Vivekanand Education Society's Institute Of Technology
Department Of Information Technology
Project Exam-Sem IV

Sem: IV Academic Year : 2021-22

Title: ScrapSonic

Mentor Name:
Mrs. Sukanya Roychowdhury

Group Members:
Member 1 : TEJAS ROKADE (46)
Member 2: PRANAV RAIKAR (43)
Member 3: GAURANG RAORANE(45)
Member 4: ANURAG GAIWAL(15)

Content

- Introduction to the project
- Problem Statement
- Objectives of the project
- Requirements of the system (Hardware, software)
- ER diagram of the proposed system
- Front End
- Implementation
- Gantt Chart
- Conclusion
- References

Introduction

What is webscraping?

- Web scraping, also known as web harvesting or data scraping, is the automated process of extracting data from websites.
- It involves using computer programs or scripts to access and extract data from web pages, typically in large quantities, and then saving that data to a local file or database for further analysis or processing.
- Web scraping can be used for a variety of purposes, such as data mining, market research, price comparison, and content aggregation.

Introduction to project

- In our project, first we register our new user and after the registration, the user can use our project to extract the information of his/her desired product.
- User can then log in with his username and password. (We're using the LoginID as the primary key for login.)
- Users have to enter the name or category of the product and the number of pages to scrape.
- Then our project will generate the url by filling the product name in the base url and will collect the data from the appropriate tags and class of web page's HTML code.
- All this information will be displayed back on the GUI and proper option for analyzing the information will be provided.

Problem Statement

- Comparing the prices of different products can be a hectic task.
- It is a time taking process of searching for items online for right price, seller and to see description for every other item and then decide is time taking and hard to predict so through this web scraping automation it can be solved in an instant.
- So to make it easy for you to search for multiple items and compare prices.

Objectives of project

- Businesses scrap product reviews. Some use it to improve their products, while others monitor their competition. Whatever the reason, scraping product reviews can provide valuable insights into customer sentiment.
- To serve the customers better, businesses need to be aware of their feedback. By collecting and analyzing the customers' reviews, businesses can know about insightful trends of customers and fine-tune their products and services accordingly.
- Through our software customers will be able to see the reviews of their favourite books at one place and in more structured and descriptive way.

Requirements of the system (Hardware)

- Core i5 and above.
- Processor: Minimum 1 Ghz processor; (Recommended 2GHz or more).
- Hard Drive: Minimum 32 GB; (Recommended 64GB or more).
- Memory (RAM): Minimum 1 GB; (Recommended 4 GB or above).

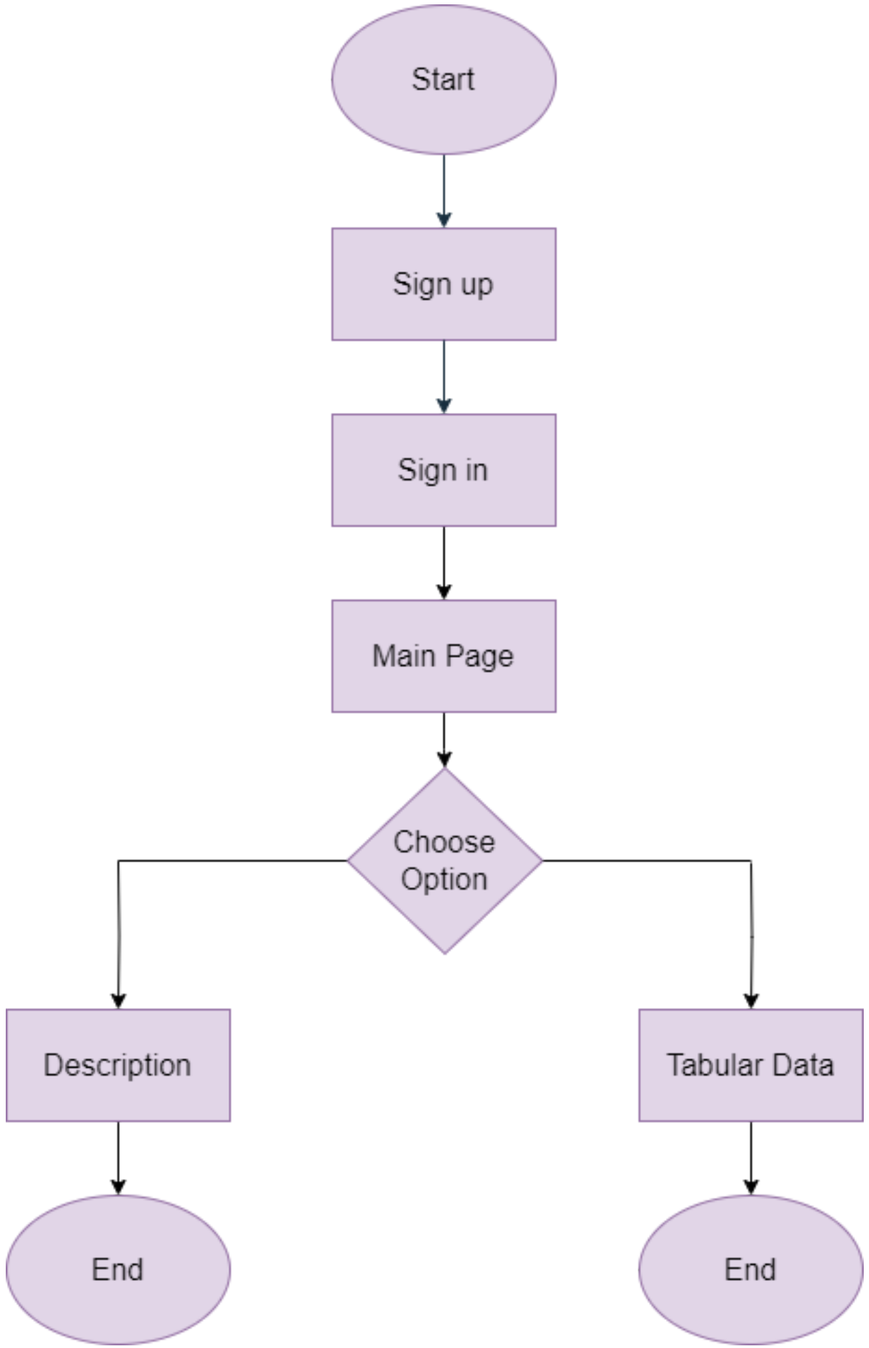
Requirements of the system (Software)

- Windows: 7 or newer
- MAC: OS X v10.7 or higher
- Linux: Ubuntu
- Any Python IDE.

