A Command Line Tool for Tracking Error Details of Program using Web Scrapper

A Poongodai, R Suhasini

Abstract Web scraping, a software technique is used to extract information from any website. The proposed system is a command line tool that uses web scrapping to extract the relevant data from stack overflow website using the library "beautiful soup". Programmers find difficulty to browse the information about the error encountered during the execution of their program. They take more time to find the appropriate information about the error. The aim of the proposed system is to help them for tracking error details thereby reducing/nullifying their browsing time for retrieving error details. This tool is implemented for five different programming languages and it works for almost all the operating system except windows 10.

I. INTRODUCTION

Most of the times though data exist, it may not be easily accessible. Many users spent a large amount of time in extracting for the effective information from the web source. Also beginner programmers may not have any idea where to browse for the details to rectify the error that occurs during program execution. There is no doubt that the internet provides so much incredible information today that we could dig out how valuable it could be.

There are varieties of software applications available on web data extraction systems for extracting the information from websites. [1] These applications are categorized into two groups: Applications at the Enterprise level and at the Social Web level. [2] At the Enterprise level, they perform data analysis, used for the business process. At the Social Web level, it gathers information from Online Social Network, Web 2.0 and Social Media to analyze human behavior at a very large scale. [4][7] Web crawling and web scraping are extremely popular methods of information retrieval. [3] Web scraper target for the specified data from the webpage or site whereas Web crawler identifies the new web page to scrap for building a large collection of web data. [3]

Web scraping is also known as Screen Scraping or Web Harvesting. It is a technique used to extract information from web source. The extracted data is saved in the personal computer or to a database in the format of spreadsheet. [5][6] The information from the web sources can only be viewed through web servers. If one wants to copy and paste the information it can only be done by manually. Manual copy of the large amount of data is a difficult task and it takes a long time (from hours to days). But web scrapping does it automatically. The web scrapping software can copy the information within a fraction of time. [6] Scrapy is used for scrapping the web data at larger scale and Beautiful soup is simple and easy to use for small scale web scrapping.

The proposed system is a command line tool for the execution of the programming language such as Java, JS, and Python etc. This tool uses web scrapping technique to retrieve information about the programming errors from the website “Stack overflow”. This proposed web scrapping tool is built using the python library Beautiful soup.

II. RELATED STUDY

2.1 Web Scraping

Web scraping is the process of scrapping the entire or required data from the website instead of copying each and every time. [8] As an example, it enables the user to search an article written by specified author published in a magazine website in any given year. [9] This software is built using python or any other programming language and the software is easy to use for nonprogrammers. Some of the commercial scraping software’s are Helium Scraper, outwit Hub, FMiner, Mozenda, RapidMiner, Beautiful Soup etc. [12] Scrapy is used for scrapping the web data at larger scale [13] and Beautiful soup is simple and easy to use for small scale web scrapping. [15][16]

Web scrapping downloads the text from one or more webpage within a single website. It recognizes various types of content in the website, acquire and store only user specified content. The web scraping technique extracts structured or unstructured data from the websites available in different format like JSON, CSV, EXCEL, HTML and XML. [10]

The major advantage of web scrapping is that it can penetrate the information too more extant than a traditional search engine. For instance, if we search for “cheapest five star Hotels accommodation in Chennai”, then Google results with uncontrollable advertisement and popular hotel search sites. Google only knows what these websites say on their content pages, but with seasoned web scrapers, users are able to know the cheapest five star hotels in Chennai in terms of cost for accommodation across a variety of websites and the best time to book the hotels. [11]

Web scrappers are very well capable of collection, processing large amount of required information from websites. This task is accomplished through an automated program written by user. This automated program tells the scrapper on which page to start, the type of text to look for, what to do with the found text, where to save the text, where

Revised Version Manuscript Received on 10 September, 2019.

A. Poongodai, CMR College of Engineering & Technology, Hyderabad, Telangana, India.
R. Suhasini, CMR College of Engineering & Technology, Hyderabad, Telangana, India.

to navigate next and when to stop. This automated program queries a web server, requests data in the form of the HTML and then parses that data to extract needed information. [14]

2.2 Beautiful soup

Beautiful soup is a Python library, designed to create a parse tree from parsed HTML and XML documents. Beautiful Soup enables the user to extract specific content from a webpage, remove the HTML markup, and save the information.

Beautiful soup is a tool for web scraping, works with the favorite parser to provide idiomatic ways of navigating, searching and modifying the parse tree. [15] It commonly saves programmers hours or days of work. It works with various python parsers such as lxml’s HTML parser, lxml’s XML parser, python’s html.parser and html5lib. [17]

Using Beautiful Soup, data can be easily accessed from the parsed HTML and prettify() is used to beautify the parsed data. Scrapy can be used to scrap the data for extracting the data from html/xml but Beautiful Soup is more preferred than the Scrapy.

2.3 Stack Overflow

Stack Overflow is a privately held largest online website. It is the flagship site of the Stack Exchange Network for professional and programmers. They share the knowledge, learn and build their careers. The site consists of questions and answers on varieties of themes in programming. The similar sites like Stack overflow are Question2answer, Quora and Scoold. More than 50 million professional and aspiring programmers visit Stack Overflow each month to help solve coding problems.

