

Web Crawler Assignment

Submitted by **Roni Thomas**

INITIAL APPROACH

Initially, the idea was to crack the assignment was to scrape the amazon website for the required data and compare the prices with Flipkart's device listing and highlight the lowest price available. Flushing all the retrieved data onto an excel file for easy viewing.

FINAL APPROACH

Started off with scrapping the listing on the Amazon website, which I managed to get done successfully got the device names perfectly but found a problem in getting the prices of specific devices as amazon's site had advertisements before the actual listing. As you can see in the image below, the actual listing starts from "Apple iPhone 12 Mini Blue, 128GB Storage" and the price starts from "68,900".

Here I was facing a problem to ignore the advertisements and come along to the actual listing. since all the prices were mentioned inside the same class name, I couldn't find a way to ignore the first three prices.

1-16 of over 1,000 results for "mobile"

Delivery Day

☐ Get it by Tomorrow

☐ Get it in 2 Days

Department

Smartphones & Basic Mobiles

Smartphones

Basic Mobiles

See All 2 Departments

Avg. Customer Review

★★★★★ & Up

★★★★★ & Up

★★★★★ & Up

★★★★★ & Up

Brand

☐ Samsung

☐ Redmi

☐ Oppo

☐ Apple

☐ Vivo

☐ Panasonic

☐ OnePlus

See more


Price

11,111 - 75,000

SAMSUNG

Galaxy M02: 6.5" HD+ Display for Cinema Experience


Save up to 13% on Samsung >



Limited time Deal

₹7,499⁰⁰ ✓prime


₹8,499.00 (12% off)



Limited time Deal

₹6,999⁰⁰ ✓prime


₹7,999.00 (13% off)



Limited time Deal

₹7,499⁰⁰ ✓prime

₹8,499.00 (12% off)



Sponsored ⓘ

Apple iPhone 12 Mini Blue, 128GB Storage

★★★★★ < 647

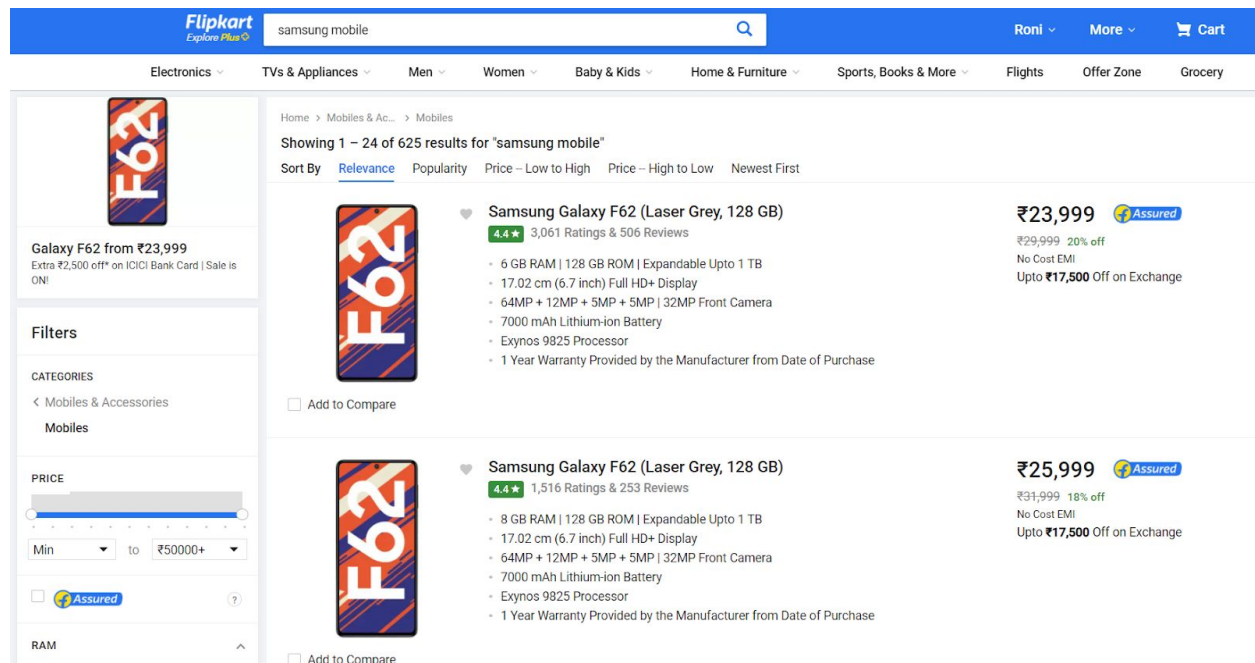
₹68,900 ₹74,900 Save ₹6,000 (8%)

Flat INR 1500 Off on HDFC Bank Cards

Get it by Tomorrow, March 7

FREE Delivery by Amazon

After trying different approaches to crack this solution I was unable to crack it. So I decided to choose another approach that is to scrape Flipkart first and compare it with amazon's listing.



Like the image speaks Flipkart doesn't have advertisements at the beginning so it was quite a good turnaround to scrape the device names and prices in a single go.

PYTHON LIBRARIES USED

1. Regex - for string formatting
2. Openpyxl - to read and write excel files
3. Selenium - browser automation tool
4. Time - to introduce a delay in the process

METHODOLOGY

Open Browser => Go to Flipkart => Scrape Data => Bind To Excel

DEVELOPMENT STAGES

1. Importing required libraries

All the libraries were imported when the need arrived. All the imports are kept at the beginning of the code for easy understanding.

2. Assigning a driver for the browser

Assigned a web driver so that it makes it easy to call the driver over and over for future uses.

3. Visiting the site & listing the data

Visiting the host site was an easy approach. The login page pop shows up sometimes. Then search for the term entering the element in the search bar clicks on a search button to get desired results. The scraped data are stored in a web object which was later looped and stored making a list element.

A function is defined inside the code "*def populate()*" to call it later in the script if the switching page is necessary for more results.

4. Creating an excel file

The extracted data were then stored in excel sheets for user preview. The titles were appended before the looping of the data. Openpyxl library is being used here as it supports reading and writing excel files.

CHALLENGES FACED

1. Closing the login page pop-up as it shows up sometimes - Used a *try & except* method to solve the problem.
2. Entering the text in the search bar & clicking the search button.
3. Switching tabs - later it was possible by using the indexing method to provide the currently active tab number.
4. Selecting the sort by option

PROBLEMS SOLVED

1. PROBLEM 1

Accepting search term, as user input. Tested with terms Samsung mobiles, Apple mobiles.

Generates an excel file containing Product Name, Storage, User Rating & Price.

2. PROBLEM 2

Accepting the number of products to be written in an excel file from the user. Max results can be obtained is 72 (3 Pages)

Users can sort by the following options. Only one can be chosen at the time because Flipkart supports only one at a time.

```
0 = Low to High
1 = High to Low
2 = Relevance #default
3 = Popularity
4 = Newest First
```

3. PROBLEM 3

To keep the entire process like human behavior delay and time gaps are introduced in the required part of the code like clicking the search button, navigating pages for more results, and so on.

REFERENCES

1. Selenium Docs - <https://www.selenium.dev/selenium/docs/api/py/api.html>
2. Openpyxl - <https://openpyxl.readthedocs.io/en/stable/>
3. Medium Blogs
4. Stack overflow