National College of Ireland

Configuration Manual

MSc Research Project
Artificial Intelligence

Praneeth Bandi
Student ID: 22183922

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Praneeth Bandi
Student ID:	22183922
Programme:	MSc in Artificial Intelligence **Year:** 2023
Module:	MSc Research Method
Lecturer:	Rejwanul Haque
Submission Due Date:	05/01/2023
Project Title:	Advanced Google Scholar Scraper: A Content-Based Filtering Approach for Literature Recommendation Using BERT Embeddings
Word Count:	**1438** **Page Count: 10**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Praneeth Bandi

Date: 05/01/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual: Advanced Google Scholar Scraper: A Content-Based Filtering Approach for Literature Recommendation Using BERT Embeddings

Praneeth Bandi
Student ID: 22183922

1 Introduction

Welcome to the user guide for Advanced Google Scholar Scraper. This guide is intended to help users install and get comfortable with our advanced tool for effective literature search. A complete literature recommendation system The Advanced Google Scholar Scraper combines web scraping, natural language processing (NLP), and an easy-to-use graphical interface into a single streamlined package. However, with the growing amount of scholarly content scattered about the Internet, our scraper equips researchers and students everywhere to quickly cut through Google Scholar's gargantuan heap.

This manual will guide you through installation, system requirements and configuration settings. Setting the program up couldn't be easier. How to wield the power of BERT embeddings and cosine-based similarity metrics, so you can receive highly accurate and relevant literature recommendations. Tkinter-based GUI: explore its user friendly intuitive interface.

You can use this application to study Natural Language Processing, Machine Learning, BERT Models and more. Its cross-discipline robustness means it is equally at home in all such fields. Jump into the guide to make full use of Advanced Google Scholar Scraper, and keep an eye open for further developments as we continue to add functionality.

2 System Requirements

Before setting up the Advanced Google Scholar Scraper, ensure that your system meets the following requirements for optimal performance:

Operating System	Windows 10 or Later
Programming Language	Python
Integrated Development Environment (IDE)	PyCharm
Memory (RAM)	4GB+
Graphical Processing Unit (GPU)	If Available
Web Application	Chrome
Processor	Minimum I3

Operating System:

- Windows 10 or later.

Programming Language:

- Python: The scraper is developed in Python, and the system should have Python installed. Download the latest version from python.org.

Integrated Development Environment (IDE):

- PyCharm: We recommend using PyCharm as the preferred IDE for running and managing the scraper. Download the latest version from JetBrains.

Memory (RAM):

- 4GB or higher: Ensure your system has a minimum of 4GB RAM for smooth execution. Higher RAM capacity will enhance performance, especially when dealing with large datasets.

Graphics Processing Unit (GPU):

- If available: While not mandatory, having a GPU can potentially accelerate certain processes, providing a more efficient experience.

Web Application:

- Google: The scraper interacts with Google Scholar, so a stable internet connection and access to Google services are essential.

Processor:

- Intel Core i3 or equivalent: The processor should be at least an Intel Core i3 or equivalent for effective execution of the scraping tasks.

Ensure that your system components align with these specifications to fully leverage the capabilities of the Advanced Google Scholar Scraper.

3 Installation Instructions

Follow these steps to set up the Advanced Google Scholar Scraper on your system:

1. Install Python:
   Ensure Python is installed on your system. Download the latest version from python.org .

2. Install PyCharm:
   Download and install PyCharm, the recommended Integrated Development Environment (IDE), from JetBrains.

3. Install Required Libraries:
   Open a terminal or command prompt and run the following command to install the necessary Python libraries:

pip install pandas requests torch beautifulsoup4 selenium scikit-learn transformers

4. Download ChromeDriver:
   Download the ChromeDriver executable suitable for your Chrome browser version from ChromeDriver Downloads.

5. Extract and Add ChromeDriver to PATH:
   Extract the downloaded ChromeDriver executable and add its location to your system's PATH environment variable.

6. Open PyCharm:
   Open PyCharm and create a new Python project.

7. Copy and Run the Code:
   Copy the provided Python code into a Python file within your PyCharm project. Run the script to ensure it executes without errors.