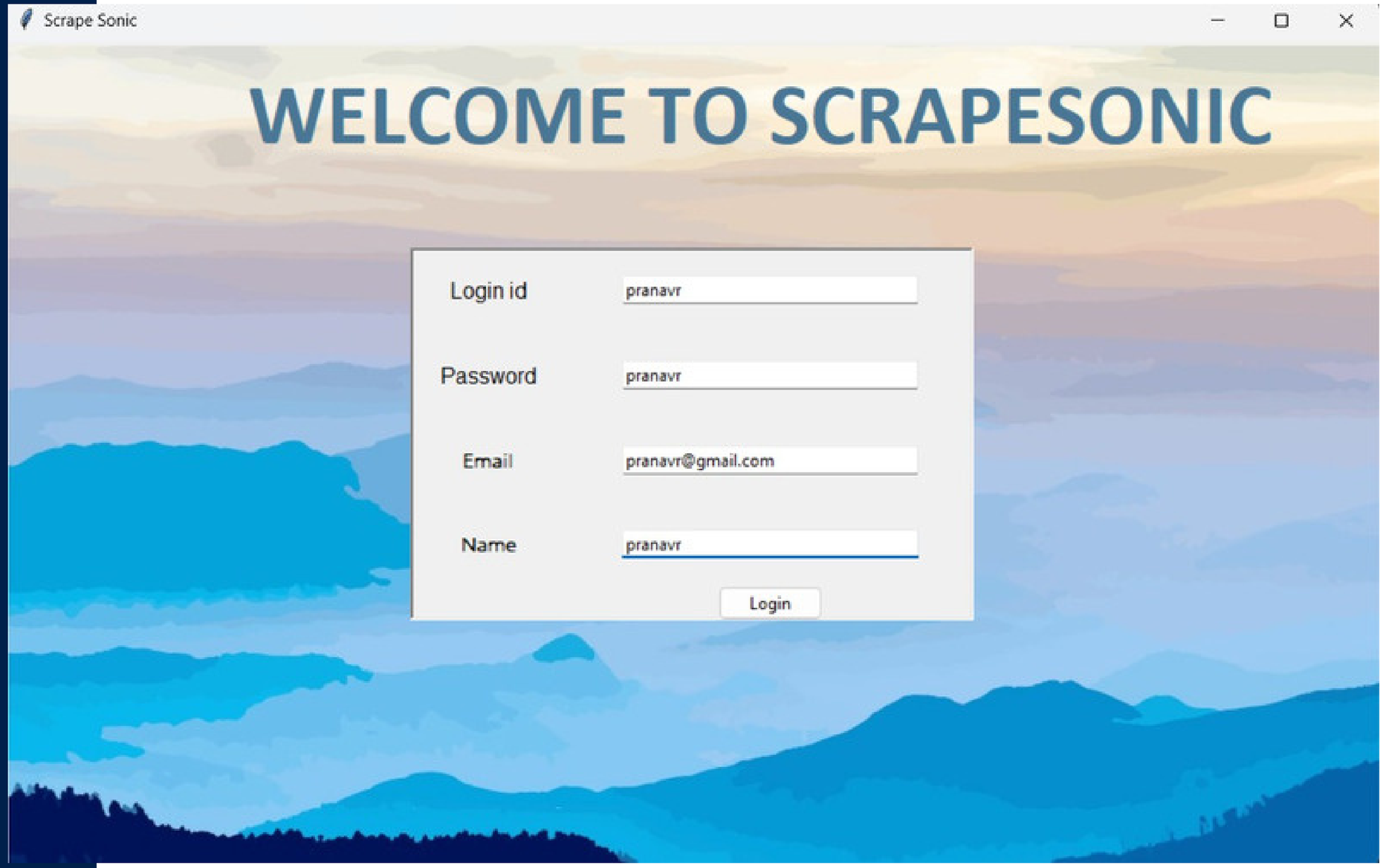
Python packages installed -

- PIL
- Pandas
- Tkinter
- pandastable
- Requests

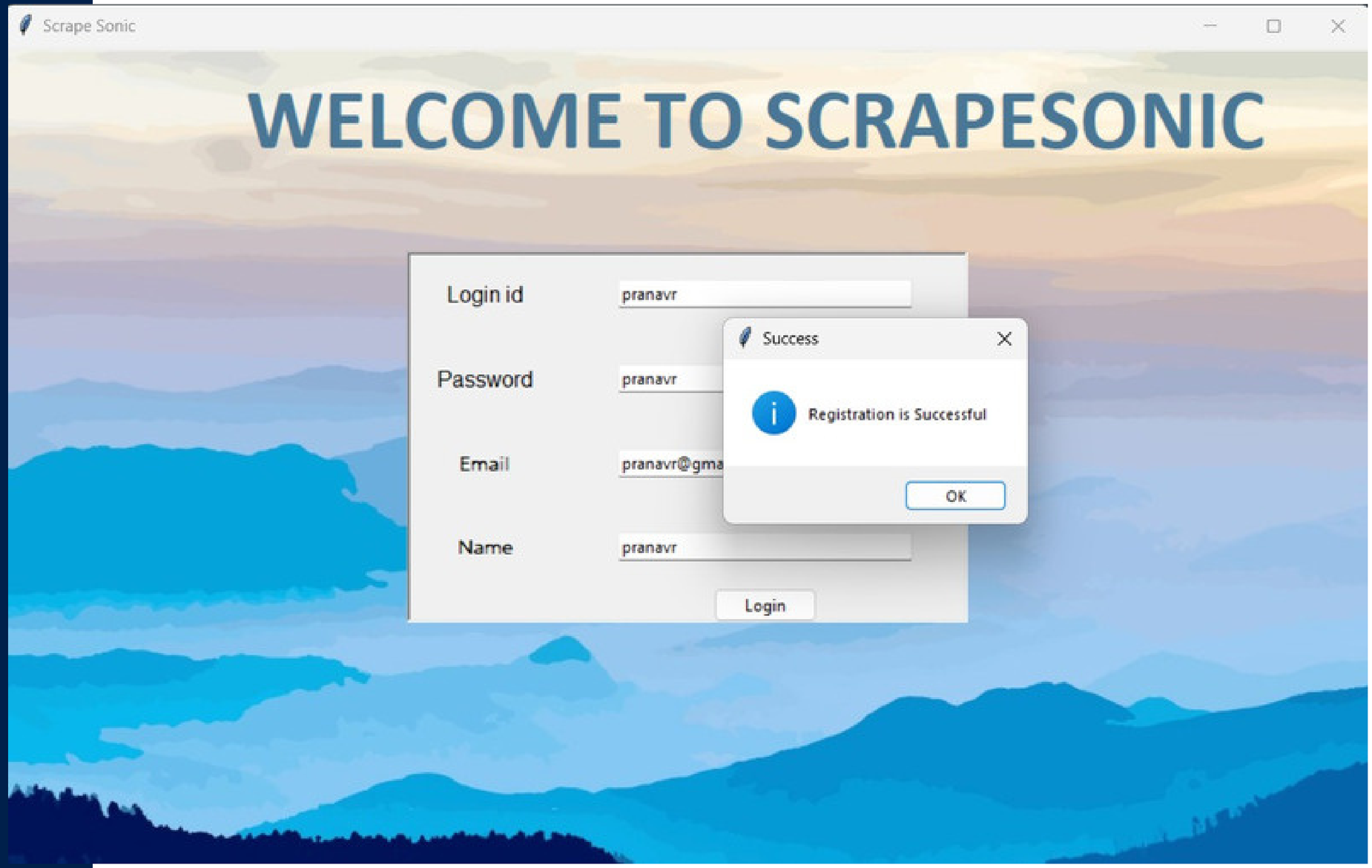
Flow Chart


Start
Sign up
Sign in
Main Page
Choose Option
Description
Tabular Data
End
End

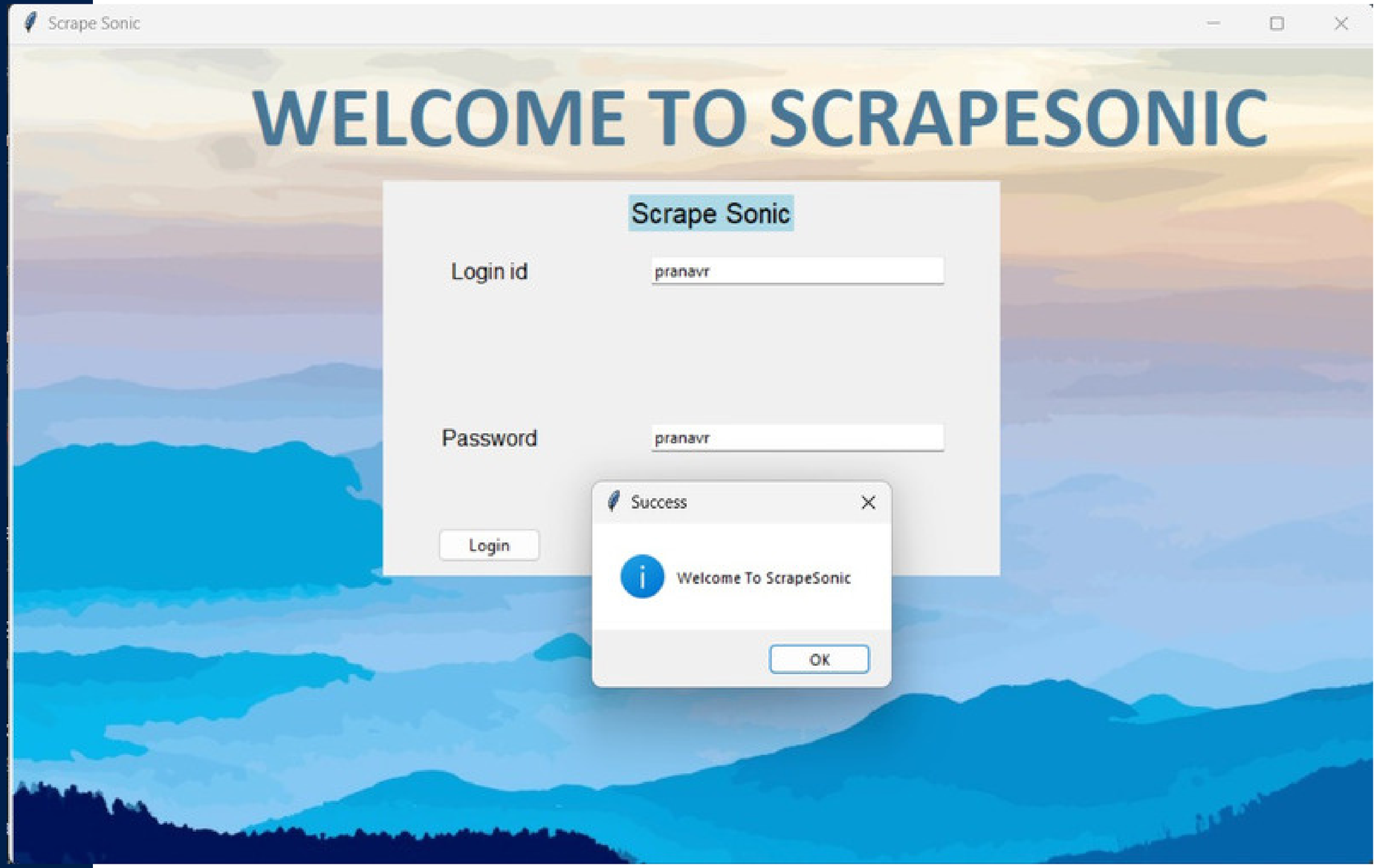
Front End


Scrape Sonic
WELCOME TO SCRAPESONIC
Login id
pranavr
Password
pranavr
Email
pranavr@gmail.com
Name
pranavr
Login

Front End


Scrape Sonic
WELCOME TO SCRAPESONIC
Login id
pranavr
Password
pranavr
Email
pranavr@gma
Name
pranavr
Login
Success
Registration is Successful
OK

