

DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' grade,

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078.



Minor Project Report

on

“Pulling data from website to identify best deals on products”

Submitted By

Rhuthu Hegde

E3-1DS18CS731

Fifth Semester B.E (CSE)

in

Emerging Technologies

18CS5DMETG

Under the guidance of

Prof. Poornima K S

Assistant Professor

Dept. of CSE

DSCE, Bangalore

Department of Computer Science and Engineering

Dayananda Sagar College of Engineering

Bangalore-78

ABSTRACT

Web Scraping is a technique employed to extract large amounts of data from websites where the data is extracted and saved to a local file in text format or table format in a database on our computer. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. The scraped data is usually in CSV, TSV or JSON format. Web scraping has traditionally been used for automating pricing solution, market research, sentimental analysis, news & content monitoring and real estates.

In this project, we have automated the process of searching for the best deals offered on products in an ecommerce website by scraping the details and the price of the desired products and sorting it according to price. This helps the customer make an informed choice on the product they going to buy and also save their time. To automate the task, we have used Robotic Process Automation (RPA) software.

CONTENTS

Page No

CHAPTER 1 INTRODUCTION	3-6
1. Introduction to RPA	3
1.1. Introduction to Automation Anywhere (AA)	3
1.1.1. RPA on Automation Anywhere Platform	3
1.2. Objectives	5
1.3. Problem Statement	5
1.4. Scope of the work and its importance	5
1.5. System Requirements Specification	6
CHAPTER 2 DESIGN/IMPLEMENTATION	6-8
2.1 Flow Diagram	6
2.1.1 The process being Automated	7
2.1.2 Why it should be automated	7
2.2 Implementation	7
CHAPTER 3 TESTING/RESULT AND ANALYSIS	9-13
3.1 Bot Execution Procedure	9
3.2 Screenshots of Instructions on Control Room	9
CHAPTER 4 CONCLUSIONS/FUTURE ENVIRONMENT	15-17
4.1 Benefits of Automation Anywhere	15
4.2 Applications of Web Scraping	16
4.3 Future Enhancements	16
REFERENCES	17

CHAPTER 1

INTRODUCTION

Emerging Technologies

1.1 Introduction to RPA

Robotic process automation is a software technology that enables users to create software robots or “bots” that can learn mimic and then execute rules-based business processes and is easy to use. RPA enables users to create bots by observing human digital actions i.e., showing them what to do. These bots can interact with any application or system the same way humans do but they can operate around the clock, nonstop, faster than humans with a hundred percent reliability and precision.

RPA can be thought of as a digital workforce that has a positive effect on business and its outcomes. It provides greater productivity by accelerating workflows, greater accuracy, cost savings and fast ROI, it integrates seamlessly across platforms, improved customer experiences, scalability and integrates Artificial Intelligence (AI).

1.1.1 Introduction to Automation Anywhere (AA)

Automation Anywhere is a leading global enterprise Robotic Process Automation (RPA) platform that provides intelligent automation software solutions to various industries. Their vision is to liberate humans from mundane repetitive tasks, allowing them to more time to use their intellect and creativity to solve higher order business challenges and perform knowledge work.

Automation anywhere was originally founded as Tethys Solutions LLC in San Jose, California by Mihir Shukla, Ankur Kothari, Neeti Mehta, and Rushabh Parmani. The company rebranded itself as Automation Anywhere, Inc in 2010. As of 2019, Automation anywhere has more than 3,500 customers including Google, LinkedIn, GM and the World Health Organization. Automation Anywhere was ranked 29th on the Forbes Cloud 100 and was named a leader in Gartner’s Magic Quadrant for RPA software in 2019.

1.1.2 RPA on Automation Anywhere Platform

Automation Anywhere’s products combine traditional RPA with cognitive elements such as natural language processing (NLP), reading unstructured data

and machine learning capabilities and business intelligence to create a digital workforce. It is a cloud -native platform that provides a web based-interface to the user. It provides both RPA- as-a-service option as well as an on-premise deployment with enterprise-class data privacy, security and encryption in each.

Automation Anywhere has three primary components

- Control Room
 - Bot Creator
 - Bot Runner
1. **Control Room:** It is a web-based platform that controls the bots. It also provides user management, source code control, analytics of the bots through the dashboard and licence management. There are two types of licences in AA
 - Dev licence: Allows you to create, edit and run a bot
 - Run licence: Allows you to run the bot but you cannot make any changes to it
 2. **Bot Creator:** Developers use Desktop-based or web-based applications to create bots. Their dev licences are checked with the one configured in the control room. On authentication, the code of the bots the create is stored in the control room.
 3. **Bot Runner:** It is the machine where you run the bot. You can have multiple bots running in parallel. The bots report back the execution logs back to the control room.

Other important components of Automation Anywhere are

1. **Bot Insights:** The tool shows statistic and display graphs to analyse the performance of every bot in the system. Here, you can also calculate the time you have saved because of the automation process.
2. **Bot Store:** Bot Store is a first digital workforce marketplace. Here, you will get lots of pre-built bots for every type of business automation.

There are three types of bots on Automation Anywhere platform

1. **Task Bots:** Task bots are bots which automate rule-based, repetitive task, in areas like document administration, HR, claims management, IT services and more. This leads to immediate improvement in productivity, error reduction, and cost saving.
2. **Meta Bots:** Meta bots are the automation building blocks. It is designed in such a way that with application updates or changes you need to make minimal edits to the bot. Changes automatically apply to any process utilizing that bot.

3. **IQ Bots:** This is an advanced tool which can learn on its own and perform a task. IQ Bot offers automation using the highly advanced cognitive technology. It works on the concept to organize an unstructured data while improving its skills and performance.