8. Install Missing Libraries:
   If any library is missing, use the following command in PyCharm's terminal to install it:

   pip install library_name

9. Configure WebDriver:
   Ensure the webdriver.Chrome() call is compatible with your Chrome browser version.

10. Run the Scraper:
    Run the script, and the Tkinter GUI for the Advanced Google Scholar Scraper will appear.

11. Enter Query and Settings:
    Input the desired article title or keyword, specify the number of pages to scrape, and choose an output folder. Optionally, browse for an output folder.

12. Click "Extract Data":
    Click the "Extract Data" button to initiate the scraping process. Monitor the progress through the displayed progress bar.

13. Completion Message:
    Once the extraction is complete, a status message will indicate success. The extracted data will be saved in a CSV file.

14. Close the Application:
    Close the application when finished.

Now you have successfully set up and used the Advanced Google Scholar Scraper on your system. Adjust the code as needed for your specific requirements.

Note: Ensure you have a stable internet connection and access to Google services for the scraper to function properly.

4 Installation Demo

1. PyCharm Installation

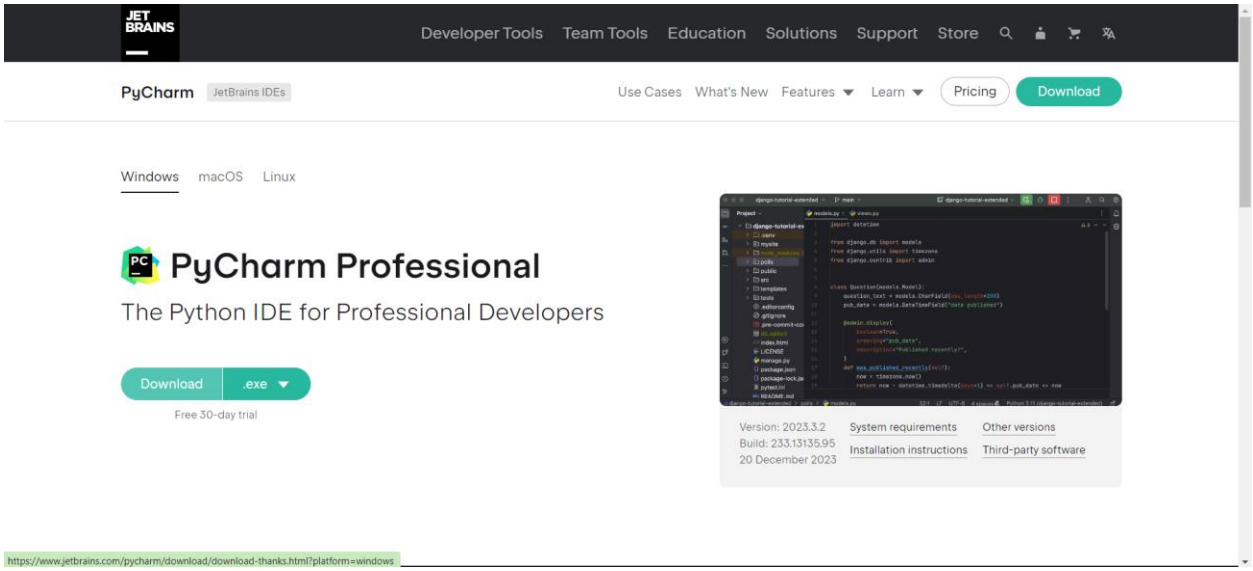

Figure 1

After going to the Website Click on the download button as in Figure 1

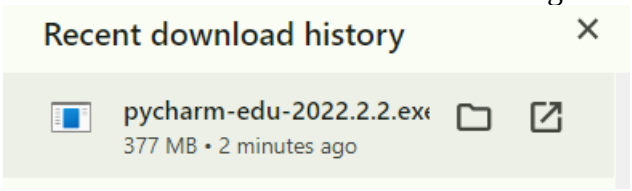

Figure 2

After Clicking, Download will be started as in Figure 2

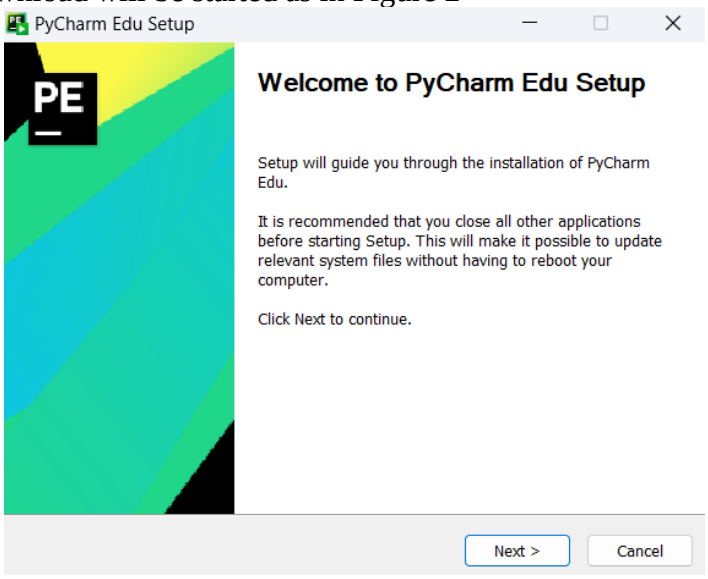

Figure 3

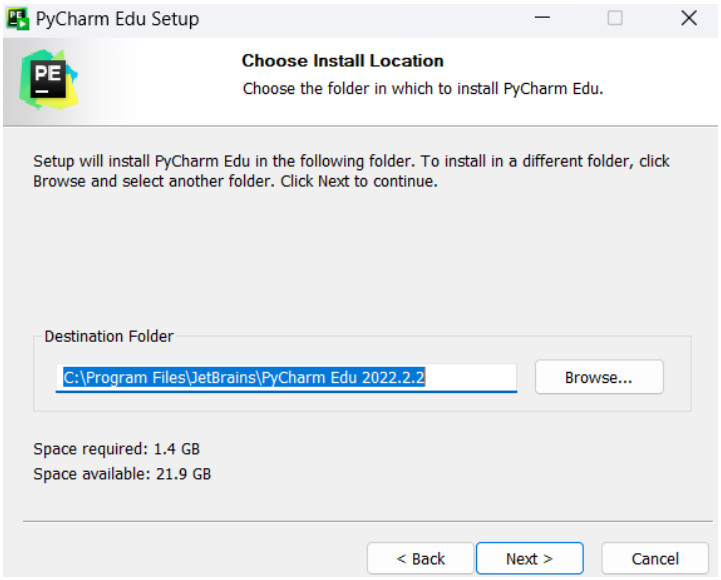

Figure 4

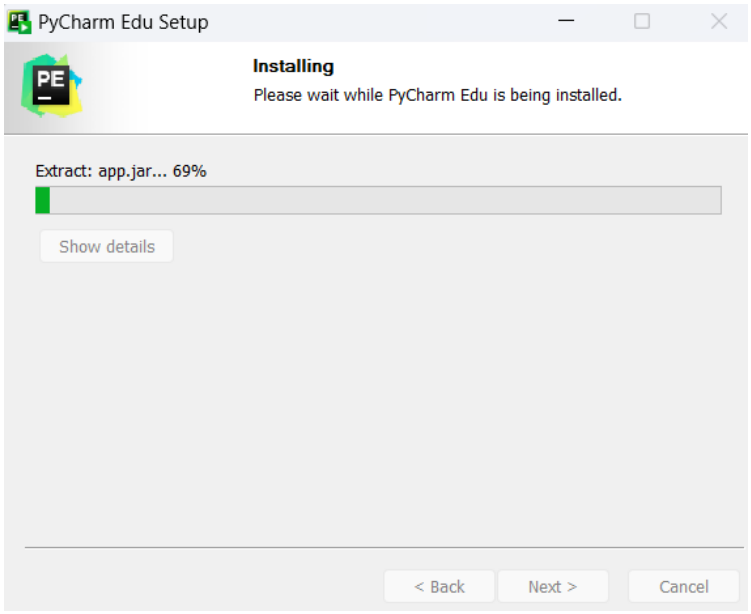

Figure 5

Click on the .exe file and complete the Installation as in Figure 3, 4 and 5.After the Setup is completed, open the folder of the Code and import the Scholar Scraping Code.