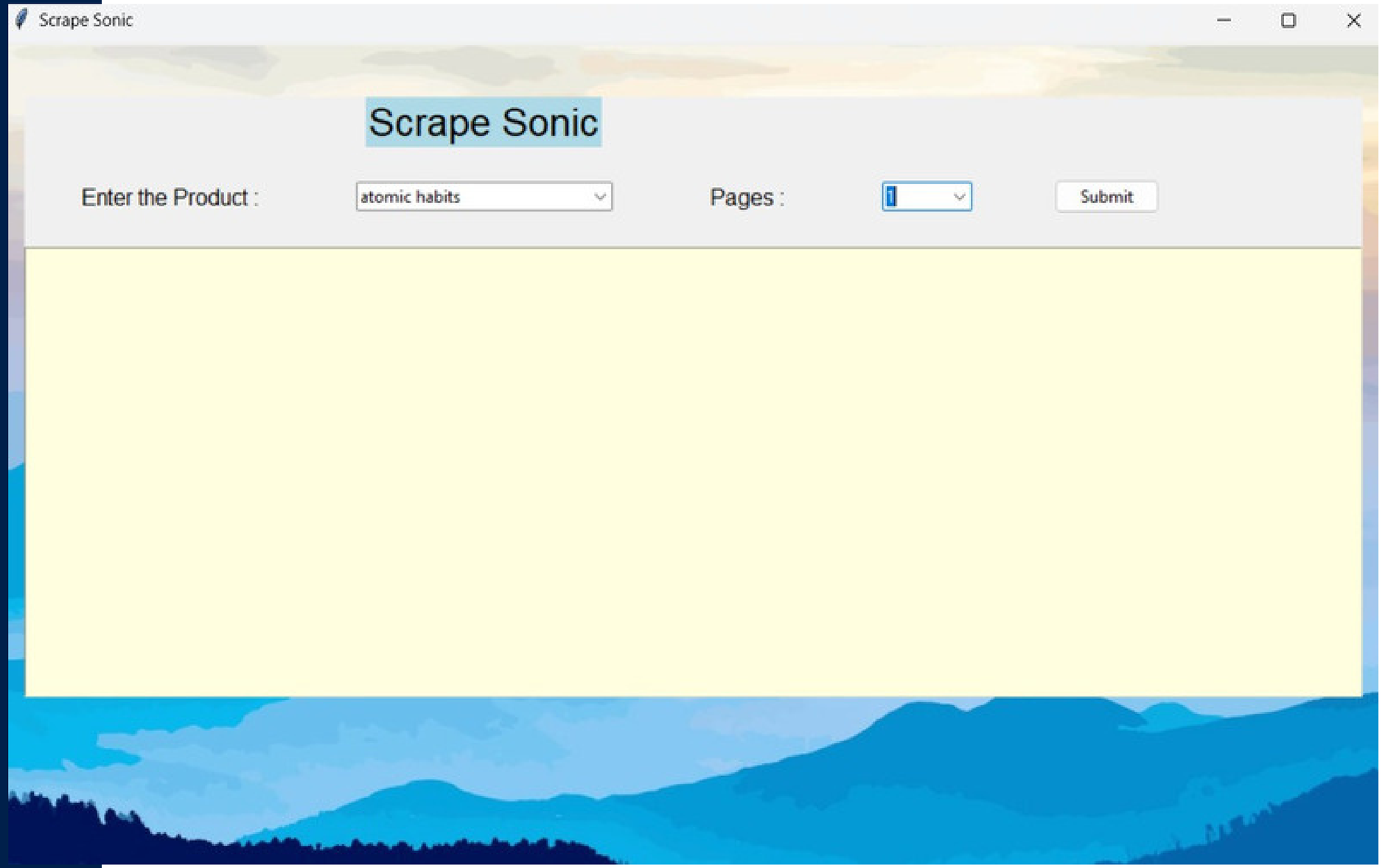
Front End


Scrape Sonic
WELCOME TO SCRAPESONIC
Scrape Sonic
Login id
pranavr
Password
pranavr
Login
Success
Welcome To ScrapeSonic
OK

