Pharmeasy

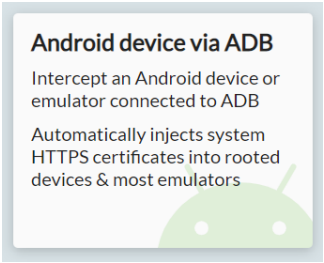
Title	Tags
Medium	App
Type of scraper	HTTP Programming
Type of crawling	Generic
Depth and Scraping Strategy	DFS Nested
Difficulty	Advance
Technology	GenyMotion HTTP Toolkit Python

PharmEasy is an Indian healthcare app and e-commerce platform for consumers involved in selling online medicines, diagnostics, telehealth, and online doctor consultation services. PharmEasy is headquartered in Mumbai.

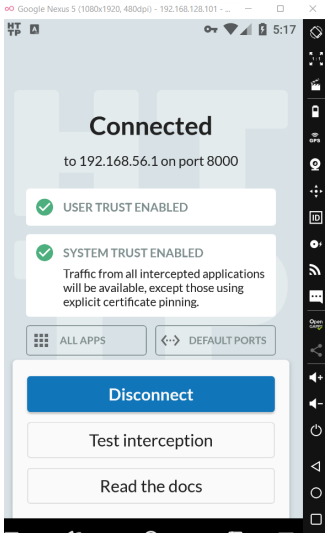
Pharmeasy app has a massive database of medicines and health care products. Our task is to scrap this data and perform an analysis of the API structure of Pharmeasy.

Process

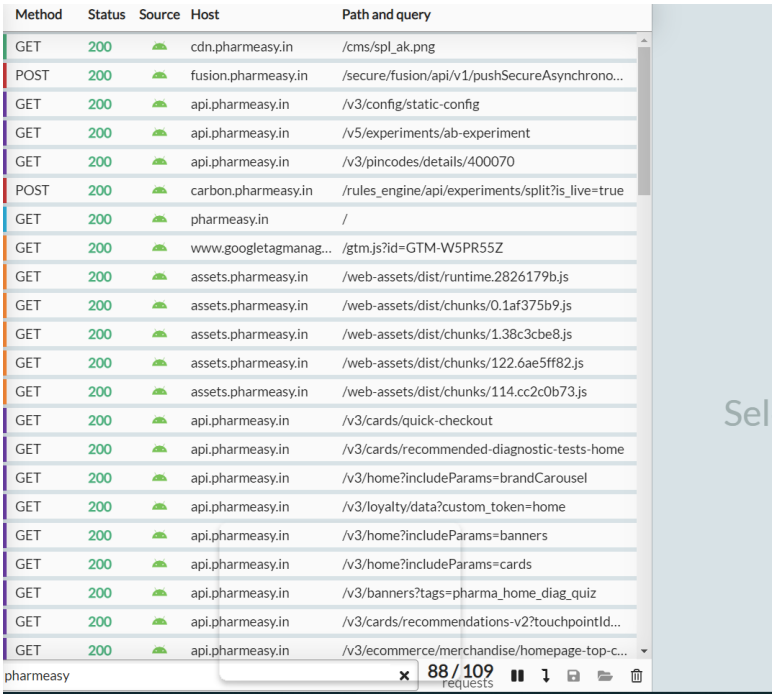
1. We will use GenyMotion for a rooted android device, HTTP Toolkit for runtime injection, and Python to scrape data.
2. Download the Pharmeasy apk from here.
3. Open the android emulator in GenyMotion.
4. Drag and drop the Pharmeasy apk file in the android emulator to install the apk.
5. Open the HTTP Toolkit. Choose the option “Android Device via ADB”. It will automatically connect to the emulator. You can verify it on the emulator.



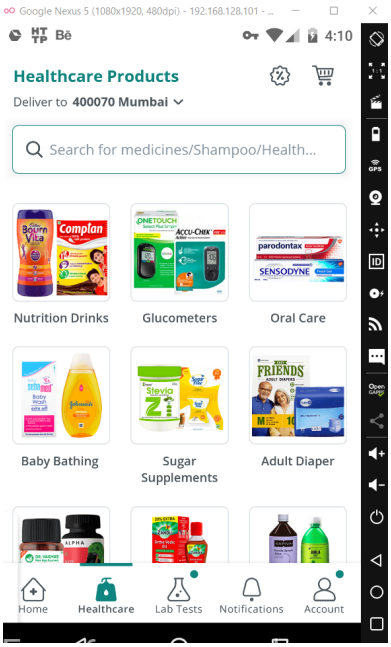
6. To verify whether our emulator device is successfully connected to the HTTP Toolkit, we can confirm the connection on the emulator itself.