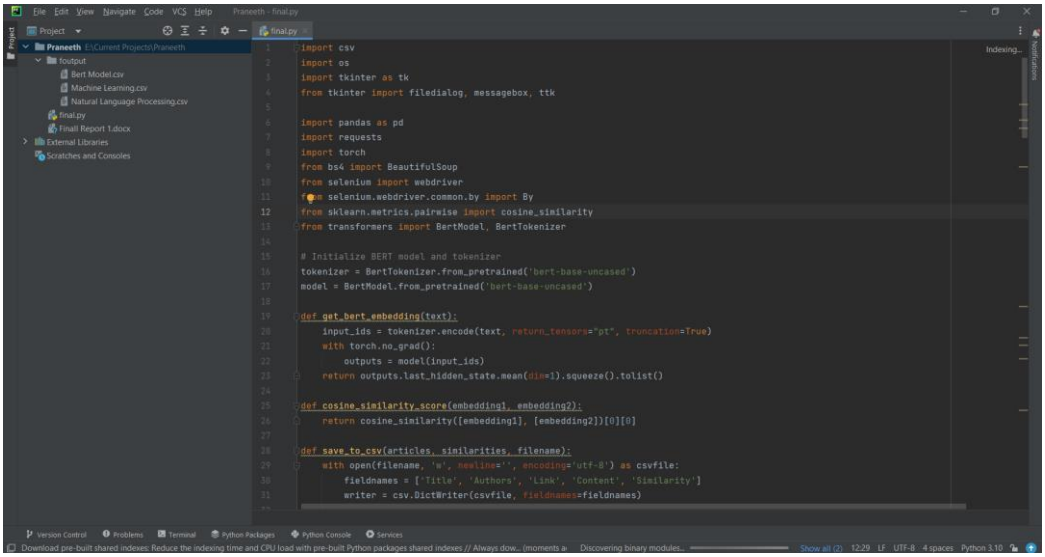

Figure 6

After Importing the Window will look like as in Figure 6.

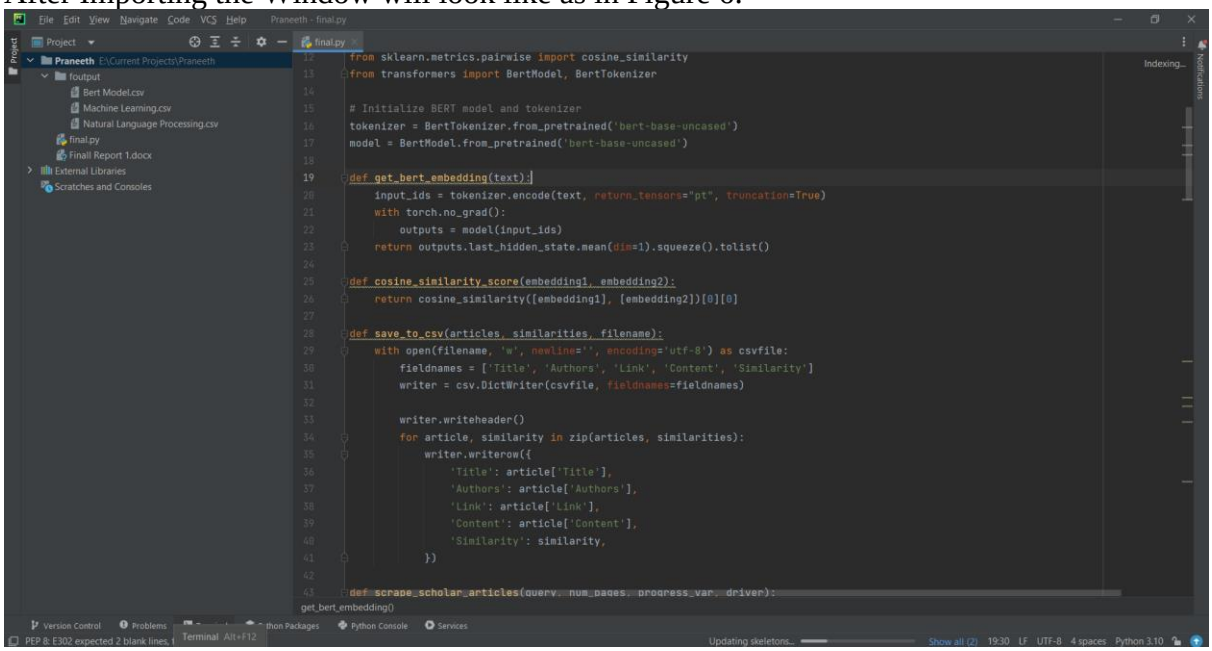

Figure 7

Now open the Terminal, Either by Alt+F12 or Clicking on the terminal in bottom left as in Figure 7.