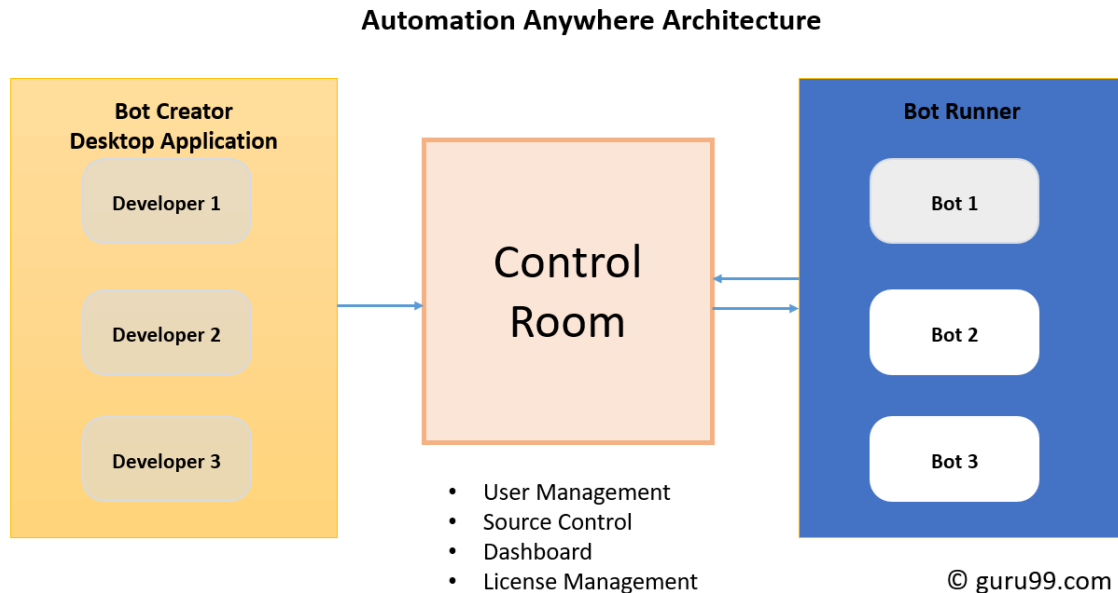


Fig 1.1.1: Automation Anywhere architecture

1.2 Objectives

The objective of the project is to help users find the best deals for products on the internet without them having to do much work.

1.3 Problem statement

Pulling prices of items from websites to identify best deals on products.

1.4 Scope of the work

E-commerce is the activity of buying and selling of products on online services over the internet. It plays a vital role in our daily lives because of the rise in the use of mobile devices all over the world. These websites offer various discounts and organize sales. Due to the plethora of e-commerce stores that have been established in the recent years, It has become difficult for the customer to keep track of all these websites, the deals offered by each of them and compare the prices and discounts. Using RPA, the task of checking different products on the websites, comparing prices and getting the best deals can be automated, making the customer's task easier and faster.

To save the user's time and money, we automate a bot which will visit the shopping websites/e-commerce website, search for the particular item in a website and scrap the data which includes the prices of the items. The scraped data with

the prices is then stored in an excel sheet and sorted according to the user's requirement. This would help the user select the best deal for an item.

1.5 System requirement specification

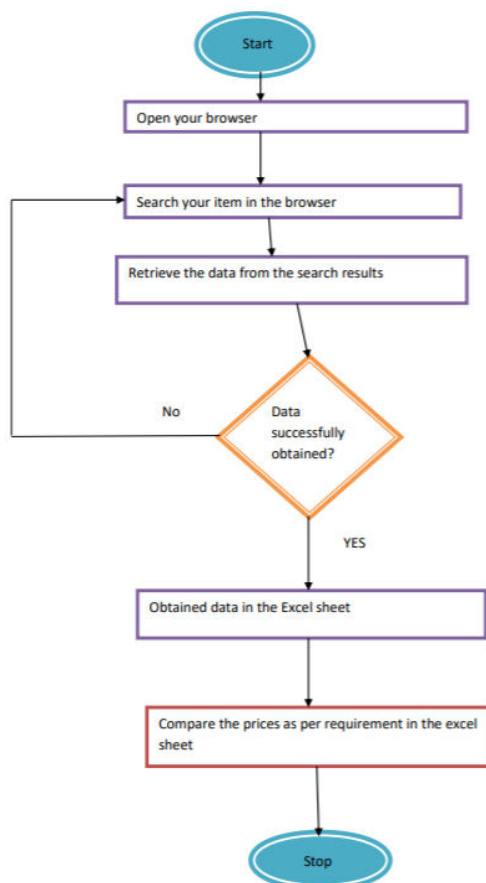
The project is executed with the help of Automation Anywhere Enterprise A2019 Community edition software on our laptops which has Windows 10 operating system, i5 processor, 1TB memory and 8GB RAM.

Automation Anywhere Enterprise A2019 Community edition software is free of cost and can be downloaded from the official Automation Anywhere website by registering our details.

CHAPTER 2

DESIGN AND IMPLEMENTATION

2.1 Flow Diagram



The flow diagram above describes the steps to be taken to automate the process.

Step 1: Open the website and launch the website.

Step 2: Search for the item in the website.

Step 3: Scrape the data from the search results and if data is successfully obtained store it in an excel sheet. Go to step 5

Step 4: If data is not successfully obtained go back and try to scrape it again.

Step 5: Sort the data stored in the excel sheet according to the requirement.

2.1.1 The process being automated:

Here, we intend to automate the process of searching from e-commerce website from typing our required object to finding all the objects related to our search and capturing them to .csv file, sorting the files and then loading our data to excel sheet.

From our data we have understood that we are scraping pattern-based data. We automate the entire process of searching and comparing the prices of different models.

2.1.2 Why it should be automated:

Extracting data from a website is fairly a simple and straightforward process. Images can be saved and text can be copied. However, this kind of data extraction is practically impossible when you need large amount of data from multiple websites for a business use case.

To crawl and extract large amounts of data continuously, an automated web scraping setup can be employed. The benefit is minimal manual interference after the initial setup and fully automated web scraping thereafter.

2.2 Implementation:

For a web scraping setup to work on full automation, the bot should be able to navigate through the different pages on a website and save the required data fields. Navigation is the key aspect when it comes to automating a web scraping task. This is because, different websites use different kinds of navigation systems and these vary greatly in terms of the complexity. While some websites use simple numbered navigation, there are some modern websites that use infinite scrolling and other AJAX based dynamic navigation techniques.