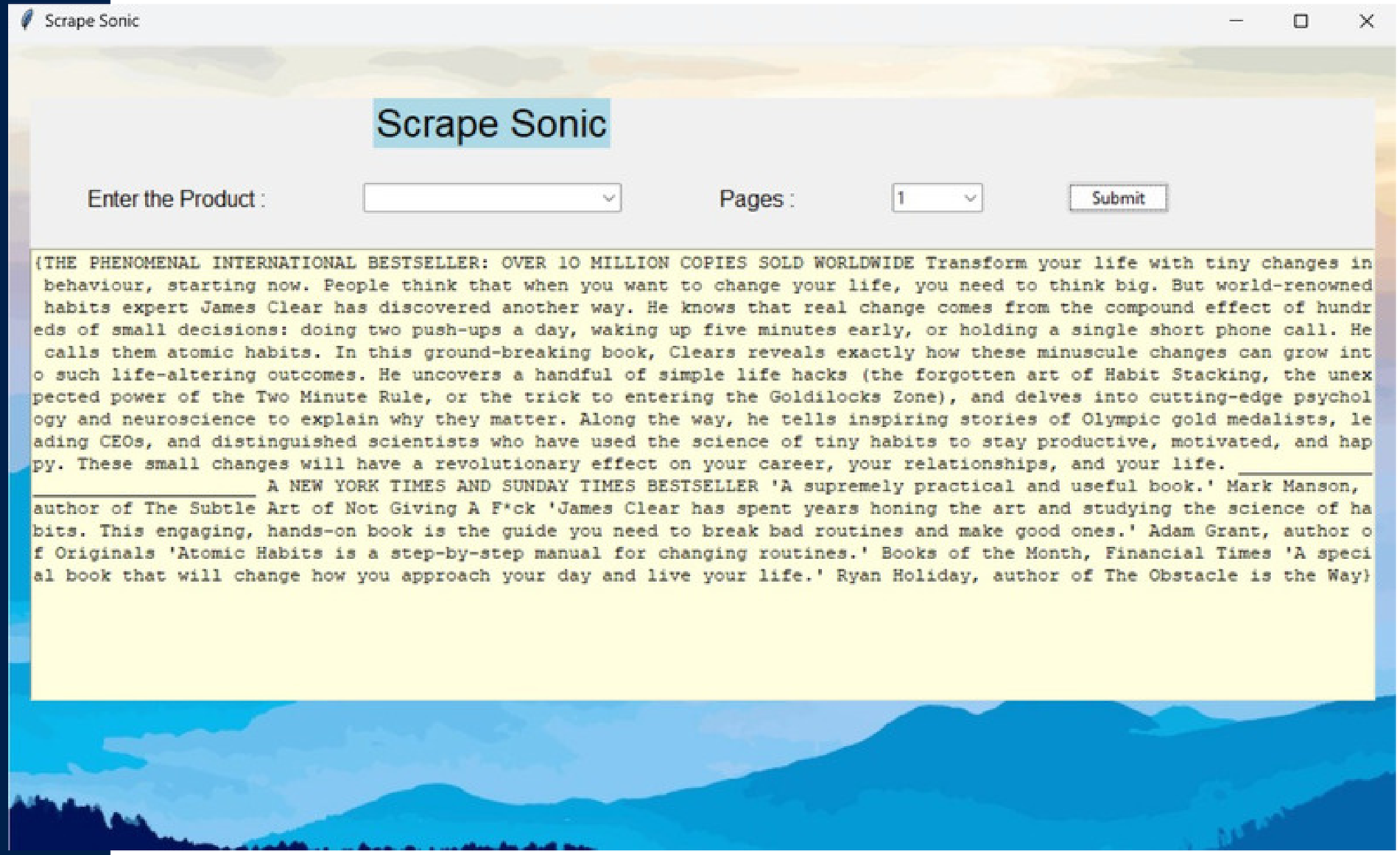
Front End


Scrape Sonic
WELCOME TO SCRAPESONIC
DESCRIPTION
TABULARDATA

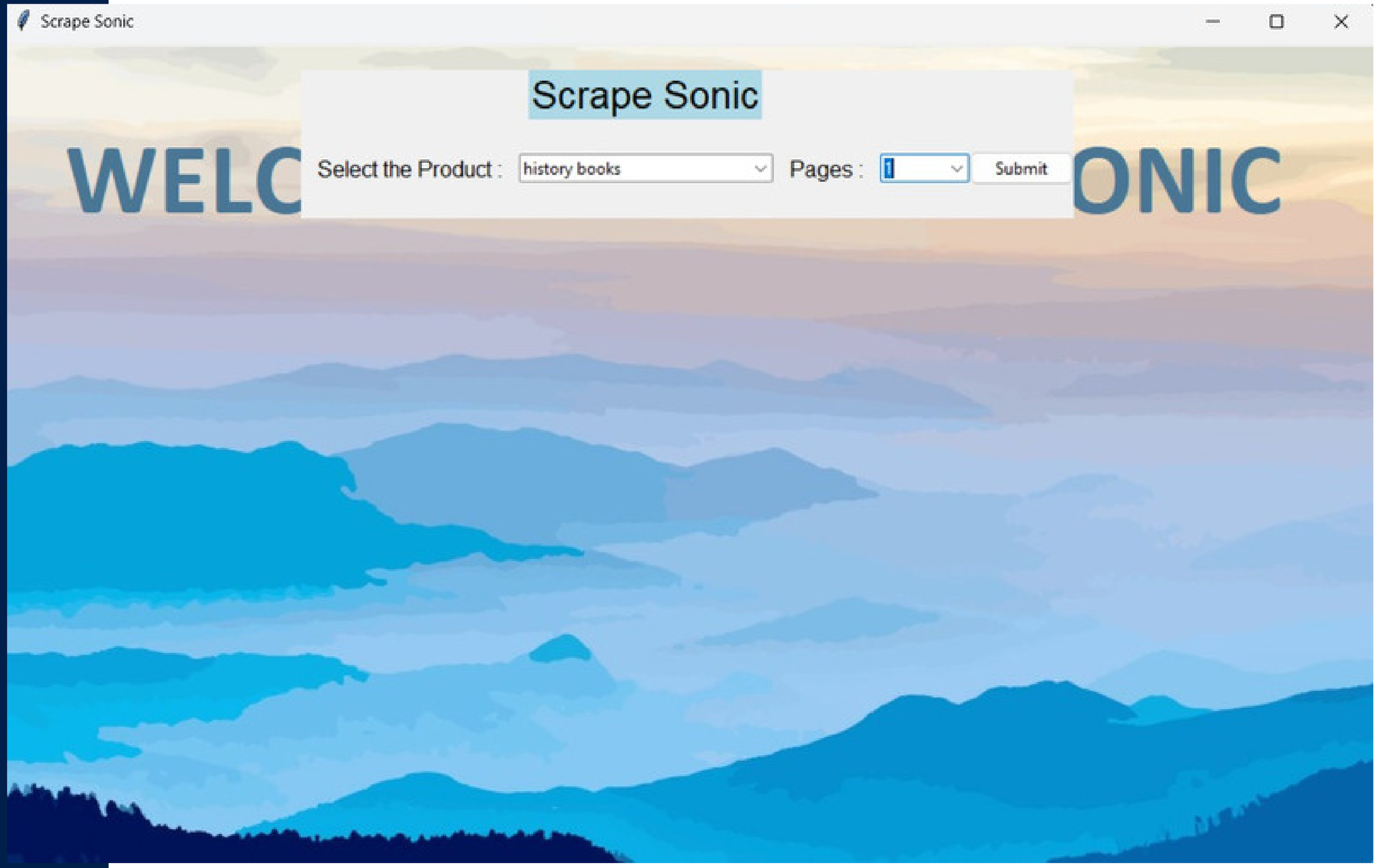
Front End


Scrape Sonic
Scrape Sonic
Enter the Product :
atomic habits
Pages :
1
Submit

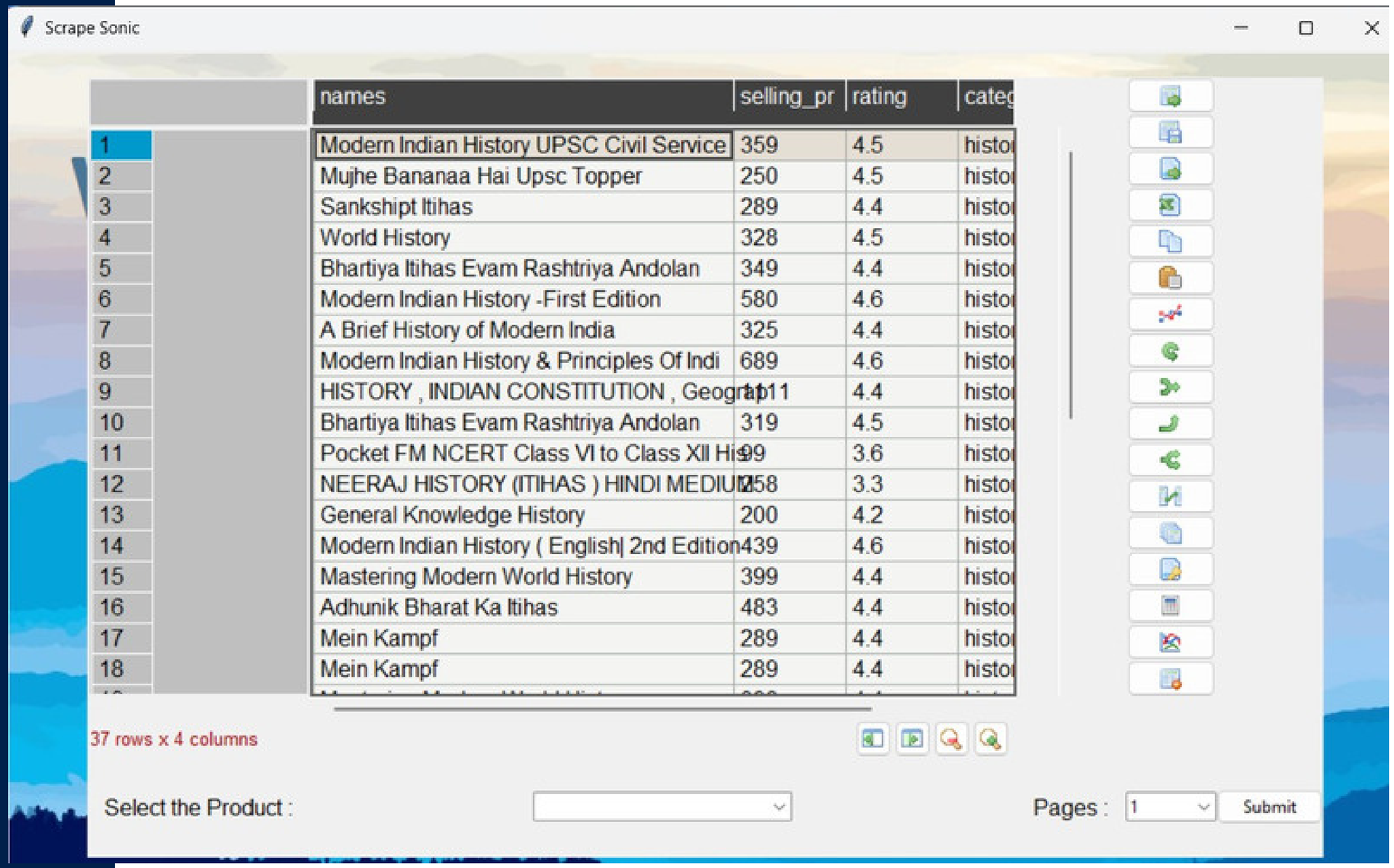
Front End


Scrape Sonic
Scrape Sonic
Enter the Product :
Pages :
1
Submit
{THE PHENOMENAL INTERNATIONAL BESTSELLER: OVER 10 MILLION COPIES SOLD WORLDWIDE Transform your life with tiny changes in behaviour, starting now. People think that when you want to change your life, you need to think big. But world-renowned habits expert James Clear has discovered another way. He knows that real change comes from the compound effect of hundreds of small decisions: doing two push-ups a day, waking up five minutes early, or holding a single short phone call. He calls them atomic habits. In this ground-breaking book, Clears reveals exactly how these minuscule changes can grow into such life-altering outcomes. He uncovers a handful of simple life hacks (the forgotten art of Habit Stacking, the unexpected power of the Two Minute Rule, or the trick to entering