

Web Crawler Assignment

Instructions

1. Please ensure that you follow all the coding conventions.
2. Use Java/C++/python to make a desktop application
3. As a part of this evaluation, you're required to provide a timeline schedule in a separate document (pdf/txt), structured into development phases with the problems faced and the solutions to it
4. You are also recommended to demonstrate your skills & expertise by enhancing any particular part of the application.
5. Any attempt at cheating, plagiarism, or any other malpractice will result in immediate rejection of your application.
6. There are a total of 6 problems, submit as many as you can.
7. In case of any doubt use your best judgment and write down the assumptions in pdf.

Prerequisites

1. Web crawling/scrapping
2. Selenium

Problem Statement

You need to create a simple application that takes the product names and parameters as input and then scrape Amazon and Flipkart for the products list.

Problem 1: Take search name as input and generate the output file

You should generate the txt/excel file containing the product name with the link, source (Flipkart/Amazon), price, product details [like model no, delivery time], category of product, etc (you can pick important details bonus points for creativity)

Problem 2: Take arguments input and search product on the basis of these filters

- a. **Number of products** (default is 10 and max is 50)
- b. **Sort by** Price low to high, high to low, relevance/featured, new arrivals (default is relevance/featured)
- c. **Price range** (default is no limit)
- d. **Delivery Pincode** (default should be 400072)

Problem 3: The browser scrapping behavior should be as close to user behavior as possible (you need to be creative on this how you will implement it) [it should be like the user is searching for products on Amazon/flipkart]

Problem 4: Need to generate another sheet that gives the lowest rate of the product of that particular model across Amazon/Flipkart

Problem 5: [Bonus] Flipkart and Amazon have protection against web-crawling which blocks your IP address if they detect any incorrect usage, you will have to implement a logic to ensure your system is

never blacklisted

Problem 6: [Bonus] Can we add unit test cases for providing the lowest rate of the product?

E.g. Input mobile

Sheet1

Product Name	Source	Price	Catagory	Model no/unique no
Samsung Galaxy M51 (Electric Blue, 8GB RAM, 128GB Storage)	Amazon	24,999	Smartphones	SM-M515FZBEINS
Vivo Y51 (Titanium Sapphire, 8GB RAM, 128GB ROM) with No Cost EMI/Additional Exchange Offers	Amazon	17,990	Smartphones	V2030
Samsung Galaxy M51 (Celestial Black, 8GB RAM, 128GB Storage)	Amazon	26,999	Smartphones	SM-M515FZKEINS
Samsung Galaxy M51 (Electric Blue, 128 GB) (8 GB RAM)	Flipkart	26,900	Smartphones	SM-M515FZBEINS
POCO M2 (Slate Blue, 64 GB) (6 GB RAM)	Flipkart	9,999	Smartphones	P15435

Sheet2

Product Name	Source	Min Price	Model no/unique product no
Samsung Galaxy M51 (Electric Blue, 128 GB) (8 GB RAM)	Flipkart	26,900	SM-M515FZBEINS
Vivo Y51 (Titanium Sapphire, 8GB RAM, 128GB ROM) with No Cost EMI/Additional Exchange Offers	Amazon	17,990	V2030
Samsung Galaxy M51 (Celestial Black, 8GB RAM, 128GB Storage)	Amazon	26,999	SM-M515FZKEINS
POCO M2 (Slate Blue, 64 GB) (6 GB RAM)	Flipkart	9,999	P15435

Note: In sheet 2 model no. can be matched or product name can be matched. In this only 1 product matched and the Flipkart has the minimum value
 [Bonus]

1. How will you make this logic generalize for all of the product categories (if we cant generalize then can we create a category-wise matching logic?)
2. How can we compare products within the same website e.g. in amazon [Samsung Galaxy M51 \(Electric Blue, 8GB RAM, 128GB Storage\)](#) and [Samsung Galaxy M51 \(Celestial Black, 8GB RAM, 128GB Storage\)](#) are the same products but the different color/vendor can we make a logic to eliminate this and do a price comparison?