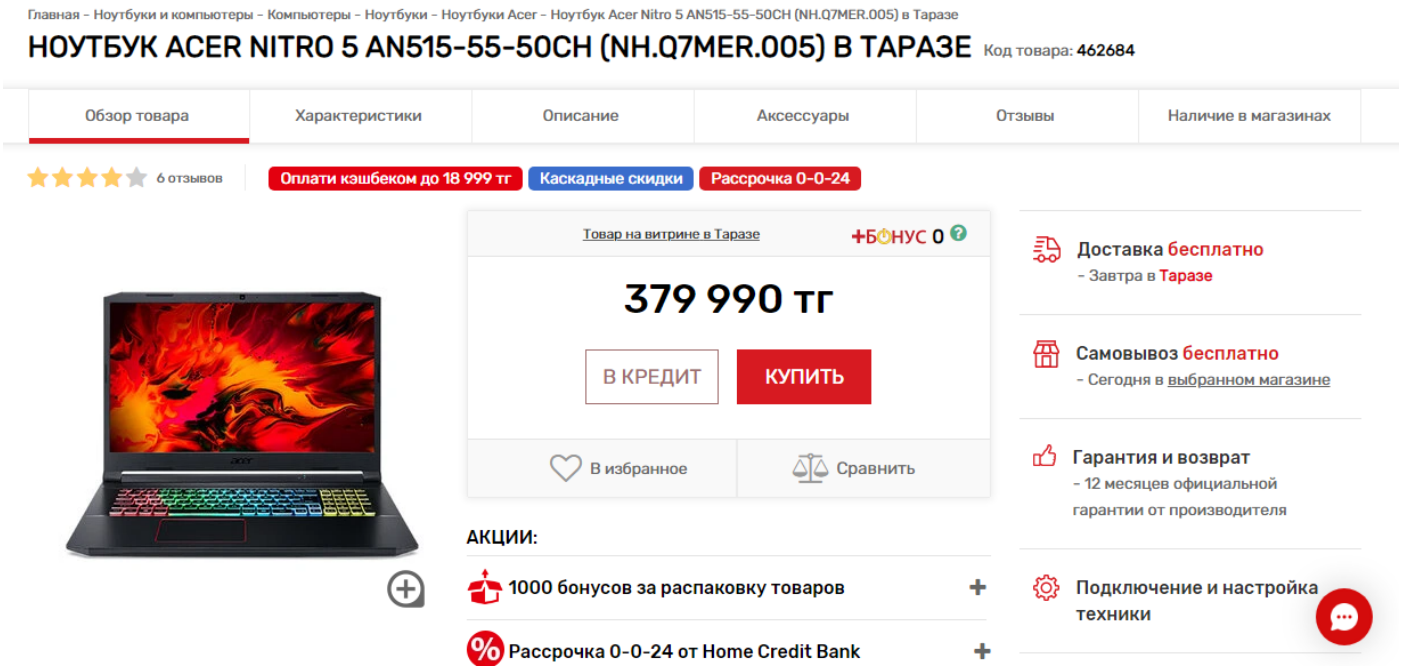
```
In [3]: productlinks = []
```

```
In [4]: for x in range(1, 13):
            r = requests.get(f'https://www.sulpak.kz/f/noutbuki/taraz?page={x}')
            soup = BeautifulSoup(r.content, 'lxml')
            productlist = soup.find_all('div', class_='goods-tiles')
            for product in productlist:
                for link in product.find_all('a', href = True):
                    productlinks.append(mainurl + link['href'])
                    productlinks = set(productlinks)
                    productlinks = [ x for x in productlinks if "#commentsTab" not in x ]
```