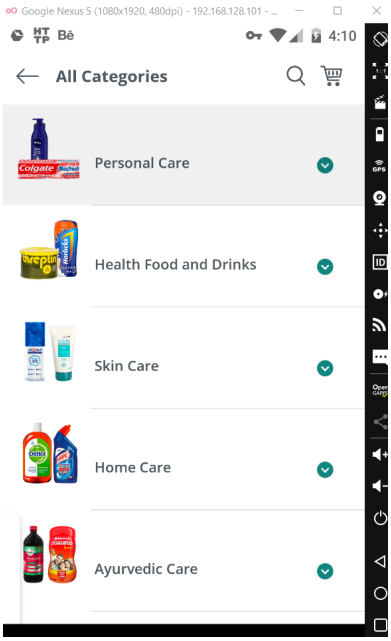
7. Open the Pharmeasy app on the emulator.
8. As soon as we start the app, we can see the requests being recorded on the HTTP Toolkit.
9. We can search for the Pharmeasy APIs in the search bar in the bottom left corner.



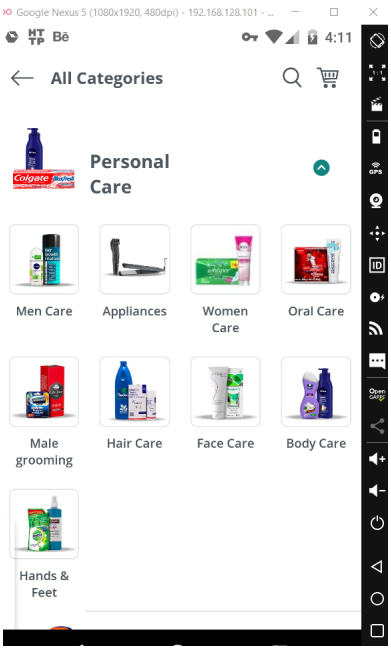
10. Switch to the emulator. Swipe to the “healthcare” section. You will find a section called “Categories” while you scroll down. We can guess that most of the data would have been stored across various categories.



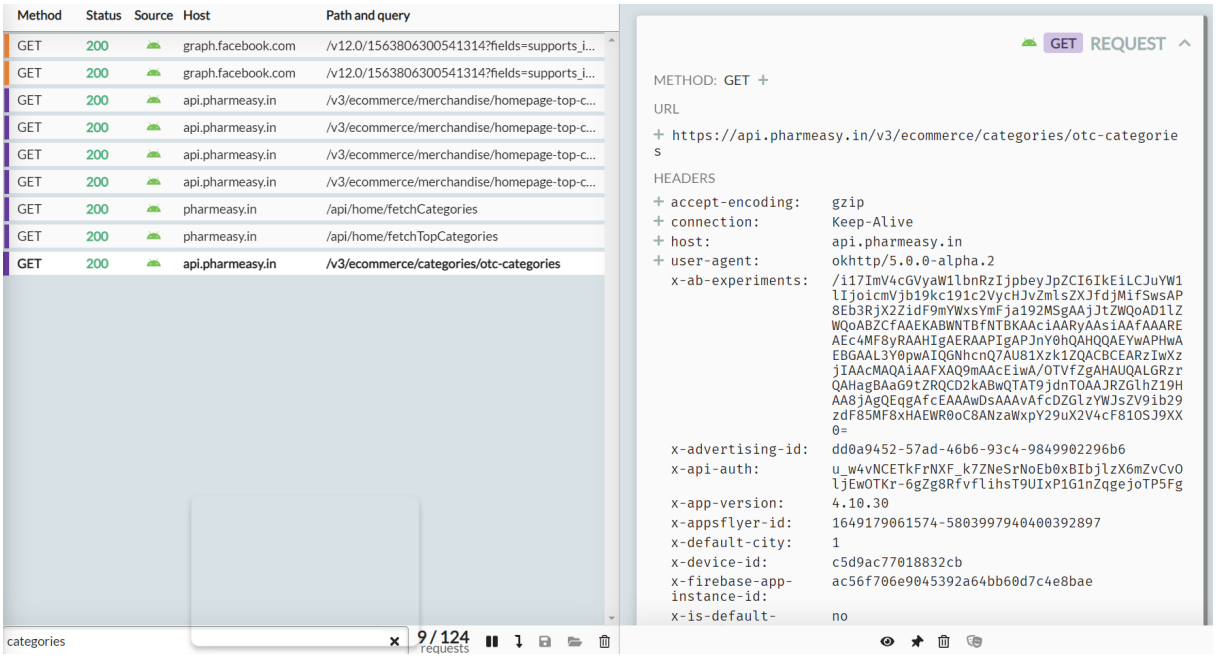
11. Choose the “see all” option to further explore other categories.