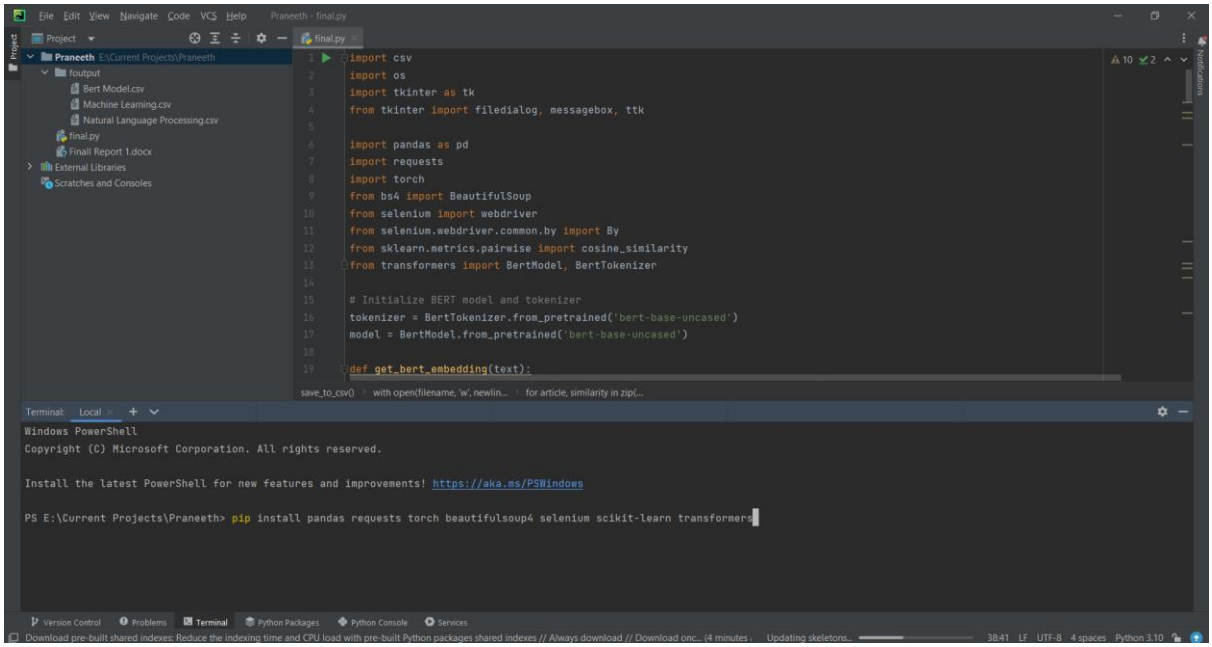

Figure 8

After opening the Terminal, Copy and Paste the pip installation command as in 3.3 Install Required Libraries or Just by simply copy, pasting below command as in Figure 8
pip install pandas requests torch beautifulsoup4 selenium scikit-learn transformers
And hit enter.

Now All the Installation work is completed.

5 How to Run

The Advanced Google Scholar Scraper gives users the option of fine-tuning many different configuration items to suit their own taste. Follow these guidelines to customize the configuration settings:
To Run:

- Open Terminal

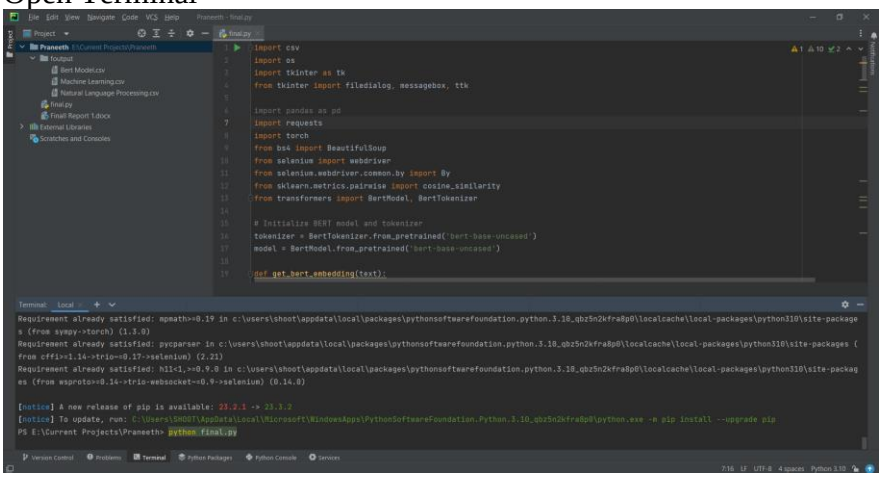

Figure 9

- Type python filename.py such as python final.py and hit enter as in Figure 9.