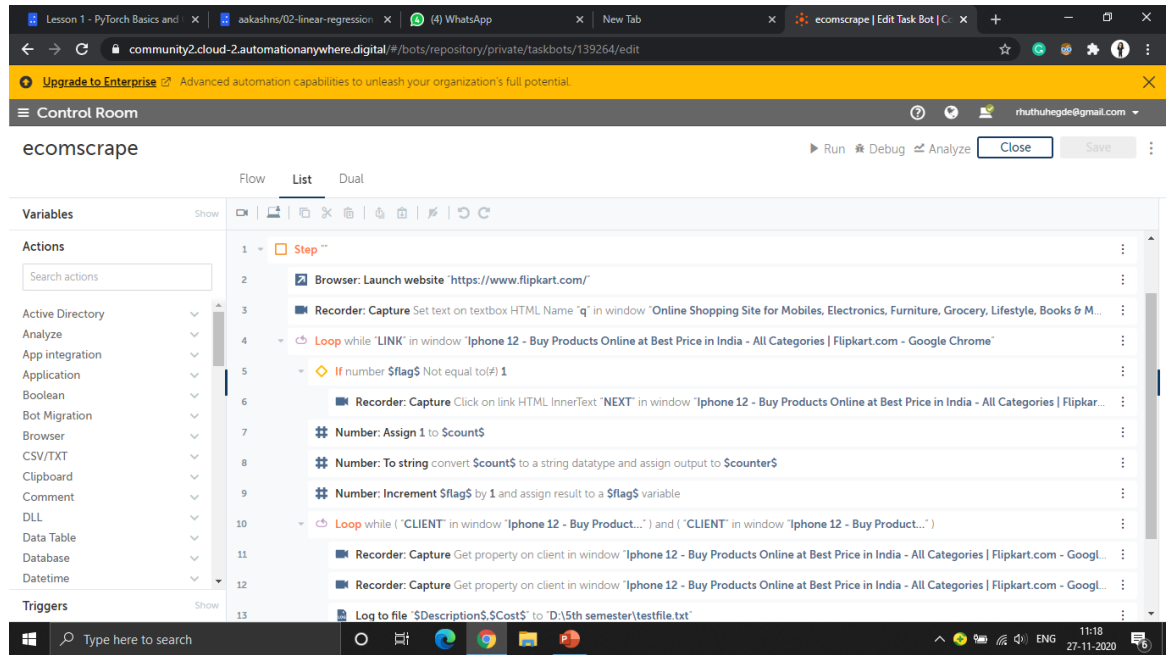


Fig 2.2.1: The implementation steps in list view.

The steps followed to implement the bot are:

1. First, we launch the e-commerce website.
2. Then, Search for the required components using the search bar in the website. We have automated it by capturing the search bar using web recorder and specifying the text to be entered and action to be performed (ex: Click or Enter) in the control room.
3. Using web recorder, we train the bot to extract the product name and product price details from the website. We use two loops one to extract the details of each product on a page and the other to move to the following page (if any). We find the Xpath of the website so that all the pages related to our search product are considered. This would enable bot to extract data from all the pages.
4. The extracted data is stored in a log file in text format. This is achieved using the log-to-file component in the Automation Anywhere control room.
5. The text data is then stored in a structured form in an excel spreadsheet. This is achieved by writing a python script and running it using Automation Anywhere control room.
6. The price data which was initially stored in string format for web scraping is converted to integer, again using the python script.
7. Finally, The product details are sorted in ascending order according to price using the python script and running the bot in the control room.

CHAPTER 3

Testing, Result and Analysis

3.1 Bot Execution Procedure

The community edition has the control room and functions to run and stop the bot from execution.

When we execute the bot, it first launches the website and searches the item. The bot extracts the data from all the webpages of the searched item. This data is collected from website and stored in a text file. The bot then converts the text file to excel file using a scripting language (Python) and finally data is formatted and sorted using a Python method and presented to the user on the excel sheet.

If the bot runs successfully, we get a message showing that bot has run successfully, otherwise we receive a message that bot did not run successfully.

3.2 Screenshots of instructions on control room

The following are the screenshots of control room. The Control Room is a centralized management point for all bots. It is a Microsoft Windows server-based web application providing a single administrator interface for Enterprise-wide bot deployment, management, and control, including the Bot Insight analytics functions and Elasticsearch search functions.

List:

This shows the detailed explanation of our workspace.