the Goldilocks Zone), and delves into cutting-edge psychology and neuroscience to explain why they matter. Along the way, he tells inspiring stories of Olympic gold medalists, leading CEOs, and distinguished scientists who have used the science of tiny habits to stay productive, motivated, and happy. These small changes will have a revolutionary effect on your career, your relationships, and your life. _____ _____ A NEW YORK TIMES AND SUNDAY TIMES BESTSELLER 'A supremely practical and useful book.' Mark Manson, author of The Subtle Art of Not Giving A F*ck 'James Clear has spent years honing the art and studying the science of habits. This engaging, hands-on book is the guide you need to break bad routines and make good ones.' Adam Grant, author of Originals 'Atomic Habits is a step-by-step manual for changing routines.' Books of the Month, Financial Times 'A special book that will change how you approach your day and live your life.' Ryan Holiday, author of The Obstacle is the Way}

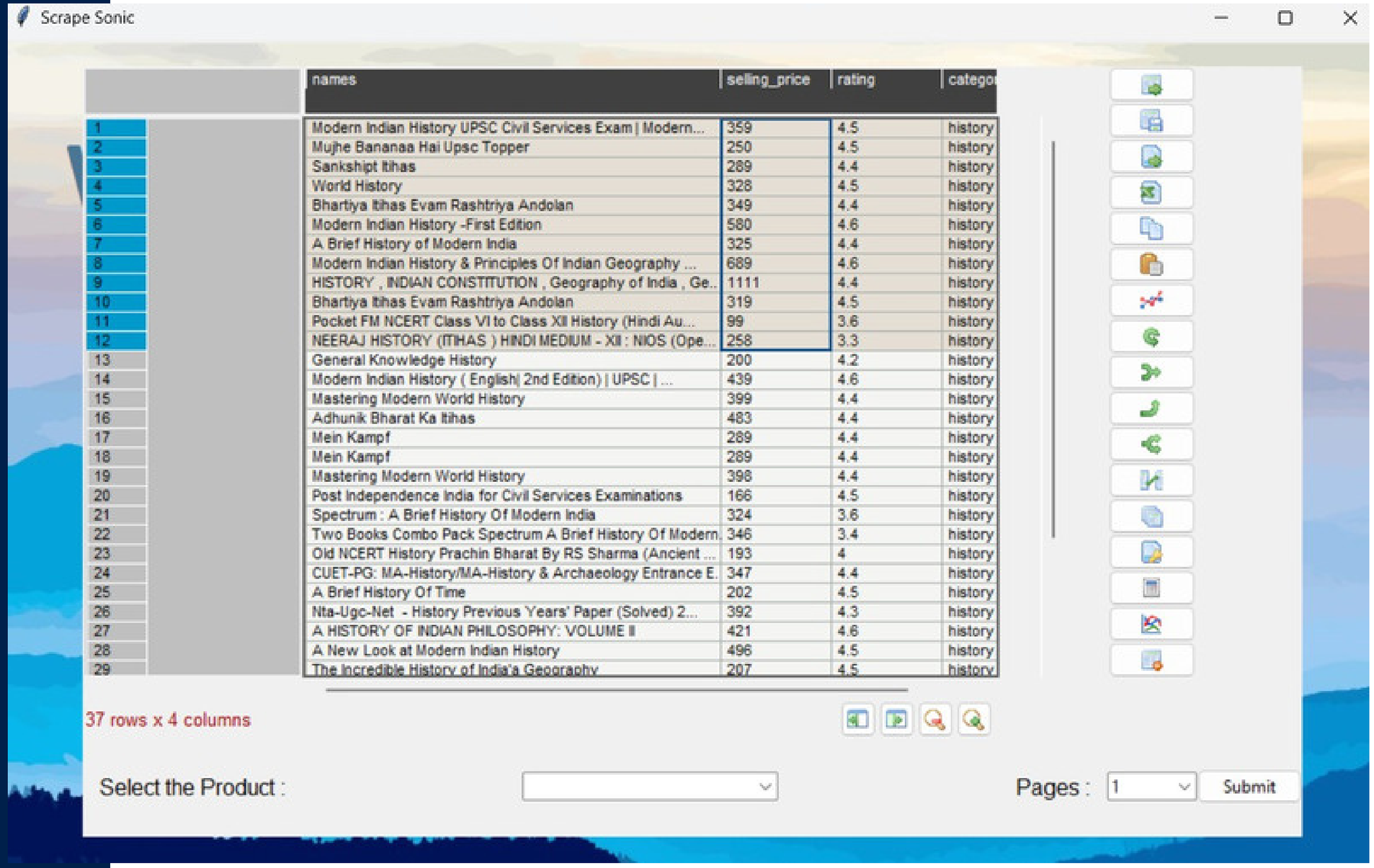
Front End


Scrape Sonic
Scrape Sonic
WELC
Select the Product :
history books
Pages :
1
Submit
ONIC