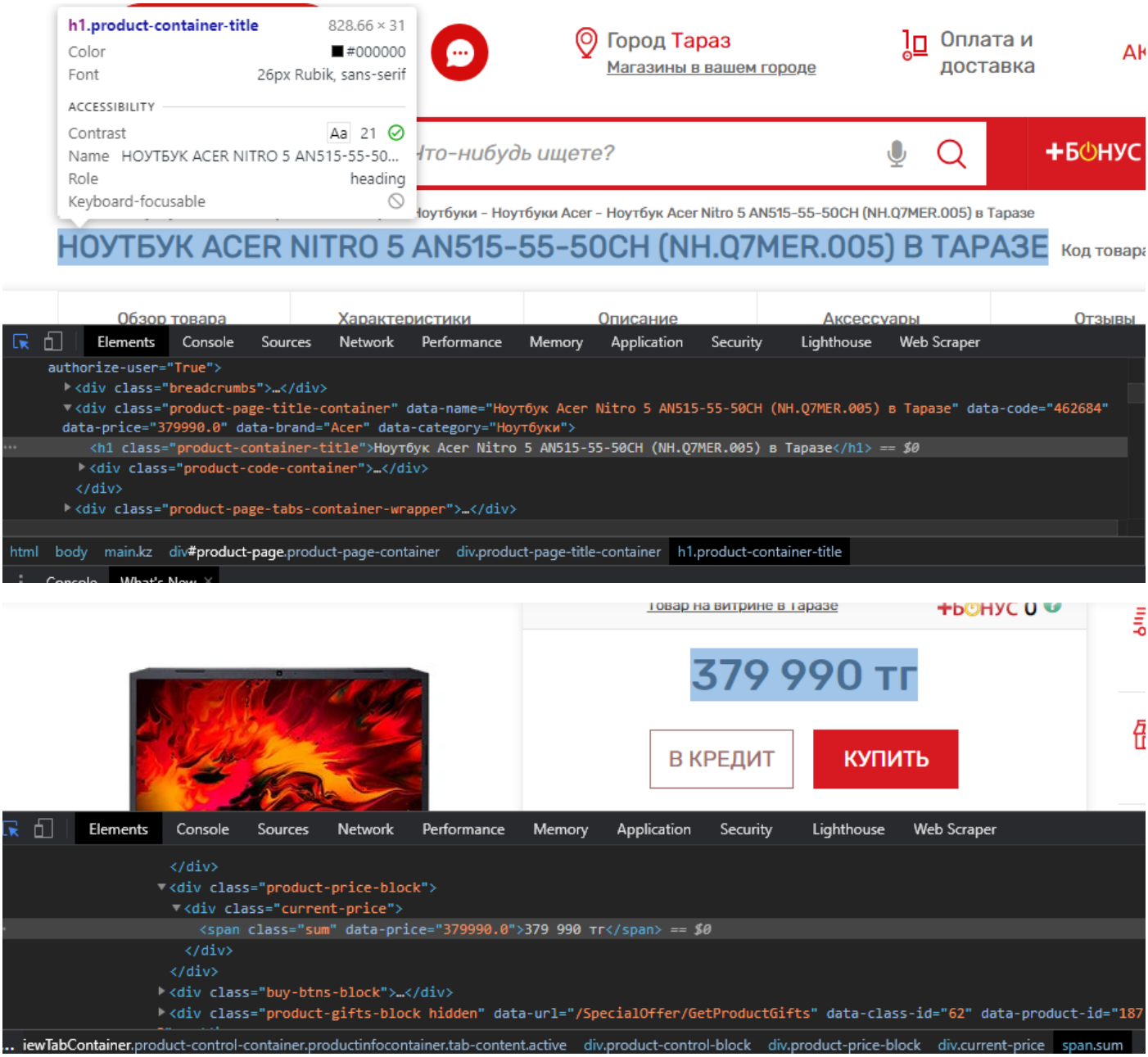
There I am checking is everything fine.