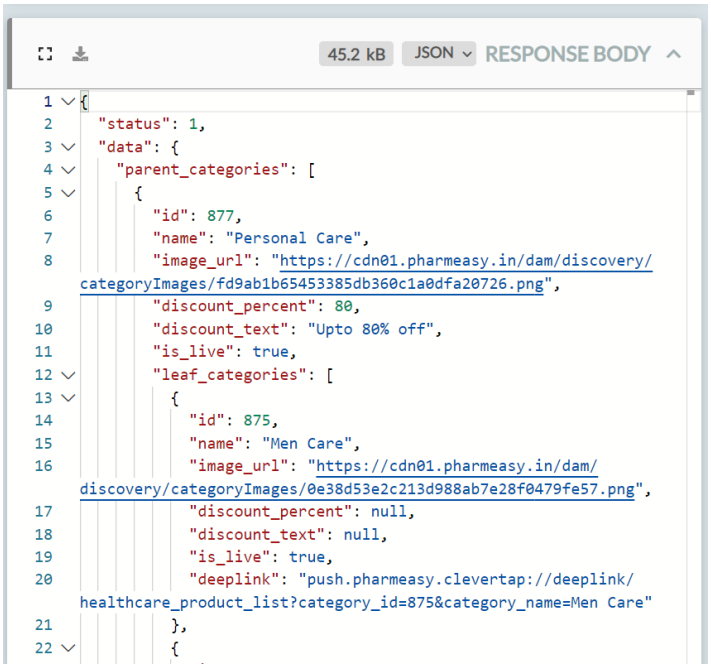
12. And tap on any of the categories to further explore the sub-categories.



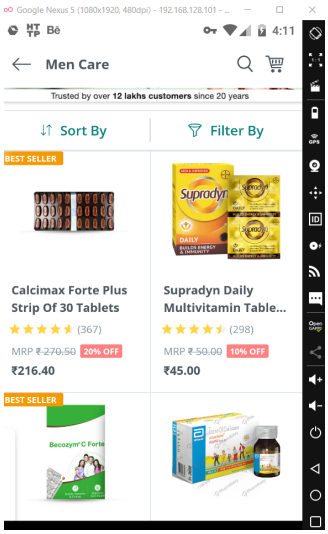
13. We need to extract an entire list of these categories but for that purpose, we will have to investigate the APIs that retrieve this data. We will search for categories in HTTP Toolkit.



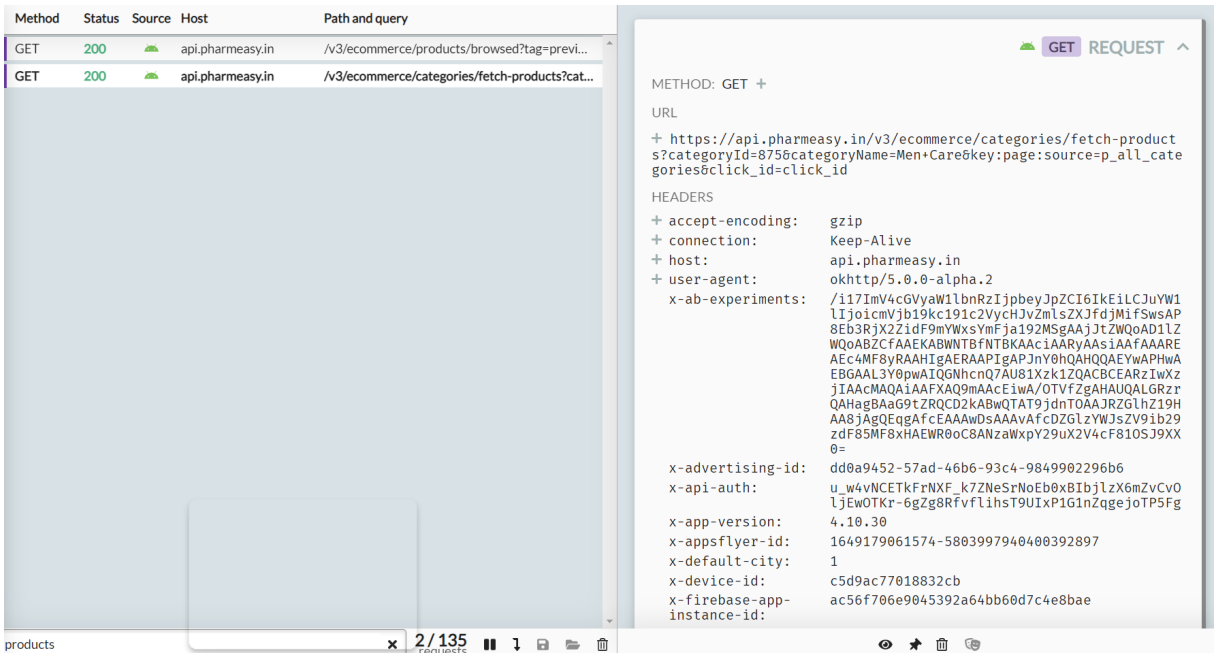
14. Post investigating, we found the right response that retrieves all the categories.