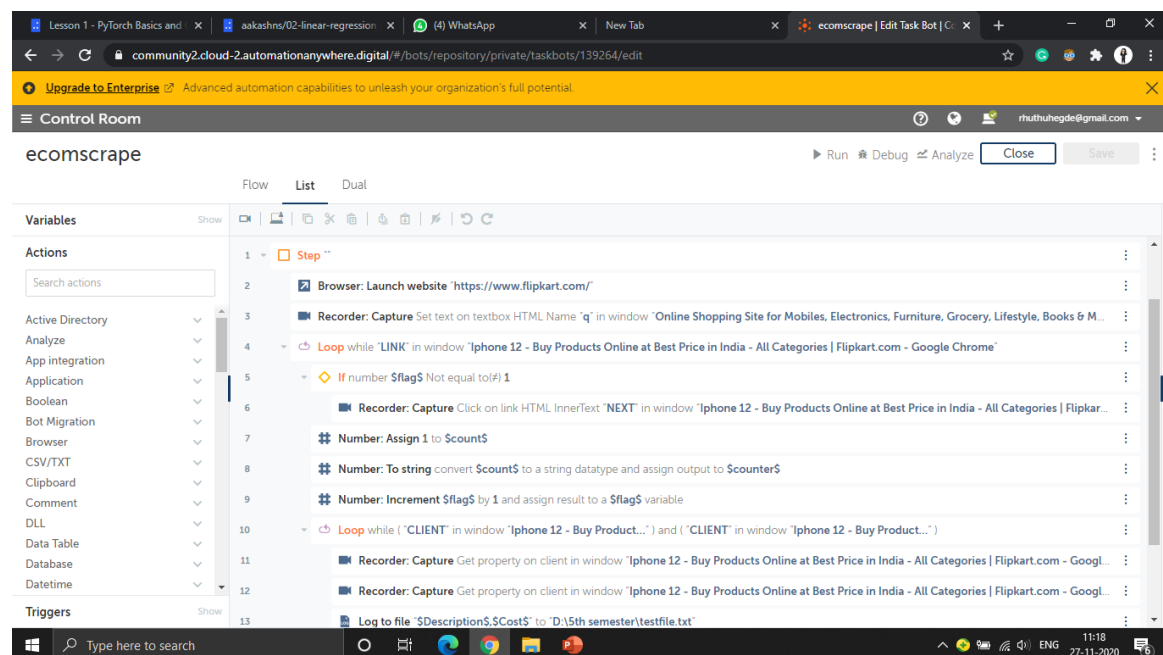


Fig 3.1.1: This list shows the steps to launch the website and capture objects using recorder.

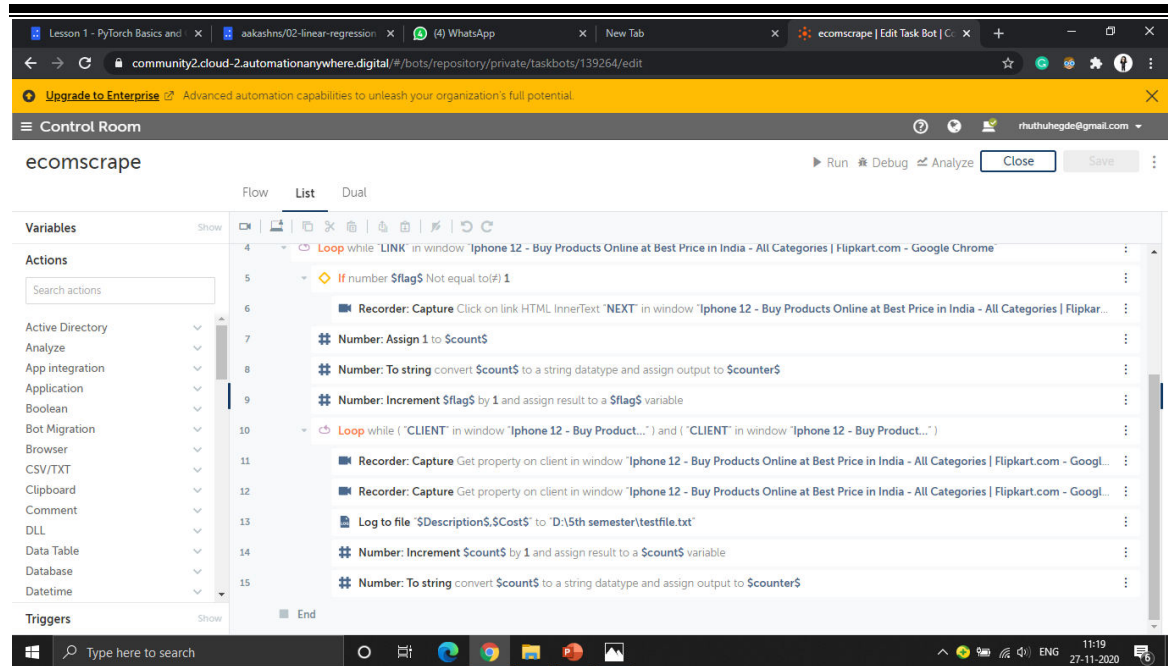


Fig 3.1.2 Extraction of data to log file takes place here.

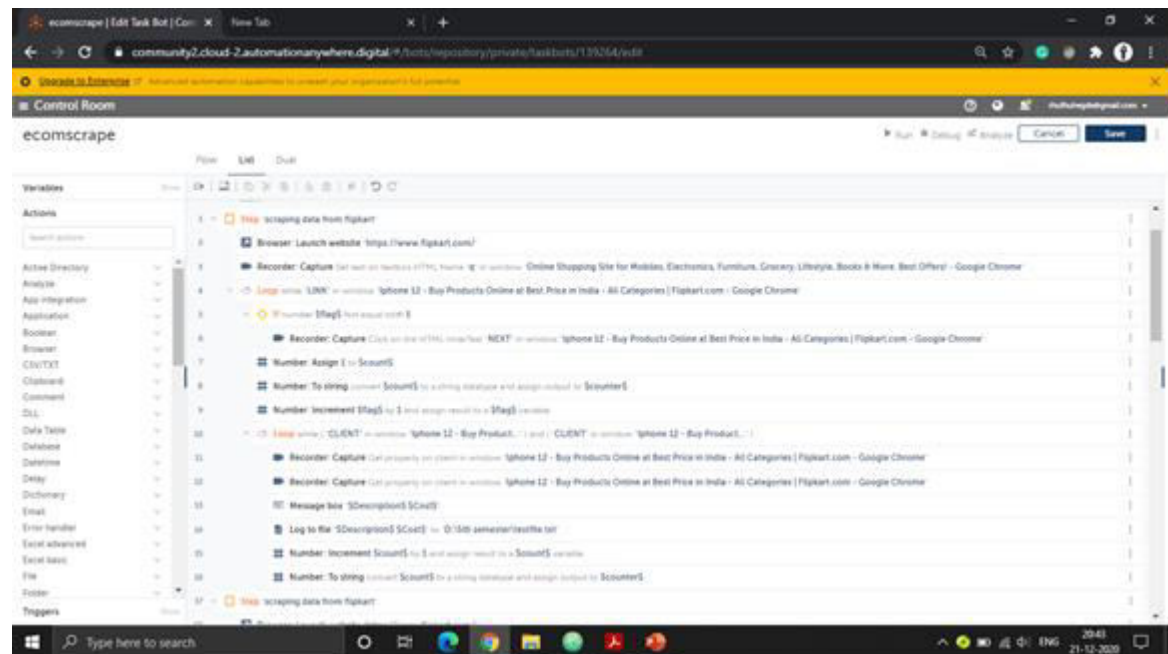


Fig.3.1.7 This is the entire list of commands we execute showing the loops and entire scraping of data

Flowchart:

This shows the pictorial representation of the instructions.