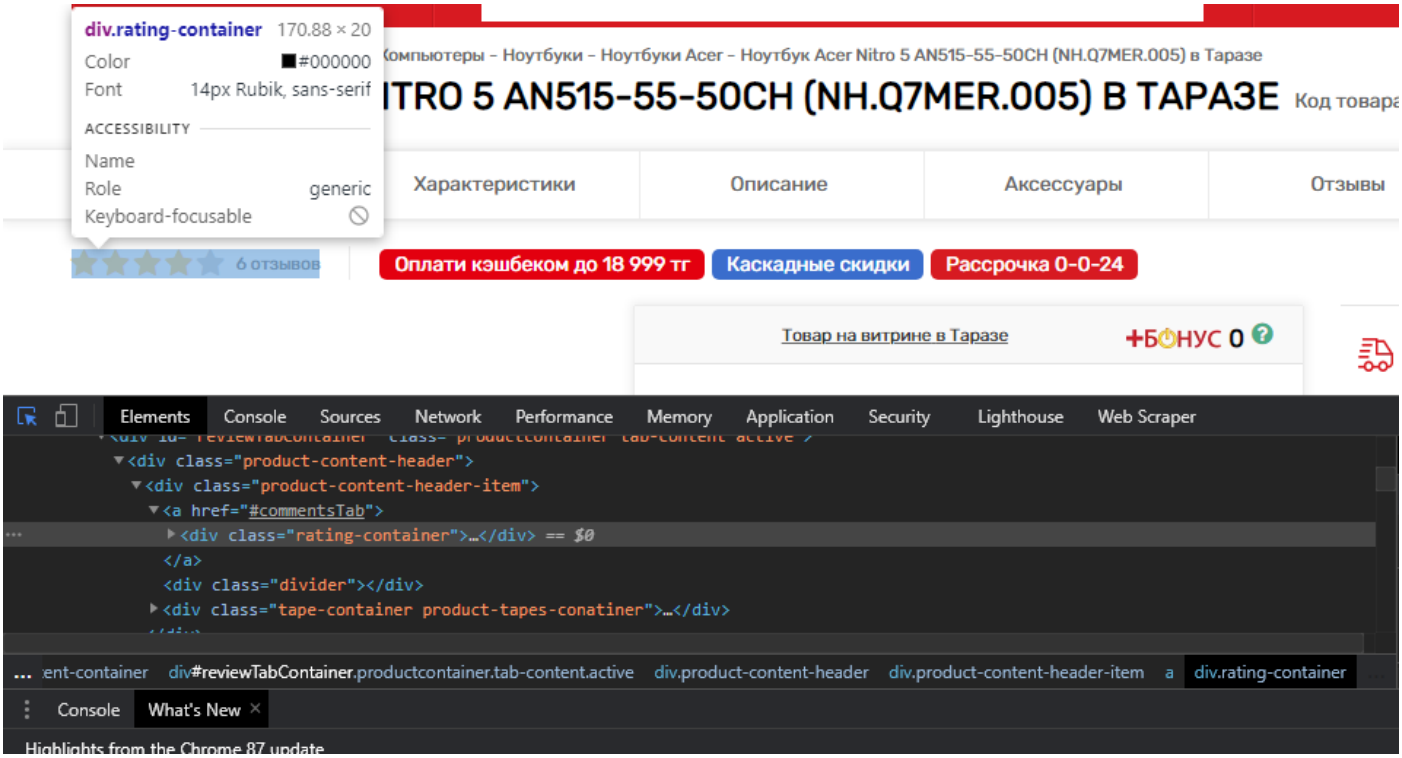
```
In [5]: print(len(productlinks))
341

In [6]: for x in productlinks:
            print(x)
https://www.sulpak.kz/g/ultrabuk_apple_macbook_12_mrqn2_62_1530
https://www.sulpak.kz/g/noutbuk_apple_macbook_pro_13_m1_3_28gb512gb_ssd_space_gray_myd92_62_2074
https://www.sulpak.kz/g/noutbuk_acer_aspire_3_a315_55_nx_heher_01g_62_1722
https://www.sulpak.kz/g/noutbuk_acer_aspire_5_a515_54g_767q_i781t1smn_silver_nx_hn5er_004_62_1932
https://www.sulpak.kz/g/noutbuk_lenovo_ideapad_s145_i341tuw_lenpn4_62_1914
https://www.sulpak.kz/g/noutbuk__nitro_5_an515_44_r67f_r585gn_nh_q9ger_008_62_1883
https://www.sulpak.kz/g/ultrabuk_apple_macbook_pro_13_touch_bar_silver_mv9a2rua_62_1531
https://www.sulpak.kz/g/noutbuk_hp_elitebook_840_g6_i585smw_8mj69ea_62_2061
https://www.sulpak.kz/g/noutbuk_asus_rog_strix_scar_17_g733qs_90nr0591_m02010_62_2088
https://www.sulpak.kz/g/noutbuk_acer_aspire_3_a315_42_r381tun_nx_hf9er_02k_62_1852
https://www.sulpak.kz/g/noutbuk_lenovo_ideapad_s145_81ut00gcrk_62_1792
https://www.sulpak.kz/g/noutbuk_xiaomi_mi_notebook_air_125_silver_jyu4047cn
https://www.sulpak.kz/g/noutbuk_hp_elitebook_830_g7_i585suw_176z1ea_62_2057
https://www.sulpak.kz/g/noutbuk_hewlett_packard_pavilion_gaming_15_dk0036ur_i585gn_7qa10ea_62_1827
https://www.sulpak.kz/g/noutbuk_asus_tuf_gaming_fx505dt_al097_r585gn_62_1749
https://www.sulpak.kz/g/noutbuk_apple_macbook_12_dyujmov_mlh72_space_gray
https://www.sulpak.kz/g/noutbuk_asus_fx505d_90nr02n2_m03000_62_1778
https://www.sulpak.kz/g/noutbuk_lenovo_ideapad_3_330_17ikb_i341tmw_81dk0088rk_62_1950
https://www.sulpak.kz/g/noutbuk_apple_macbook_air_2020_gold_i385ux_mvh52_62_1818
https://www.sulpak.kz/g/noutbuk_acer_extensa_ex215_51k_i585un_nx_efner_017_62_1764
```