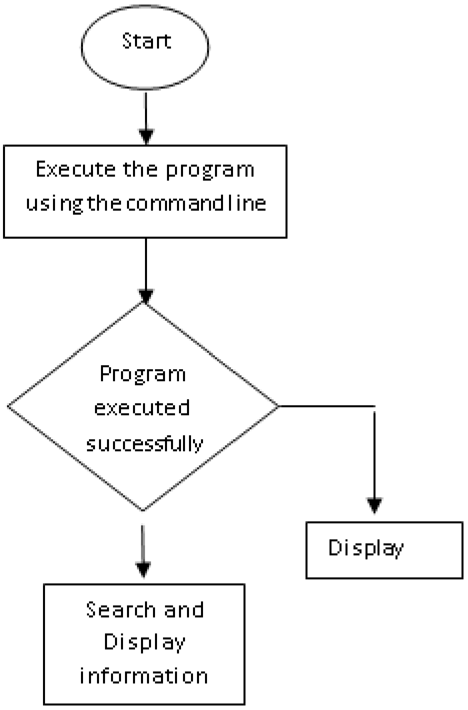
The Stack Overflow website serves as a platform for users to ask and answer questions, and, through membership and active participation, to vote and edit questions and answers. It only accepts questions about programming that are tightly focused on a specific problem. Questions of a broader nature–or those inviting answers that are inherently a matter of opinion– are usually rejected by the site's users, and marked as closed.

III. PROPOSED SYSTEM

The proposed system is a command line tool that instantly fetches relevant data from stack overflow website whenever error encounters while executing a program in command line. The proposed tool is developed using web scrapping technique. This helps a lot in reducing the user’s browsing time regarding the error occurred during program execution. This tool supports nearly 5 programming languages: Python, JAVA, GO, Ruby and JS.

Figure 1 shows the work flow of the proposed command line tool. The following are the steps involved in the work flow process

1. Execute the program using the command line tool
2. If no error, Display the output.
3. If error occurs, fetch the details of error using web scrapper and display it


Start
Execute the program using the command line
Program executed successfully
Display
Search and Display information

In the proposed tool, fetching the error information is carried out using web scrapping technique through Beautiful soup and python.

The following are the steps involved in the process.

a. Installing required packages

The required packages to implement this tool are packages "requests" and "beautifulsoup4". They are installed using pip install command.

```
Pip install requests
pip install bs4
```

b. Collection of the web pages

The web pages are collected using request.get() method. The url of the website to be scrapped is passed to that method.

```
import requests
url=https://stackoverflow.com
page = requests.get(url)
```

To work with web data, the text-based content of web files should be accessed. It can be accessed using page.text or page.content

```
page.text
```

Once the above code is executed, the output is the html file, an unreadable format (not be able to read by human eye).

c. Run the web page document through Beautiful Soup

The document (unreadable format) is passed to the “beautifulsoup” constructor and it is parsed by using python's built-in parser "html.parser". It returns BeautifulSoup object, which represents the document as a nested data structure (parse tree). The content of Beautiful soup object is transformed into formatted unicode string using function prettify().

```
from bs4 import BeautifulSoup
soup=BeautifulSoup(page.text, 'html.parser')
print(soup.prettify())
```

d. Finding Tags by Class and ID

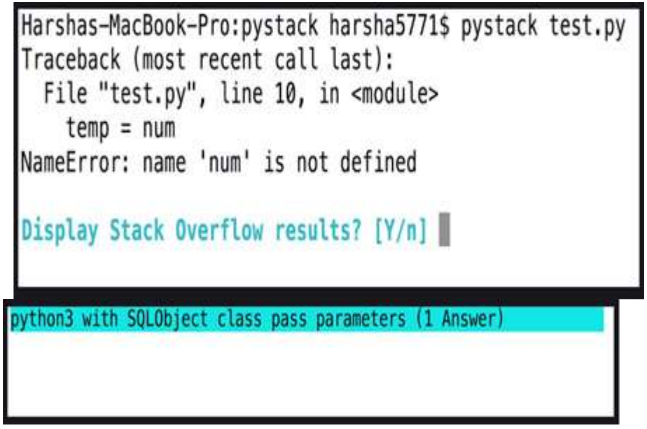
From the tree structure (beautiful soup object), search for the corresponding error that took place in the command line while executing the program and displaying the data that is present in the paragraph tag and also header tags.

```
soup.find_all( 'p' , class= 'App' )
```

The above code retrieves all the data from the p tag. Similar search operation is performed and all the data collected is question and answers. The question data is collected the from the header tag and the answer data from the paragraph tag after searching of corresponding keyword in the stackoverflow api.

IV. RESULT

The command line tool is used to execute the programs of five programming language. Figure 2 illustrate the execution of a python program. The tool executes the program, asks for the user suggestion and then it display the error details which is shown in Figure 2.