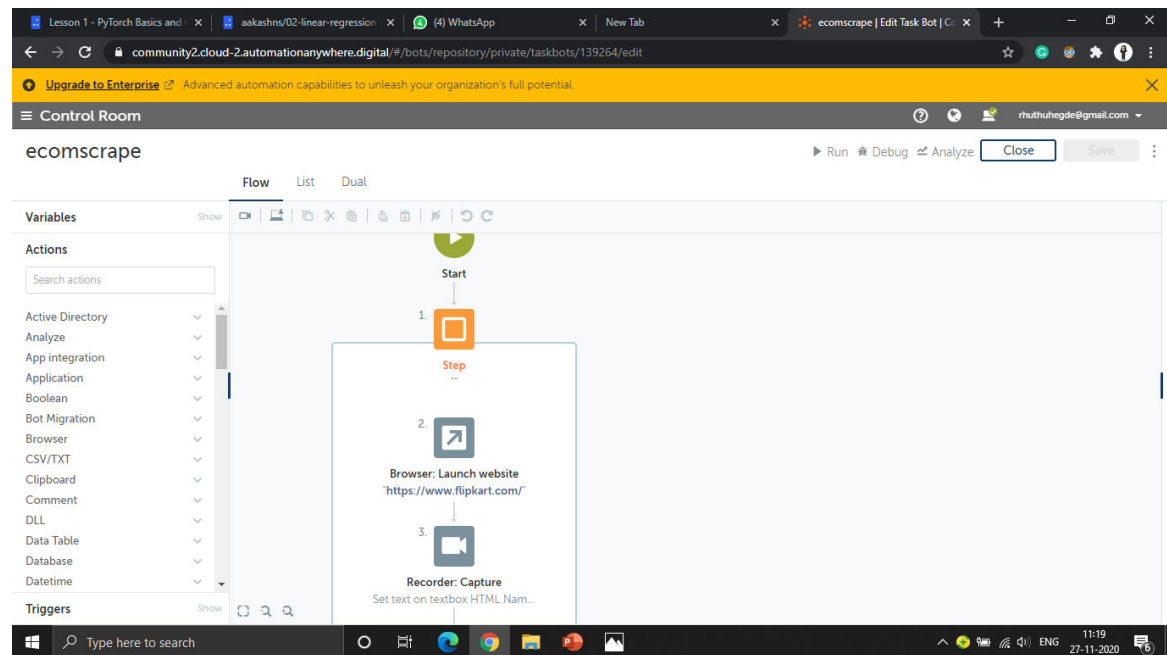


Fig.3.1.3 This shows launching and capturing textbox/search bar from the website.

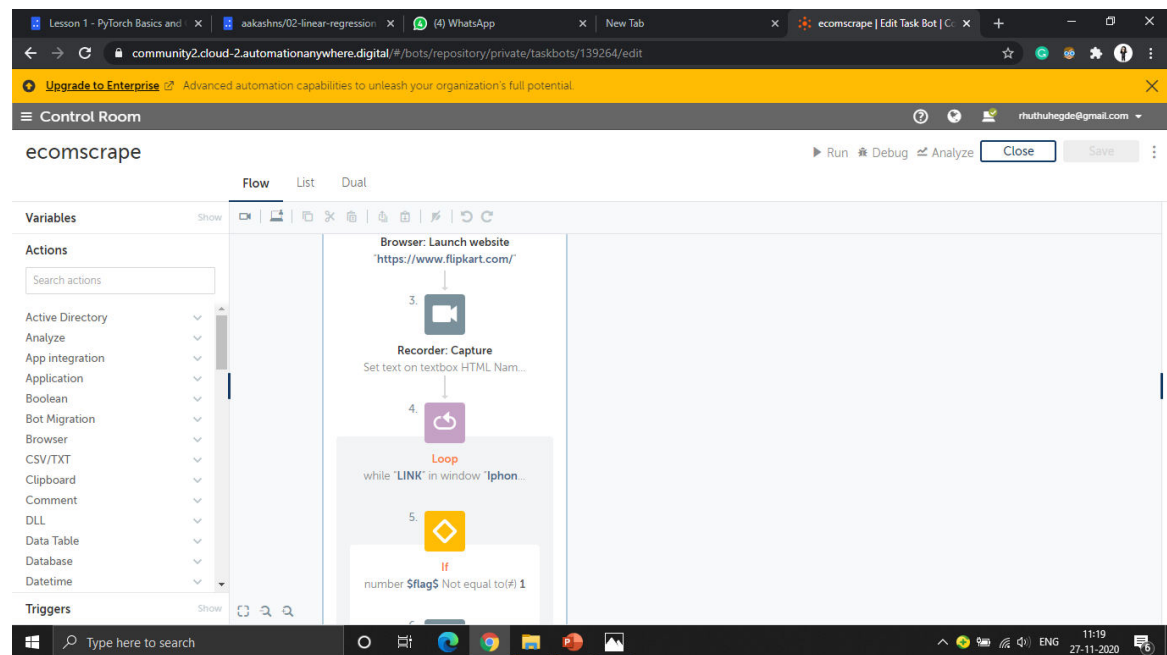


Fig 3.1.4 This shows loop where it captures NEXT button and checks the condition using if statement.