The next step was to detect where name, price and reviews are laid out.

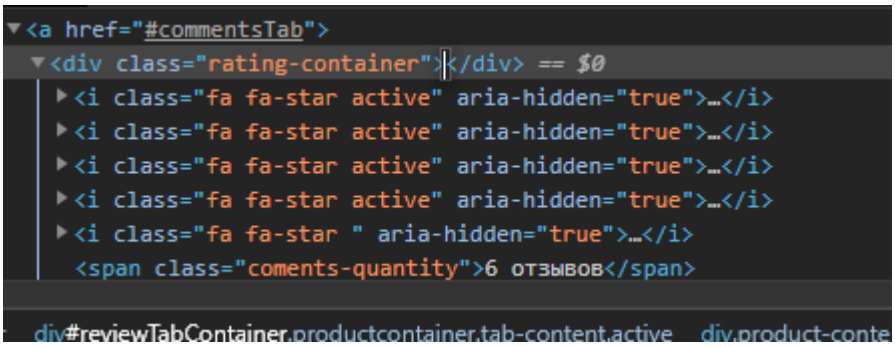


Again using hotkey ctrl+shift+i I started inspecting the page and found out where they are.





To identify how many stars a product gained I counted how many times a string "fa fa-star active" did appear.



So here is my code for all these steps.

In [50]:
```
#testlink = 'https://www.sulpak.kz/g/noutbuk_acer_aspire_3_a315_34_c41tuw_nx_he3er_011_62_1982/'
laptoplist = []
for link in productlinks:
    r = requests.get(link, headers = headers)
    soup = BeautifulSoup(r.content, 'lxml')
    name = soup.find('h1', class_='product-container-title').text.strip()
    try:
        price = soup.find('span', class_='sum').text.strip()
    except:
        price = 'Sold Out'
    try:
        review = soup.find('div', class_='rating-container').text.strip()
    except:
        review = 'No Reviews'
    try:
        rating = soup.find('div', class_='rating-container')
        rating = str(rating)
        star = rating.count('fa fa-star active')
    except:
        star = 'No Ratings'

    laptop = {
        'Name' : name,
        'Price' : price,
        'Reviews' : review,
        'Rating out of 5' : star
    }

    laptoplist.append(laptop)
```

This is a scraped data frame that we finally got.

In [11]:
```
import pandas as pd
```

In [51]:
```
df = pd.DataFrame(laptoplist)
df
```

Out[51]:

	Name	Price	Reviews	Rating out of 5
0	Ультрабук Apple MacBook 12" Core M3/8GB/256GB ...	383 790 тг	2 отзывов	5
1	Ноутбук Apple MacBook Pro 13" M1 3.2/8Gb/512GB...	825 990 тг	0 отзывов	0
2	Ноутбук Acer Aspire 3 A315-55 (NX.HEHER.01G)	189 990 тг	1 отзывов	1
3	Ноутбук Acer Aspire 5 A515-54G-767Q Silver (NX...	349 990 тг	0 отзывов	0
4	Ноутбук Lenovo IdeaPad S145 I341TUW (81W800KNRK)	219 990 тг	0 отзывов	0
...	...	...	...	...
336	Ультрабук Acer Swift 1 SF114-32 (NX.GXUER.00E)	184 990 тг	0 отзывов	0
337	Ноутбук Acer Predator Helios 300 PH317-54-76VC...	979 990 тг	0 отзывов	0
338	Ноутбук Asus VivoBook 15 X512D R341TUW (90NB0L...	149 990 тг	0 отзывов	0
339	Ноутбук ASUS Laptop 15 M509D R381TUN (90NB0P51...	179 990 тг	0 отзывов	0
340	Ноутбук Asus Rog Strix G512L (90NR0351-M04890)	599 990 тг	0 отзывов	0

341 rows × 4 columns