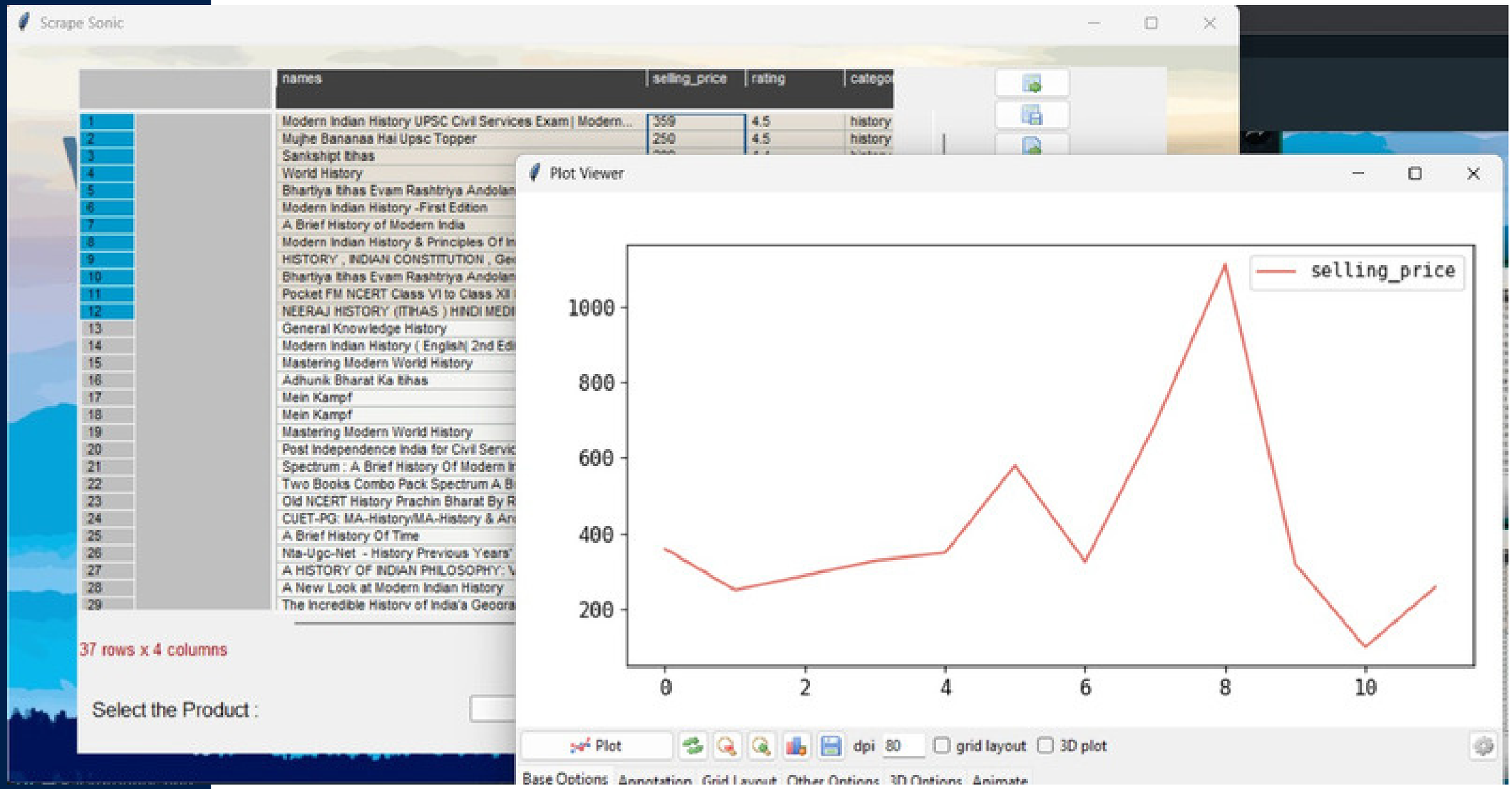
Front End


Scrape Sonic
names
selling_pr
rating
categ
1 Modern Indian History UPSC Civil Service 359 4.5 histo
2 Mujhe Bananaa Hai Upsc Topper 250 4.5 histo
3 Sankshipt Itihas 289 4.4 histo
4 World History 328 4.5 histo
5 Bhartiya Itihas Evam Rashtriya Andolan 349 4.4 histo
6 Modern Indian History -First Edition 580 4.6 histo
7 A Brief History of Modern India 325 4.4 histo
8 Modern Indian History & Principles Of Indi 689 4.6 histo
9 HISTORY , INDIAN CONSTITUTION , Geogra 1011 4.4 histo
10 Bhartiya Itihas Evam Rashtriya Andolan 319 4.5 histo
11 Pocket FM NCERT Class VI to Class XII His 99 3.6 histo
12 NEERAJ HISTORY (ITIHAS) HINDI MEDIUM 258 3.3 histo
13 General Knowledge History 200 4.2 histo
14 Modern Indian History (English| 2nd Edition 439 4.6 histo
15 Mastering Modern World History 399 4.4 histo
16 Adhunik Bharat Ka Itihas 483 4.4 histo
17 Mein Kampf 289 4.4 histo
18 Mein Kampf 289 4.4 histo
37 rows x 4 columns
Select the Product :
Pages :
1
Submit

Front End


Scrape Sonic
names
selling_price
rating
categor
1 Modern Indian History UPSC Civil Services Exam | Modern... 359 4.5 history
2 Mujhe Bananaa Hai Upsc Topper 250 4.5 history
3 Sankshipt Itihas 289 4.4 history
4 World History 328 4.5 history
5 Bhartiya Itihas Evam Rashtriya Andolan 349 4.4 history
6 Modern Indian History -First Edition 580 4.6 history
7 A Brief History of Modern India 325 4.4 history
8 Modern Indian History & Principles Of Indian Geography ... 689 4.6 history
9 HISTORY , INDIAN CONSTITUTION , Geography of India , Ge.. 1111 4.4 history
10 Bhartiya Itihas Evam Rashtriya Andolan 319 4.5 history
11 Pocket FM NCERT Class VI to Class XII History (Hindi Au... 99 3.6 history
12 NEERAJ HISTORY (ITIHAS) HINDI MEDIUM - XII : NIOS (Ope... 258 3.3 history
13 General Knowledge History 200 4.2 history
14 Modern Indian History (English| 2nd Edition) | UPSC | ... 439 4.6 history
15 Mastering Modern World History 399 4.4 history
16 Adhunik Bharat Ka Itihas 483 4.4 history
17 Mein Kampf 289 4.4 history
18 Mein Kampf 289 4.4 history
19 Mastering Modern World History 398 4.4 history
20 Post Independence India for Civil Services Examinations 166 4.5 history
21 Spectrum : A Brief History Of Modern India 324 3.6 history
22 Two Books Combo Pack Spectrum A Brief History Of Modern. 346 3.4 history
23 Old NCERT History Prachin Bharat By RS Sharma (Ancient ... 193 4 history
24 CUET-PG: MA-History/MA-History & Archaeology Entrance E. 347 4.4 history
25 A Brief History Of Time 202 4.5 history
26 Nta-Ugc-Net - History Previous Years' Paper (Solved) 2... 392 4.3 history
27 A HISTORY OF INDIAN PHILOSOPHY: VOLUME II 421 4.6 history
28 A New Look at Modern Indian History 496 4.5 history
29 The Incredible History of India'a Geography 207 4.5 histor
37 rows x 4 columns
Select the Product :
Pages :
1
Submit