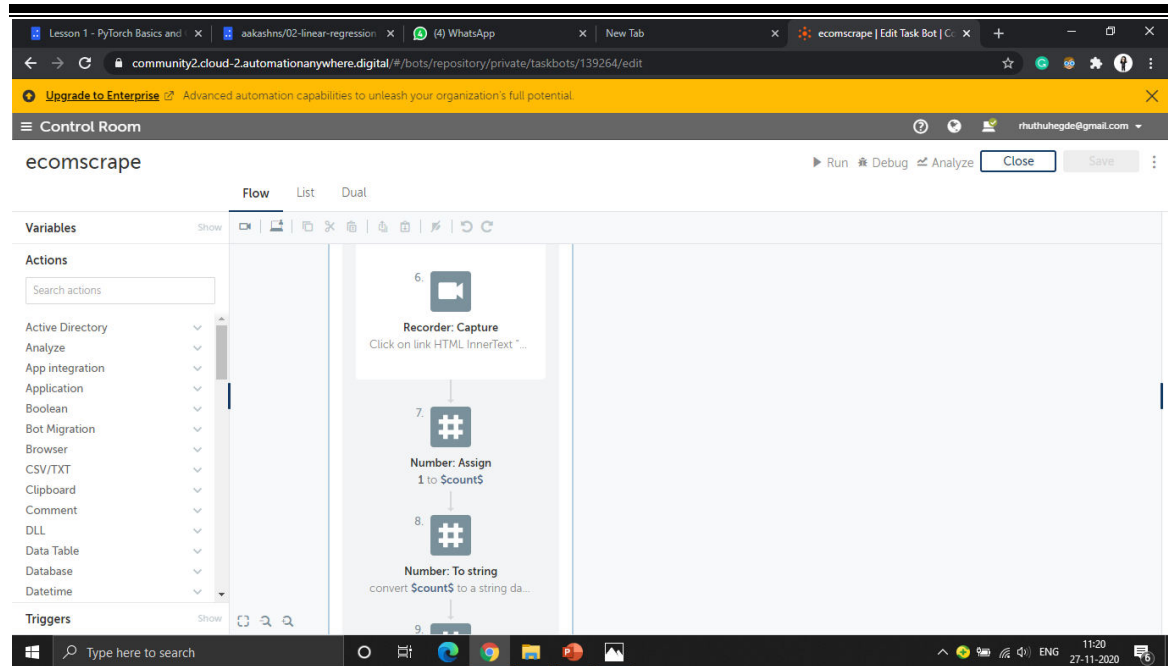


Fig 3.1.5 The count variable is assigned so that we can keep track of capturing different objects on a particular page

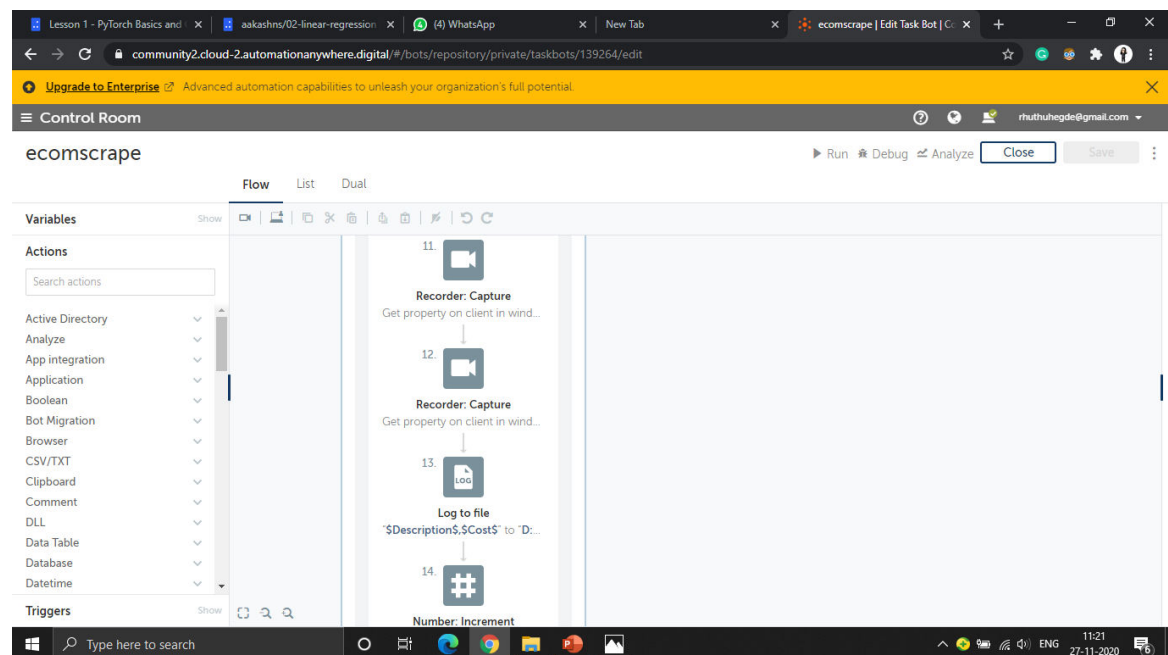
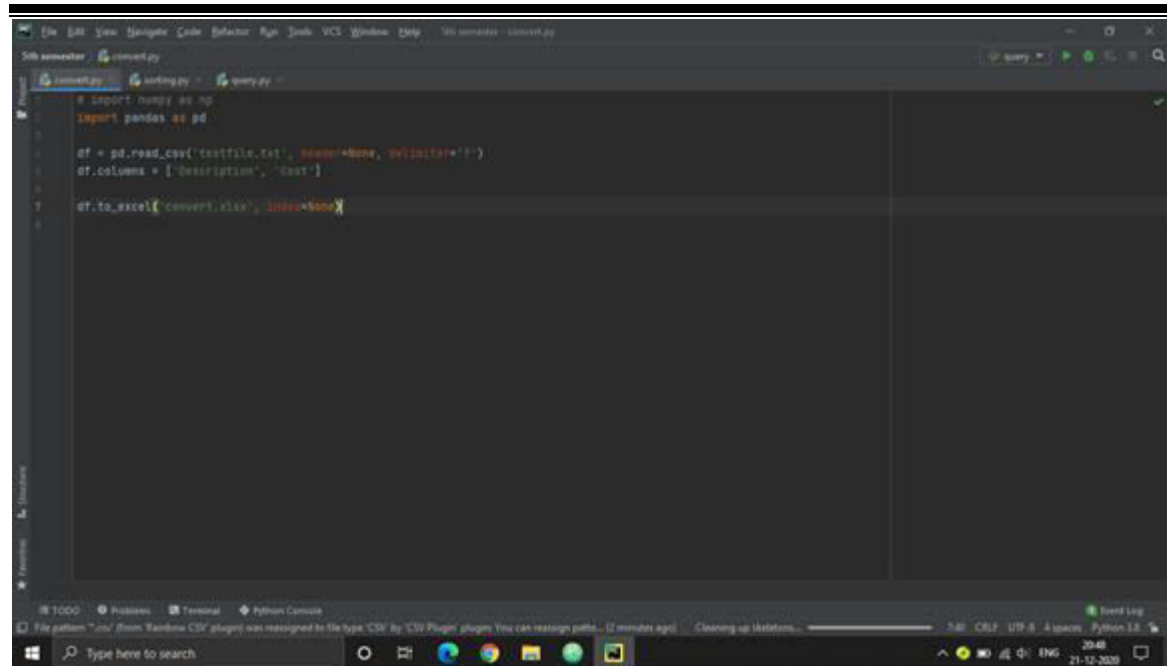


