

1619313040/3301 (Data Science Oriented Programming Language-Python)

SEMESTER 2 (Spring), 2021/2022

LABORATORY WORK TWO

NAME: MD Omar Faruq

STUDENT NO.: 201939060073

SUBMITTED TO:

Assoc. Prof. FANG YU

MARKING SCHEME (TO BE FILLED BY THE LECTURER)

|  |  |
| --- | --- |
| **CRITERIA** | **MARKS** |
| CONTENTS | /6 |
| ORGANIZATION/STRUCTURE | /2 |
| WRITING MECHANICS | /2 |
| **TOTAL** | **/10** |

**DATA ACQUISITION PRACTICE**

In this section, students must display originality of their writings by jotting down their understanding of crawling. Students need to fully describe and discuss in terms of:

1. **The structure analysis for targeted html page, and flowcharts for your python codes**
2. **Screenshot your running results (better applying video clips), upload your source codes and your lab work report.**

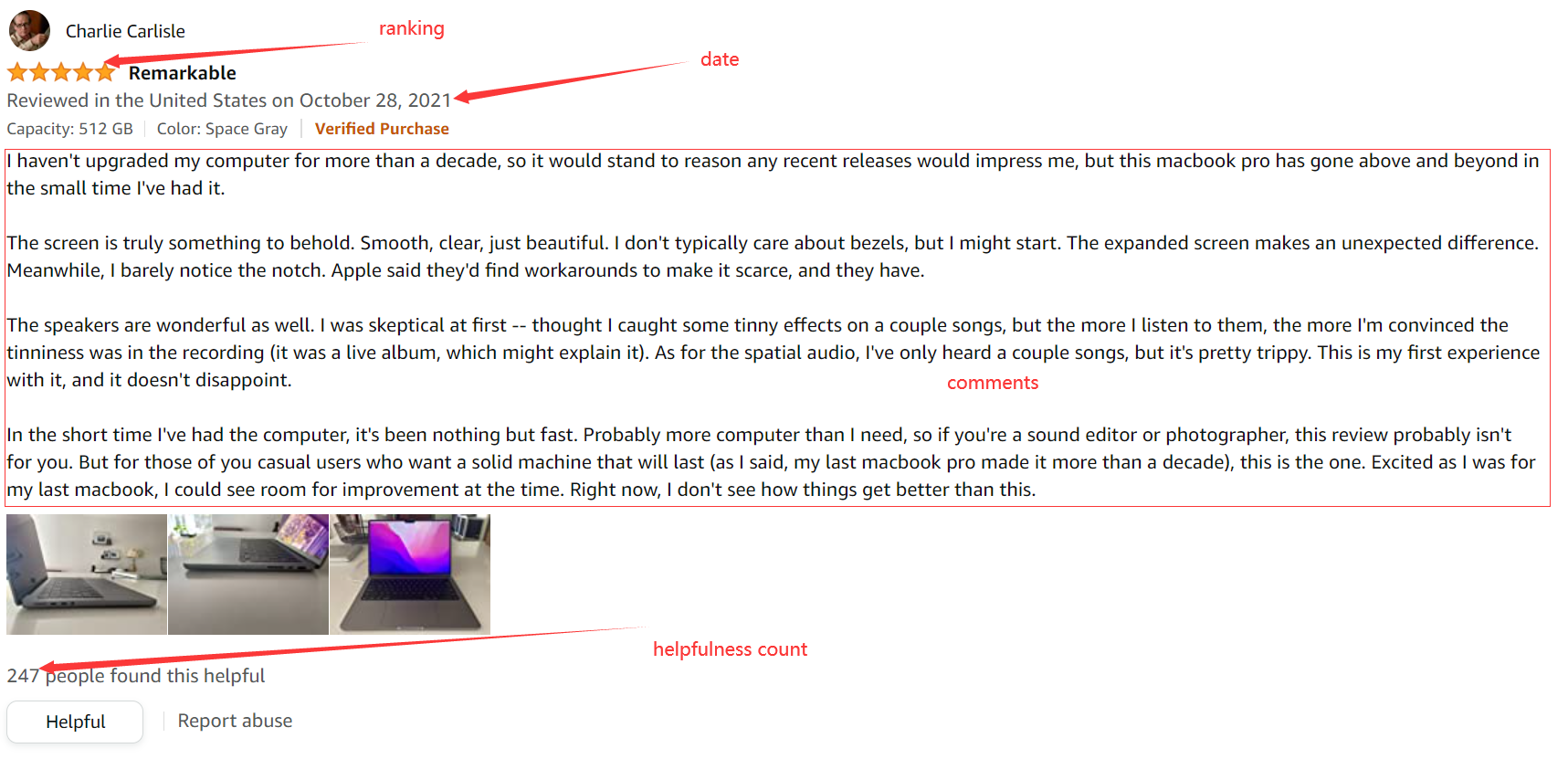
*FOLLOWING IS THE LABWORKS (INCLUDING 3 VERSIONS OF THE GAME)*

**Description about the task:**

1. Obtain all the comments (currently 524 entries) from following link:

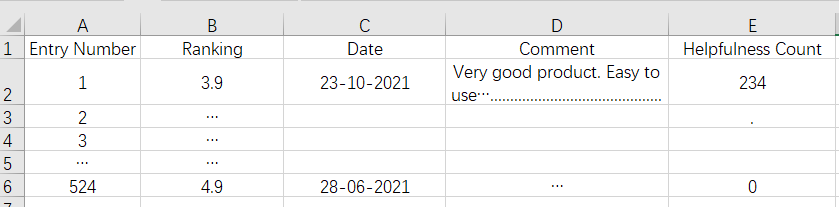
[Amazon.com: Customer reviews: 2021 Apple MacBook Pro (14-inch, Apple M1 Pro chip with 8‑core CPU and 14‑core GPU, 16GB RAM, 512GB SSD) - Space Gray](https://www.amazon.com/Apple-MacBook-14-inch-8%E2%80%91core-14%E2%80%91core/product-reviews/B09JQSLL92/ref=cm_cr_dp_d_show_all_btm?ie=UTF8&reviewerType=all_reviews)

Including: ranking, date, comment details, helpfulness count.



1. There is no limitation for crawling technique, you can use any crawlers you preferred.
2. Store all information into a CSV or Excel file.

The headers should be following figure:



1. Make a summary about this task.

This is an interesting task. I am using Selenium and BeautifulShup package to collect information from the given web page. The task is about collecting comments from an Amazon product. For this product, there are 566 total ratings, 105 with reviews until now. So, I have to collect 105 comments. Firstly, I insert Chrome web driver. Then I put the website link to driver. After that I use BeautifulShup package to find html elements. Then I collect all the information and put into in an exel file by using openpyxl package. I have learned mane new things from the task.

**Source Code**

from selenium import webdriver  
