DIGITAL ASSIGNMENT - 5

MAY 13, 2021

RISHABH SHARMA
20MAI0082

Github Link :- https://github.com/rishabh5197/Data-Mining/tree/main/Assignment-5

Name :- Rishabh Sharma

Registration Number :- 20MAI0082

Assignment - 5

Web Mining

Importing Libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from bs4 import BeautifulSoup as bs
import requests
from warnings import filterwarnings
filterwarnings("ignore")
```

Using Web Scraping to collect the data

In [2]:

```
while True:
    try:
        link = 'https://www.flipkart.com/search?q=laptops&otracker=search&otracker1=se
        print(link)
        page = requests.get(link)
        print(page)
        break
    except:
        pass
```

```
https://www.flipkart.com/search?q=laptops&otracker=search&otracker1=search&marketpla
ce=FLIPKART&as-show=on&as=off (https://www.flipkart.com/search?q=laptops&otracker=s
earch&otracker1=search&marketplace=FLIPKART&as-show=on&as=off)
<Response [200]>
```

In [3]:

```
soup = bs(page.content,'html.parser')
```

In [4]:

```
# items = soup.find_all('div',class_="_4rR01T")
# items=[i.get_text() for i in items]
# items = [i.split(" - (")[0] for i in items]
```

In [5]:

```
details = soup.find_all("li",class_="rgWa7D")
details = [i.get_text() for i in details]
```

In [6]:

```
lis,newlis=[],[]
for i in details:
    if ("Intel ")in i or ("AMD ") in i or ("M1") in i :
        if newlis:
            lis.append(newlis)
            newlis=[]
            newlis.append(i)
        else:
            newlis.append(i)
    else:
        newlis.append(i)
newlisnew=[]
for i in lis:
    newlisnew.append(" ".join(i))
```

In [7]:

```
items = []
for i in newlisnew:
    if ' GB DDR4' in i:
         items.append(i.split(" GB DDR4")[0][:-2])
    elif ' GB DDR3' in i:
        items.append(i.split(" GB DDR3")[0][:-2] )
    elif ' GB LPDDR4X' in i:
        items.append(i.split(" GB LPDDR4X")[0][:-2] )
    else:
        items.append("")
company = [i.split()[0] for i in items]
```

In [8]:

```
pages= soup.find_all('a',class_='ge-49M',href=True)
pages = [str(i).split(">")[0][31:] for i in pages ]
```

In [9]:

```
ratings = soup.find_all("div",class_='_3LWZlK')
ratings = [i.get_text() for i in ratings]
```

In [10]:

```
prices =  soup.find_all("div",class_="_30jeq3 _1_WHN1")
prices = [i.get_text().replace("₹","").replace(",","") for i in prices]
```

In [11]:

```
# details = soup.find_all("li",class_="rgWa7D")
# details = [i.get_text() for i in details]
```

In [12]:

```
# indepth_link = soup.find_all("a",class_ = '_1fQZEK')
```

In [13]:

```
# start= str(indepth_link[0]).index("href=")
# end = str(indepth_link[0]).index(" rel=")
# links = ['https://www.flipkart.com'+str(str(i)[start+6:end-1]) for i in indepth_link]
```

In [14]:

```
lis,newlis=[],[]
for i in details:
    if ("Intel ")in i or ("AMD ") in i or ("M1") in i :
        if newlis:
            lis.append(newlis)
            newlis=[]
            newlis.append(i)
        else:
            newlis.append(i)
    else:
        newlis.append(i)
newlisnew=[]
for i in lis:
    newlisnew.append(" ".join(i))
```

Collecting data and filtering everything

In [15]:

```
processor_brand=[]
processor = []
ram = []
ram_type=[]
os =[]
screen_size=[]
ssd_present=[]
ssd_capacity=[]
hdd_capacity=[]
# count=0
for i in newlisnew:
    if "Intel " in i:
#           print(count)
        processor_brand.append("Intel")
        processor.append(i.split()[1]+" "+i.split()[2])
    elif 'AMD' in i:
#           print(count)
        processor_brand.append("AMD")
        processor.append(i.split()[1]+" "+i.split()[2])
    elif "M1" in i:
#           print(count)
        processor_brand.append("M1")
        processor.append(i.split()[1]+" "+i.split()[2])
    else:
        processor_brand.append("")
        processor.append("")
#       count+=1
    if ' GB DDR4 RAM' in i:
        index = i.index(' GB DDR4 RAM')
        ram.append(i[index-2:index])
        ram_type.append("DDR4")
    elif  ' GB DDR3' in i:
        index = i.index(' GB DDR3 RAM')
        ram.append(i[index-2:index])
        ram_type.append("DDR3")
    else:
        ram.append("")
    if ' Operating System' in i:
        a = i.split()
        index = a.index("Operating")
        if a[index-1]=="10":
                os.append("Windows 10")
        elif a[index-2]=='Mac':
            os.append("Mac Os")
        else:
            os.append(a[index-1])
    else:
        os.append("")
    if "inch" in i:
        a = i.split()
        if 'inches)' in a:
            index = a.index("inches)")
        else:
            index = a.index("inch)")
        screen_size.append(a[index-1].strip("(")+" inch")
    if (" GB SSD" in i) or (" TB SSD" in i):
        ssd_present.append("Yes")
        if " GB SSD" in i:
            start = i.index(" GB SSD")-3
            end=i.index(" GB SSD")
            ssd_capacity.append(i[start:end])
        elif' TB SSD' in i:
            index = i.index(" TB SSD")
            ssd_capacity.append(int(i[index-1])*1024)
    else:
        ssd_present.append("No")
        ssd_capacity.append("")
    if " HDD" in i:
        index = i.index(" HDD")
        hdd_capacity.append(i[index-4])
    else:
        hdd_capacity.append("")
```

Creating a dataframe in order to carry out further steps

In [16]:

```
data = pd.DataFrame({
'items':items[:len(processor)],                              # Name of Laptop
'company':company[:len(processor)],                          # Company of laptop  (or Laptop
'ratings out of 5':ratings[:len(processor)],                  # What are the rating mentione
'prices':prices[:len(processor)],                            # Price of the laptop
'processor_brand':processor_brand,                           # what is the processor of the l
'processor':processor,                                       # Type of processor of the lapto
'ram':ram ,                                                  # how much ram does it have (Ca
'ram_type':ram_type,                                         # what is the type of ram (Categ
'operating_system' :os,                                      # consist of which operating sys
'screen_size':screen_size,                                   # Screen-size of the laptop (Ca
'ssd_present':ssd_present,                                   # is SSD present in the laptop (
'ssd_capacity':ssd_capacity,                                 # What is the capacity of the S
'hdd_capacity_in_TB':hdd_capacity,                            # What is the capacity of HDD
'Purchased': [np.random.choice(["No","Yes"]) for i in range(0,len(processor))]})  # this i
data.to_csv("Dataset.csv",index=False)                                              # St
```

In [17]:

```
data.shape
```

Out[17]:

```
(23, 14)
```

In [18]:

```
data =data.replace("",np.nan)
```

In [19]:
```
data.head()
```

Out[19]:

	items	company	ratings out of 5	prices	processor_brand	processor	ram	ram_type	operating_s
0	AMD Ryzen 5 Quad Core Processor (3rd Gen)	AMD	4.4	48990	AMD	Ryzen 5	8	DDR4	Windc
1	Intel Core i3 Processor (10th Gen)	Intel	4	35990	Intel	Core i3	8	DDR4	Windc
2	Intel Core i5 Processor (9th Gen)	Intel	4.5	52990	Intel	Core i5	8	DDR4	Windc
3	Intel Core i3 Processor (10th Gen)	Intel	4.2	33490	Intel	Core i3	4	DDR4	Windc
4	Intel Core i3 Processor (10th Gen)	Intel	4.2	35990	Intel	Core i3	8	DDR4	Windc

In [20]:
```
data.isnull().sum()
```

Out[20]:
```
items                0
company              0
ratings out of 5     0
prices               0
processor_brand      0
processor            0
ram                  0
ram_type             0
operating_system     0
screen_size          0
ssd_present          0
ssd_capacity         4
hdd_capacity_in_TB  14
Purchased            0
dtype: int64
```

Filling null values

In [21]:

```
data = data.replace(np.nan,0)
data.head()
```

Out[21]:

	items	company	ratings out of 5	prices	processor_brand	processor	ram	ram_type	operating_s
0	AMD Ryzen 5 Quad Core Processor (3rd Gen)	AMD	4.4	48990	AMD	Ryzen 5	8	DDR4	Windc
1	Intel Core i3 Processor (10th Gen)	Intel	4	35990	Intel	Core i3	8	DDR4	Windc
2	Intel Core i5 Processor (9th Gen)	Intel	4.5	52990	Intel	Core i5	8	DDR4	Windc
3	Intel Core i3 Processor (10th Gen)	Intel	4.2	33490	Intel	Core i3	4	DDR4	Windc
4	Intel Core i3 Processor (10th Gen)	Intel	4.2	35990	Intel	Core i3	8	DDR4	Windc