Front End


Scrape Sonic
names
selling_price
rating
1 Modern Indian History UPSC Civil Services Exam | Modern... 359 4.5 history
2 Mujhe Bananaa Hai Upsc Topper 250 4.5 history
3 Sankshipt Itihas
4 World History
5 Bhartiya Itihas Evam Rashtriya Andolan
6 Modern Indian History -First Edition
7 A Brief History of Modern India
8 Modern Indian History & Principles Of In
9 HISTORY , INDIAN CONSTITUTION , Ge
10 Bhartiya Itihas Evam Rashtriya Andolan
11 Pocket FM NCERT Class VI to Class XII
12 NEERAJ HISTORY (ITIHAS) HINDI MEDI
13 General Knowledge History
14 Modern Indian History (English| 2nd Edi
15 Mastering Modern World History
16 Adhunik Bharat Ka Itihas
17 Mein Kampf
18 Mein Kampf
19 Mastering Modern World History
20 Post Independence India for Civil Servic
21 Spectrum : A Brief History Of Modern I
22 Two Books Combo Pack Spectrum A B
23 Old NCERT History Prachin Bharat By R
24 CUET-PG: MA-History/MA-History & An
25 A Brief History Of Time
26 Nta-Ugc-Net - History Previous Years'
27 A HISTORY OF INDIAN PHILOSOPHY: V
28 A New Look at Modern Indian History
29 The Incredible Historv of India'a Geogra
37 rows x 4 columns
Select the Product :
Plot Viewer
selling_price
1000
800
600
400
200
0
2
4
6
8
10
Plot
dpi 80
grid layout
3D plot
Base Options Annotation Grid Layout Other Options 3D Options Animate

Implementation

Web scraping tool

We used the Beautiful Soup library, along with the Requests library for making HTTP requests to the Flipkart website, to extract product information from the website.

Implementation

Data extraction

To extract the desired data from Flipkart, we inspected the website's HTML structure and identified the relevant tags and attributes for each piece of information. We then used Beautiful Soup to parse the HTML and extract the following data:

- Product name
- Selling price
- Ratings
- Product category

We used PIL library to use and manipulate the backgroud image used in the GUI

Implementation

Data cleaning

Once we had extracted the data, we performed some cleaning and formatting to make it more usable. This included converting the price and discount strings to float values, and removing any extraneous characters or spaces.

Data storage

We stored the scraped data in a Pandas DataFrame, which allowed us to easily manipulate and analyze the data. We also used the numpy library to perform some statistical analysis on the data, such as calculating the average price and discount across all products.

Implementation

User interface

To display the scraped data to the user, we used the Tkinter library to create a simple graphical user interface (GUI). We utilized the PandasTable library to display the scraped data in a table format within the GUI. The user can interact with the table by sorting and filtering the data, and can also export the data to a CSV file.

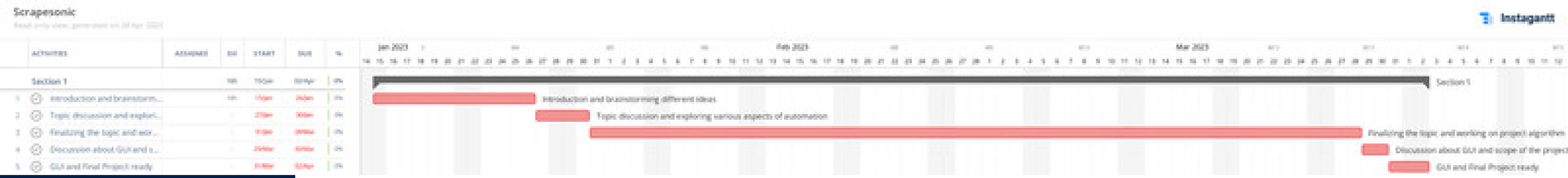
Implementation

Testing and debugging

We conducted several tests to ensure that the web scraping code was working correctly and handling errors appropriately. We also used the logging library to log any errors or warnings that occurred during the scraping process. We also introduced try/except blocks to handle any errors or exceptions that arise during the scraping process, and implemented a retry mechanism in case of network errors.

Overall, our web scraping project successfully scraped product information from Flipkart and presented it to the user in a user-friendly format. We were able to retrieve and analyze a large amount of data, and the GUI allowed users to easily interact with the data and export it for further analysis.

Gantt Chart


Scrapesonic
Instagantt
ACTIVITIES
ASSIGNED
DAYS
START
DUE
%
Jan 2023
Feb 2023
Mar 2023
Section 1
Introduction and brainstorming different ideas
Topic discussion and exploring various aspects of automation
Finalizing the topic and working on project algorithm
Discussion about GUI and scope of the project
GUI and Final Project ready

Conclusion

In conclusion, we successfully scraped product information from Flipkart using the Beautiful Soup library and other Python libraries. We extracted product name, selling price, ratings, and product category for each product, and implemented a user-friendly GUI for displaying the data.

Our project showed that Flipkart offers a wide variety of products across different categories, with varying prices and ratings. However, it is important to perform web scraping responsibly and within legal boundaries.

In future work, we could expand our web scraping efforts to include other websites and sources of data, or explore more advanced techniques for analyzing and visualizing the data.

References

- S.Dhoke,A.Sakhare,S.Sharma.”EFFICIENT SCRAPING OF DATA FROM WEBSITES USING SELENIUM”jetirhttps://www.jetir.org/papers/JETIRFM06063.pdf (accessed Feb. 2, 2023).

- C.lotfi,S.Srinivasan,M.Srinivasan,I.Latrous.”Web Scraping Techniques andApplications: A Literature Review”publications. https://www.publications.scrs.in/uploads/final_menuscript/863dc5628ae9215e611c22943d061742.pdf (accessed Feb. 2, 2023).

References

- S.Sirisuriya,.”A Comparative Study on Web Scraping”kdu.ac.in http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y (accessed Feb. 2, 2023).

- M.Shreesha,S.Shrikara,R.Manjesh.”A Novel Approach for News extraction usingWeb Scrapping”https://www.ijert.org/research/a-novel-approach-for-news-extraction-uspdf (accessed Feb. 2, 2023).