Fig 3.1.6 After capturing description and cost we store it tin a log file

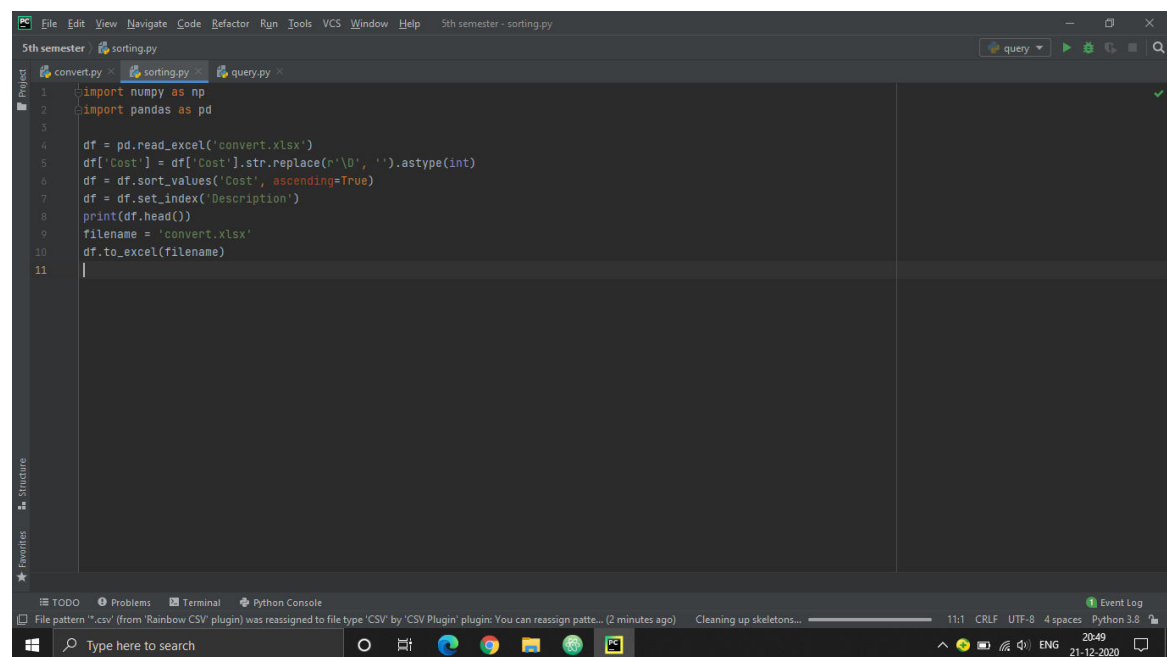


```

1 import numpy as np
2 import pandas as pd
3
4 df = pd.read_csv('testfile.txt', header=None, delimiter=';')
5 df.columns = ['Description', 'Cost']
6
7 df.to_excel('convert.xlsx', index=None)

```

Fig 3.1.7: Python file which converts text file to excel file



```

1 import numpy as np
2 import pandas as pd
3
4 df = pd.read_excel('convert.xlsx')
5 df['Cost'] = df['Cost'].str.replace(r'\D', '').astype(int)
6 df = df.sort_values('Cost', ascending=True)
7 df = df.set_index('Description')
8 print(df.head())
9 filename = 'convert.xlsx'
10 df.to_excel(filename)
11

```

Fig. 3.1.8 The python file which sorts our data.

Dataset:

This is the scraped data from the e-commerce websites.

```

testfile - Notepad
File Edit Format View Help
Apple iPhone 12 (Black, 64 GB) ₹79,900
Apple iPhone 12 (Black, 64 GB) ₹79,900
Apple iPhone 12 (Blue, 64 GB) ₹79,900
Apple iPhone 12 (White, 64 GB) ₹79,900
Apple iPhone 12 (White, 128 GB) ₹84,900
Apple iPhone 12 (Green, 64 GB) ₹79,900
Apple iPhone 12 (Red, 64 GB) ₹79,900
Apple iPhone 12 (Green, 128 GB) ₹84,900
Apple iPhone 12 (Black, 128 GB) ₹84,900
Apple iPhone 12 (Red, 128 GB) ₹84,900
Apple iPhone 12 (White, 256 GB) ₹94,900
Apple iPhone 12 (Blue, 256 GB) ₹94,900
Apple iPhone XR ((PRODUCT)RED, 64 GB) (Includes EarPods, Power Adapter) ₹44,900
Apple iPhone 12 (Green, 256 GB) ₹94,900
Apple iPhone 12 (Black, 256 GB) ₹94,900
Apple iPhone 12 (Red, 256 GB) ₹94,900
Apple iPhone 12 Mini (Blue, 64 GB) ₹69,900
Apple iPhone 12 Pro (Pacific Blue, 128 GB) ₹1,19,900
Apple iPhone 12 Mini (Black, 64 GB) ₹69,900
Apple iPhone 12 Mini (White, 64 GB) ₹69,900
Apple iPhone 12 Pro Max (Silver, 256 GB) ₹1,39,900
Apple iPhone 12 Mini (Blue, 128 GB) ₹74,900
Apple iPhone 12 Mini (Black, 128 GB) ₹74,900
Apple iPhone 12 Pro (Pacific Blue, 256 GB) ₹1,29,900
Apple iPhone 12 Pro Max (Graphite, 512 GB) ₹1,59,900
Apple iPhone 12 Pro Max (Graphite, 512 GB) ₹1,59,900
Apple iPhone 12 Mini (Green, 64 GB) ₹69,900
Apple iPhone 12 Pro Max (Silver, 512 GB) ₹1,59,900
Apple iPhone 12 Pro Max (Pacific Blue, 128 GB) ₹1,29,900
Apple iPhone 12 Pro Max (Graphite, 128 GB) ₹1,29,900
Apple iPhone 12 Mini (Red, 64 GB) ₹69,900
Apple iPhone 12 Pro Max (Pacific Blue, 256 GB) ₹1,39,900
Apple iPhone 12 Pro (Graphite, 128 GB) ₹1,19,900
Apple iPhone 12 Mini (White, 128 GB) ₹74,900
Apple iPhone 12 Pro Max (Graphite, 256 GB) ₹1,39,900
Apple iPhone 12 Mini (Red, 128 GB) ₹74,900
Ln 1, Col 1 100% Windows (CRLF) UTF-8
  
