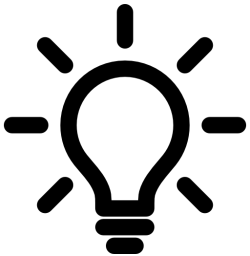
E-commerce Web Crawler

Group 14
Li Yufei (A0162548Y)
Gautam Rajulu (A0168916N)
Hilda Anggraeni (A0170754A)
Choong Yong Xin (A0171596U)

Project Idea



Selected Question:

1. Useful applications with Web crawler (extension of Assignment 1 part D)

Idea:

Building a web crawler to compare products across different e-commerce websites

Motivation

With the advent of e-commerce, there are a large number of available e-commerce websites available, including Lazada, Ezbuy, Qoo10 and many more. When one wants to buy a product online, comparing items across different websites can be a tedious process due to the large number of websites. Therefore, our group aims to build a web crawler to automate the process by visiting various ecommerce websites and listing down the prices and links for the product that user wants to buy.

Current Implementation

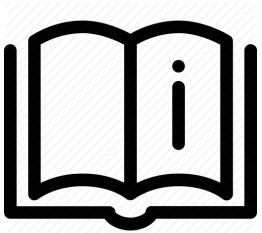
- Our program is implemented as a CLI application .
- We chose three websites to crawl:

  https://www.lazada.sg, https://www.qoo10.sg and https://www.amazon..sg.
- For each website, we append the product name that user inputs as a search keyword to the URL, e.g.:
  https://www.lazada.sg/catalog/?q=bag,
  https://www.qoo10.sg/s/BAG?keyword=bag,
  https://www.amazon.sg/s?k=bag

Current Implementation

- Then, we crawl the URLs to the websites in parallel and get the HTML pages containing our search result.
- Next, we parse each HTML page to obtain a list of product names, prices, and links to the product pages. The length of this list is either the maximum number of product per website supplied by user, or the number of products returned by the search result (if less than the user input).
- Finally, we combine the list for each website, sort the results from lowest to highest price, and display them to user.

Instructions to Run the Program



1. Install all the required packages:
   pip install requests
   pip install 'beautifulsoup4==4.8.1'
   pip install json
   pip install parsel
2. Run the program with the command `**python main.py**`.
3. Input the product that you want to compare prices for and the maximum number of results that you want from each website when prompted, as demonstrated in the next slide.

```
root@H /mnt/d/SEM5/CS3103/webcrawler master python3.6 main.py
What product do you want to compare prices for?
iPhone X
How many results do you want from each website at most?
5

Comparing prices for: iPhone X

Total results: 10


Product: Tempered Glass Screen Protector IPhone Huawei Xiaomi Samsung XS Max Plus
Price: $2.99
Link: https://www.qoo10.sg/item/TAOZI-TEMPERED-GLASS-SCREEN-PROTECTOR-IPHONE-HUAWEI-XIAOMI-SAMSUNG/414488681


Product:  BASEUS FREE Screen Protector!!! Apple iPhone 11 Pro Max / XR XS XS MAX Case Tempered Glass [SG]
Price: $3.9
Link: https://www.qoo10.sg/item/BASEUS-BASEUS-FREE-SCREEN-PROTECTOR-APPLE-IPHONE-11-PRO-MAX-XR-XS-XS/553386266


Product: Nanotech Screen Protector iPhone 11 Pro Max/Samsung Note 10 9 S10 Plus S10e
Price: $4.9
Link: https://www.qoo10.sg/item/NANOTECH-NANOTECH-SCREEN-PROTECTOR-IPHONE-11-PRO-MAX-SAMSUNG-NOTE-10-9/419142022


Product:  KnightShield Iphone 11 Pro Max Screen Protector XS Max XR X Samsung Note 10 Note 9 Note 8 S10 Plus
Price: $4.9
Link: https://www.qoo10.sg/item/KNIGHTSHIELD-KNIGHTSHIELD-IPHONE-11-PRO-MAX-SCREEN-PROTECTOR-XS-MAX-XR-X/490183990


Product: Sanptoch CaseMe Business Luxury Flip Leather Phone Case For iPhone X Xs Max XR Wallet Card Slots Cases Cover For iPhone 8 7 6 6s plus Holder
 Protective Casing Shell
Price: $9.7
Link: www.lazada.sg/products/sanptoch-caseme-business-luxury-flip-leather-phone-case-for-iphone-x-xs-max-xr-wallet-card-slots-cases-cover-for-iphone-
8-7-6-6s-plus-holder-protective-casing-shell-i326274129-s686376151.html?search=1


Product: Apple iPhone X Smart Phones 64/256GB Rom 5.8Inch Face ID 3GB Ram 12MP Dual Camera A11 Bionic Hexa
Price: $800.0
Link: https://www.qoo10.sg/item/APPLE-IPHONE-X-SMART-PHONES-64-256GB-ROM-5-8INCH-FACE-ID-3GB/670666527
```

Note: Since Lazada adopted a security feature that prevents bot from accessing and being on the site, if you run the program too often, there won't be any results from lazada. To avoid this, try to run the program after a time interval between each run, say 1 minute.

Challenges Faced

- Had difficulty getting a response from the e-commerce host as unlike Assignment 1D, the user-agent needs to be declared as part of the request.
- There were multiple href links in each listing which required some studying of the various possible links in the different listings to filter out the correct ones.
- Lazada has adopted a security feature that prevents bot from buying stuff, and being on the site. So if request for lazada pages too often using our crawl tool, it may be blocked as bot.

Possible Extensions

- Add other e-commerce websites to crawl. Could possibly extend to other types of products e.g hotels and flights as well.
- Create a simple GUI to make the program more user-friendly.
- Currently, the results chosen are based on the default list which sorts items by relevance. In the future, we could allow users to specify whether they want to view their results based on popularity or price etc.
- Implement technologies like user agent rotating and proxy/ip address rotating, to prevent being recognized as a bot by the server.

Thank you

Github repository for source code can be found here:
https://github.com/choongyx/webcrawler.git