from selenium.webdriver.common.by import By  
from time import sleep  
from bs4 import BeautifulSoup  
import re, openpyxl  
  
# Creating Exel file  
exel = openpyxl.Workbook()  
sheet = exel.active  
sheet.title = "Amazon Review"  
sheet.append(['Entry Number', 'Ranking', 'Date', 'Comment', 'Helpfulness Count'])  
  
web = webdriver.Chrome()  
web.maximize\_window()  
sleep(1)  
  
web.get('https://www.amazon.com/Apple-MacBook-14-inch-8%E2%80%91core-14%E2%80%91core/product-reviews/B09JQSLL92/ref=cm\_cr\_dp\_d\_show\_all\_btm?ie=UTF8&reviewerType=all\_reviews')  
  
sleep(2)  
page\_count = 0  
comment\_count = 0  
  
for i in range(11):  
 try:  
 page\_count += 1  
 print(f'[ Page: {page\_count} ]')  
  
 sleep(3)  
 pageSource = web.page\_source  
 soup = BeautifulSoup(pageSource, 'html.parser')  
 reviews = soup.find('div', id="cm\_cr-review\_list").find\_all('div', class\_='a-section celwidget')  
  
 for review in reviews:  
 comment\_count += 1  
  
 # collect all ratings  
 rating = review.find('span', class\_='a-icon-alt').text  
 f\_rating = rating[:3]  
 print(f"Rating: {f\_rating}")  
  
 # collect all dates  
 date = review.find('span', class\_='a-size-base a-color-secondary review-date').text  
 f\_date = date[33:]  
 print(f"Date: {f\_date}")  
  
 # collect all review contents  
 message = review.find('div', class\_='a-row a-spacing-small review-data').text  
 f\_message = message.strip()  
 print(f"Message: {f\_message}")  
  
 # collect all helpfull counts  
 if review.find('span', class\_='a-size-base a-color-tertiary cr-vote-text'):  
 helpfull\_count = review.find('span', class\_='a-size-base a-color-tertiary cr-vote-text').text  
 if helpfull\_count[:3] == 'One':  
 f\_helpfull\_count = 1  
 else:  
 f\_helpfull\_count = re.sub('[^0-9]', '', helpfull\_count)  
 else:  
 f\_helpfull\_count = 0  
 print(f"Helpfull count: {f\_helpfull\_count}")  
  
 # add information to the exel sheet  
 sheet.append([comment\_count, f\_rating, f\_date, f\_message, f\_helpfull\_count])  
  
 print("\n")  
 print("\n\n--------------------------------\n\n")  
  
 # click next button  
 web.find\_element(by=By.CLASS\_NAME, value="a-last").click()  
  
 except:  
 continue  
  
# save the exel file  
exel.save('Amazon Review.xlsx')  
web.close()

**Running result**