Query Configuration:

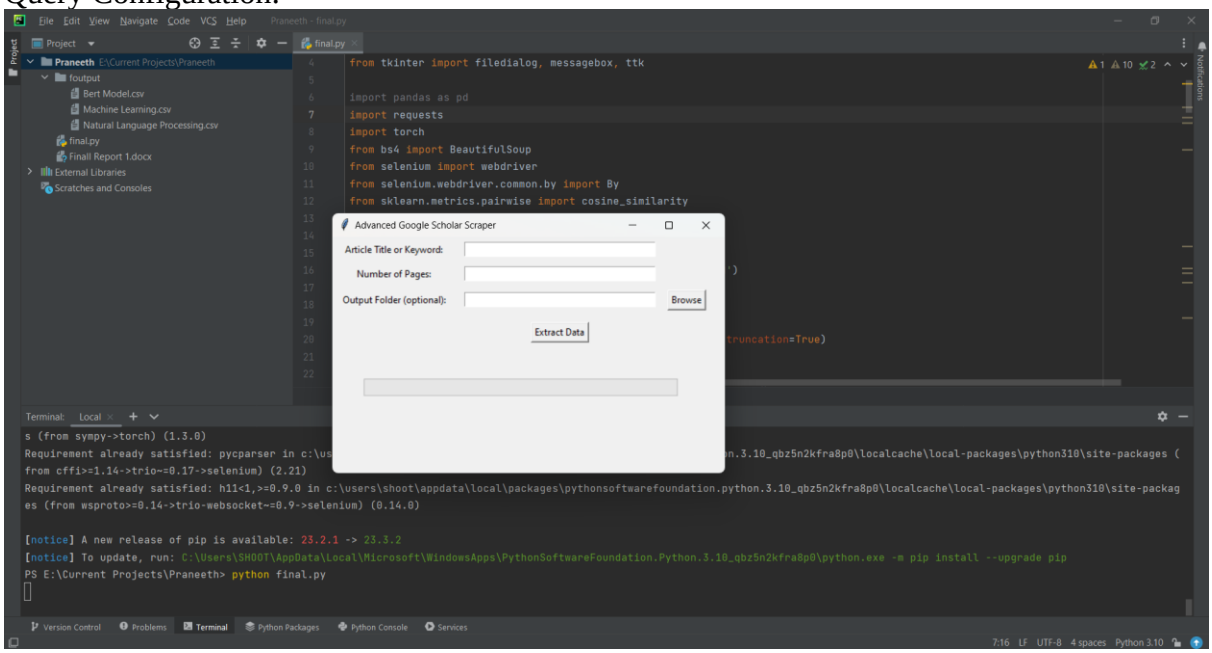

Figure 10

- Locate the entry box labeled Article Title or Keyword in the GUI as in Figure 10.
- Type in your preferred article title or key word to formulate the search term.
- Modify the query to suit your own needs.

Number of Pages to Scrape:

- In the GUI, find the "Number of Pages" entry box.
- Select the number of pages to scrape from Google Scholar.
- Tune the number to the extent of literature search you require.

Output Folder Specification:

- In the GUI, find the entry box for "Output Folder (optional)."
- You can optionally specify the folder in which the extracted data will be stored.
- Click the "Browse" button to browse and choose an output folder.
- If there is no output folder specified, then the script will save the data in current working directory.

Through these simple steps, users can successfully customize the Advanced Google Scholar Scraper to meet their research needs. An easy-to-use GUI and clearly written code that is well documented make customization and adjustment simple, even for technically illiterate users.

References

Python Software Foundation. (n.d.). Python.org. Retrieved from https://www.python.org/

JetBrains. (n.d.). PyCharm: Download. Retrieved from https://www.jetbrains.com/pycharm/download/?section=windows

Chromium Projects. (n.d.). ChromeDriver - WebDriver for Chrome. Retrieved from https://chromedriver.chromium.org/downloads