```

Fig.3.1.9: This is the excel file or the data set after the converting the log file and sorting process

Description	Cost
Samsung Galaxy Star Advance (Black, 4 GB)	3500
Samsung Galaxy Star (Ceramic White, 4 GB)	3599
Samsung Galaxy Star Pro (White, 4 GB)	4199
Samsung Galaxy Star (Noble Black, 4 GB)	4199
Samsung Galaxy Star Pro (Midnight Black, 4 GB)	4299
Moto C (Starry Black, 16 GB)	4399
Samsung Galaxy Star 2 (Iris Charcoal, 4 GB)	4890
Samsung Galaxy Star 2 (White, 4 GB)	4890
Samsung Galaxy J1 Ace (White, 4 GB)	4988
Samsung Galaxy J1 Ace (Black, 4 GB)	4988
Moto E (2nd Gen) 3G (White, 8 GB)	4999
Samsung Galaxy Trend (Midnight Black, 4 GB)	4999
Samsung Galaxy Grand Neo Plus (White, 8 GB)	4999
Samsung Galaxy Trend (Ceramic White, 4 GB)	4999
Moto E (1st Gen) (Black, 4 GB)	4999
Samsung Galaxy Ace NXT (Black, 4 GB)	5000
Samsung Galaxy Pocket Neo (Metallic Silver, 4 GB)	5000
Moto G5 (Lunar Grey, 16 GB)	5450
Samsung Galaxy Core Prime (Silver, 8 GB)	5490
Samsung Galaxy Ace NXT (Ceramic White, 4 GB)	5490
Samsung Galaxy Core Prime (Charcoal Grey, 8 GB)	5490
Samsung Galaxy Star Advance (White, 4 GB)	5499

Fig.3.1.10: This is the excel file or the data set after the converting the log file and sorting process

CHAPTER 4

Conclusions and Future Enhancements

4.1 Benefits of the Automation Anywhere

The cloud version of Enterprise A2019 offers numerous advantages. Here are six of the most important:

1. **Lower total cost of ownership (TCO):** With Enterprise A2019, Automation Anywhere reduces the barrier to automation by hosting and managing the underlying cloud infrastructure. You no longer have to make huge upfront capex investments to build and maintain your own on-site IT infrastructure.

Instead, you move to a predictable opex model and get started right away. This significantly drives down your total cost of ownership.
2. **Easy access:** Without the need to download and install software on your local machine, you can access the RPA application from virtually anywhere, on any device, with just a web browser and internet connectivity. This allows you to move from an application ownership model to an application access model, further reducing your TCO by eliminating the need to upgrade and maintain local machines.
3. **Ease of use:** Automating a business process is as easy as logging in, clicking, and automating. A highly intuitive modern interface optimized for every user — with features such as built-in product learning, drag-and-drop AI actions, different design views, and more — makes designing automation workflows extremely easy.
4. **Fast time to value:** You no longer need to wait months or years before realizing any value from your initial RPA investment. With Enterprise A2019, you get instant value. You can log in, set up automation, and go live within minutes, giving you immediate return on your investment.

5. **High scalability:** Because Enterprise A2019 is a highly scalable application, you can start small by deploying a few bots and automating some business processes. As more business processes need to be automated, you can easily scale up (to unlimited bots) by deploying additional bots — without having to worry about additional computing resources. Once done automating, you can just as easily scale down by spinning down the bots.
6. **Business agility:** Automation Anywhere updates Enterprise A2019 every four weeks, which means you'll receive automatic and regular updates to the latest software version. This will help you drive new innovative projects, launch solutions to market faster, and maintain a competitive edge in the marketplace.

4.2 Applications of Web scraping

The applications of the web scraping are:

1. Scrape product details (price, images, rating, reviews etc.) from retailer/manufacturer/eCommerce websites (Ex: Amazon, eBay, AliExpress, Alibaba etc.) to show on own websites, to provide price comparisons, to perform a price watch on competing sellers etc.
2. Scrape property details and agent contact details from real estate websites.
3. Scrape contact details of businesses as well as individuals from yellow pages websites.
4. Scrape hospital/clinic websites to build a catalog of health physicians including their contact details.

4.3 Future Enhancements

We intend to add multiple websites in the same bot.

We have been working on dynamic websites like DELL website and also working on websites where products are listed horizontally.

Furthermore, we intend to compare and bring out the best deal by comparing the data extracted from multiple websites.

REFERENCES

1. <https://mindmajix.com/30-rpa-examples>
2. <https://docs.automationanywhere.com/bundle/enterprise-v11.3/page/enterprise/topics/aae-client/bot-creator/creating-an-automation-task/extracting-data-from-websites.html>
3. <https://docs.automationanywhere.com/bundle/enterprise-v11.3/page/enterprise/topics/aae-client/bot-creator/creating-an-automation-task/extracting-pattern-based-data.html>
4. <https://docs.automationanywhere.com/bundle/enterprise-v2019/page/enterprise-cloud/topics/aae-client/bot-creator/commands/cloud-python-command.html>
5. <https://www.flipkart.com/>