15. After retrieving the categories, our task is to retrieve products from each category. So choose any category and navigate to the product result page.



16. We will start investing in HTTP Toolkit to find the API requesting the product data.



Here, we have the API that retrieves products page-wise.

17. Here is a snapshot of the final data retrieved:

Before moving to the implementation, we shall understand the characteristics of our scraping approach for Pharmeasy.

- As we are interacting with API URLs using HTTP requests, the data scraper is classified into HTTP Programming based scraper.

- Our objective is to maximize the data scraped. So, we have set no filters on the crawler. Thus, we classify it as a generic crawling.
- First, we scraped categories and later products were derived from categories (we'll clarify in the implementation process). This makes the scraper explore 2 different layers in BFS fashion. Thus, the scraper is nested and follows the DFS fashion to scrape data.

Implementation

1. For this implementation, we shall write a python script to make HTTP requests to GET data from the Pharmeasy database. We'll initiate by assigning the category and products to variables.

   categoryApiUrl = "https://api.pharmeasy.in/v3/ecommerce/categories/otc-categories"
   productApiUrl = "https://api.pharmeasy.in/v3/ecommerce/categories/fetch-products?paginated=1&page="+str(pageNumber)+"&categoryId="+str(categoryIdArray[categorySerialNumber])+"&page_id=p_home_9128b2be-4e99-803c-28fccfebd6ba&key:page:source=p_home"

2. We have set the API parameters to variable so we can iterate over them to change pages and categories.
3. Analyzing and retrieving the list of all the categories.

```
In [128]: for i in range(0,len(parsedResponse['data']["parent_categories"])):
              for j in range(0,len(parsedResponse['data']["parent_categories"][i]['leaf_categories'])):
                  categoryIdArray.append(parsedResponse['data']["parent_categories"][i]['leaf_categories'][j]['id'])

In [129]: len(categoryIdArray)
Out[129]: 129

In [130]: categoryIdArray
Out[130]: [875,
           873,
           871,
           874,
           876,
           872,
           900,
           902,
           903,
           1053,
           641,
           980,
           646,
           1052,
           946,
           647,
           1130,
           96,
           215,
```

4. Scraping products in DFS fashion.

```
while categorySerialNumber < 129:
    while len(products) != 0 and pageNumber < 3:

        for i in range(0, len(products)):
            data = [products[i]]
            with open('Pharmeasy.csv', 'a', encoding="utf-8") as csvfile:
                writer = csv.DictWriter(csvfile, fieldnames=field_names)
                writer.writerows(data)
        pageNumber = pageNumber + 1
    pageNumber = 1
    categorySerialNumber = categorySerialNumber + 1
```

5. Here is the snapshot of the dataset

	productId	name	slug	manufacturer	productType	mrpDecimal	salePriceDecimal	discountDecimal	discountPercent
0	19501	Calcimax Forte Plus Strip Of 30 Tablets	calcimax-forte-plus-strip-of-30-tablets-19501	CALCIMAX	2	270.50	216.40	54.10	20
1	3514609	Supradyn Daily Multivitamin Tablets 15s	supradyn-daily-multivitamin-tablets-15s-3514609	SUPRADYN	2	50.00	45.00	5.00	10
2	36999	Becozym C Forte (Vitamin B Complex + Biotin +...	becozym-c-forte-vitamin-b-complex-biotin-vitam...	BECOZYM C FORTE	2	25.00	23.00	2.00	8
3	209565	Arachitol Nano 60k Sugar Free Bottle Of 5ml So...	arachitol-nano-60k-sugar-free-bottle-of-5ml-so...	ARACHITOL	2	86.66	74.53	12.13	14
4	227487	Ultra D3 Drops - Vitamin D3 For Infants Vitami...	ultra-d3-drops---vitamin-d3-for-infants-vitami...	ULTRA D3	2	37.85	24.60	13.25	35
5	235886	Liveasy Wellness Calcium, Magnesium, Vitamin D...	liveasy-wellness-calcium-magnesium-vitamin-d3-...	LIVEASY	2	599.00	539.10	59.90	10
6	25749	Revital H Woman Health Supplement	revital-h-woman-health-supplement-	REVITAL	2	345.00	293.25	51.75	15

The mini version of the dataset can be accessed here.

Thank you!