Modern Programming Language.

Midterm Project

Carrefour toilet paper price comparison

現代程式語言 . 期中報告

家樂福衛生紙價格比較

第八組:111210518 陳辰鈺、111210520 吳佳泰、111210524 藍健洲

Modern Programming Language

angucsie112

Crawling Target

def goal_01():

一開始我們想搜尋全聯小時達的衛生紙,但我們發現此網站是採用React框架,並且使用了Next.js來進行伺服器端渲染。這種設計使得網頁載入速度更快,網頁會動態地更新內容而不改變URL,這給我們的爬蟲之路上造成了很多曲折

於是,我們學會了適時轉彎

```
def goal_02():
```

後來我們決定搜尋家樂福線上購的衛生紙,決定要爬取這個網站的目標的原因是因為我們想要比較各家衛生紙在此網站上面的價格並將它繪製出來

找標籤好像在考古一樣XD

count = 2

```
## Preliminary Tasks for Crawling
```

```
爬取的前置作業
```

def process():

- 配置VScode環境
- 偽裝成正常的瀏覽器客戶端以利爬取資料

```
headers = {"User-Agent" : "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
```

AppleWebKit/537.36(KHTML, like Gecko) Chrome/80.0.3897.132

Safari/537.6"}

```
def module used():
                 ##讀取網頁檔案
 request
                 ##解析
 BeautifulSoup
                 ##資料庫創建與使用
 sqlite3
                 ##讀取資料庫之元素
 pandas
 matplotlib.pyplot ##繪圖
```

Crawling Steps

def step_01():

```
import requests
from bs4 import BeautifulSoup
import sqlite3
list1 = []
list2 = []
url = "https://online.carrefour.com.tw/zh/search/?q=%E6%8A%BD%E5%8F%96%20%E8%A1%9B%E7%94%9F%E7%B4%99"
headers = {"User-Agent" : "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36(KHTML, like Gecko) Chrome/80.0.3897.132 Safari/537.6"}
html = requests.get(url , headers=headers)
html.encoding = "UTF-8"
sp = BeautifulSoup(html.text , "lxml")
data1 = sp.find all("div" , class ="commodity-desc") ##name
data2 = sp.find all("em") ## price
                                                                                                                                                     Python
```

- def step_01():
 - 導入所需模組
 - 創建兩個空串列備用
 - 爬取並提取網頁內所需元素
 - 將衛生紙的名字存在data1
 - 將衛生紙的價格存在data2

def step_02():

```
def step_02():
```

def step 02():

- 將衛生紙的名稱依序存入list1
- 移除衛生紙價格的"\$"和",", 再存入list2
- 創建一個tissue.db的資料庫
- 在資料庫中創建一個TISSUE的表格
- 將list1和list2的內容依序存入表格

```
def step_03():
```

```
import sqlite3
import pandas
import matplotlib.pyplot as plt

plt.rcParams['font.sans-serif'] = ['Microsoft JhengHei']
plt.rcParams['axes.unicode_minus'] = False

with sqlite3.connect ("tissue.db") as con:
    c = con.cursor()
    tissue = pandas.read_sql("SELECT * FROM tissue" , con=con)

list1 = list(tissue.name)
    list2 = list(tissue.price)

print(list1)
    print(list2)
```

Python

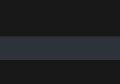
def step_03():

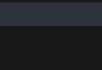
- 在Matplotlib圖表中顯示中文
- 連接到資料庫tissue.db
- 提取資料庫內所需元素
- 將提取出元素(name和price)分別存入list1和 list2

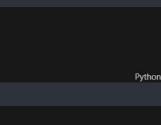
```
plt.bar(list1 , list2)
plt.title("家樂福衛生紙價格比較")
plt.xlabel("品牌")
plt.ylabel("價格(元)")
```

```
plt.barh(list1 , list2)
plt.title("家樂福衛生紙價格比較")
plt.xlabel("品牌")
plt.ylabel("價格(元)")
```









Python

Python

Modern Programming Language Onquesie112

14

Experience Sharing

def experience_Chen():

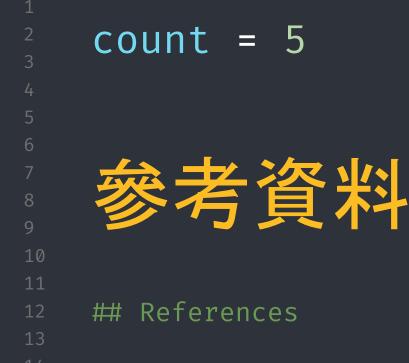
在爬取的過程中我問了AI很多次, 我發現它跑出來的結果不一定是能用的, 我們還需要幫它改很多地方才能使用, 他很喜歡在自己的觀點裡面跳針。當我終於爬出結果時, 我很開心, 以後比價時不用再開很多分頁了

def experience Wu():

在這次爬蟲作業中,除了鞏固課本的知識外,還學到了 新的知識,像是sqlite3和panadas模組的用法、以及 如何將在Python的資料存到資料庫,再從資料庫取出 資料來繪圖. 讓我學習到了很多

def experience Blue():

我們在過程中有詢問CHATGPT,效果不佳,它給的資 訊可能是從自別的網頁的一段擷取上來,除了會有種 少了甚麼東西的感覺照著打最後也沒辦法順利執行



def references():

- Kao Jia. (2023, May 28). 【資料分析】如何用 Python執行SQL語法. Medium. https://medium.com/@kaojia/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E5%A6%82%E4%BD%95%E7%94%A8python%E4%B8%8A%E5%9F%B7%E8%A1%8Csql%E8%AA%9E%E6%B3%95-60dfbdd6d3dd
- 日常tips. (2020, July 30). Python读取sqlite数据并导入DataFrame的两种方法. CSDN.

https://blog.csdn.net/veritasalice/article/details/107682782

- 鄧文淵. (2021). Python初學特訓班 Python Beginner Course (第四版). 碁峯.
- 李天胜. (2022). Python量化交易实战 (第一版). 中國水利水电出版社.

```
Thanks; {
    第八組:
         111210518 陳辰鈺
         111210520 吳佳泰
         111210524 藍健洲
                 CREDITS: This presentation template was
                 created by Slidesgo, including icons by
                 Flaticon, and infographics & images by Freepik
                 < Please keep this slide for attribution >
```