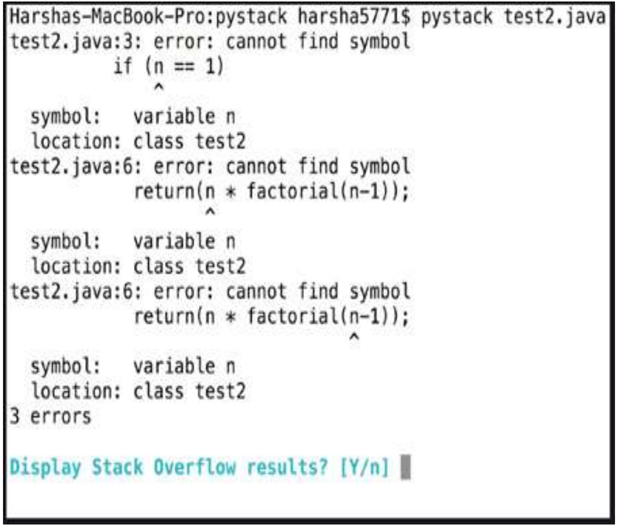
```
Harshas-MacBook-Pro:pystack harsha5771$ pystack test.py
Traceback (most recent call last):
  File "test.py", line 10, in <module>
    temp = num
NameError: name 'num' is not defined

Display Stack Overflow results? [Y/n]
```

```
python3 with SQLObject class pass parameters (1 Answer)
```

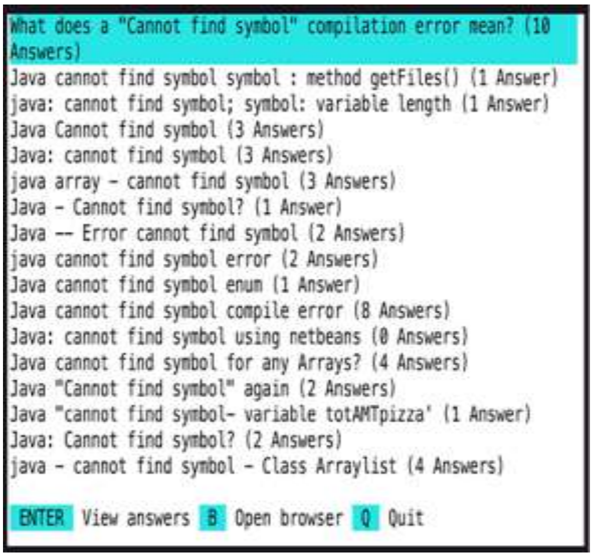
Figure 2: Execution of Python Program

As another example, Figure 3 shows the execution of Java program by the newly developed command line tool. On execution, errors are displayed and on clicking 'yes', the information about the errors from stack overflow site is displayed. In a similar way the tool works for JS, Ruby and Go (GoLang developed by Google).



```
Harshas-MacBook-Pro:pystack harsha5771$ pystack test2.java
test2.java:3: error: cannot find symbol
          if (n == 1)
              ^
  symbol:   variable n
  location: class test2
test2.java:6: error: cannot find symbol
          return(n * factorial(n-1));
                 ^
  symbol:   variable n
  location: class test2
test2.java:6: error: cannot find symbol
          return(n * factorial(n-1));
                                   ^
  symbol:   variable n
  location: class test2
3 errors

Display Stack Overflow results? [Y/n]
```



```
What does a "Cannot find symbol" compilation error mean? (10
Answers)
Java cannot find symbol symbol : method getFiles() (1 Answer)
java: cannot find symbol; symbol: variable length (1 Answer)
Java Cannot find symbol (3 Answers)
Java: cannot find symbol (3 Answers)
java array - cannot find symbol (3 Answers)
Java - Cannot find symbol? (1 Answer)
Java -- Error cannot find symbol (2 Answers)
java cannot find symbol error (2 Answers)
Java cannot find symbol enum (1 Answer)
Java cannot find symbol compile error (8 Answers)
Java: cannot find symbol using netbeans (0 Answers)
Java cannot find symbol for any Arrays? (4 Answers)
Java "Cannot find symbol" again (2 Answers)
Java "cannot find symbol- variable totAMTpizza' (1 Answer)
Java: Cannot find symbol? (2 Answers)
java - cannot find symbol - Class Arraylist (4 Answers)

ENTER View answers  B Open browser  Q Quit
```

Figure 3: Execution of Java Program

IV. CONCLUSION

Beautiful soup play a major role in developing the tool and the user interface is done using "urwid". This tool works with almost all the operating systems except for windows 10. It works only for few programming languages and hence additional work has to be done in future to include other languages also. Also, as this tool extract the data only from Stack Overflow website it can be further enhanced to extract from the other similar website such as Question2answer, Quora and Scoold.

REFERENCES

1. H. D. Pham, M. Drieberg and C. C. Nguyen, "Development of vehicle tracking system using GPS and GSM modem," in IEEE Conference on Open Systems (ICOS), Kuching , 2013.
2. Mashood Mukhtar, "GPS based Advanced Vehicle Tracking and Vehicle Control System", I.J. Intelligent Systems and Applications, 2015, 03, 1-12
3. Albert Alexe, R. Ezhilarasie, "Cloud Computing Based Vehicle Tracking Information Systems", ISSN: 2229 - 4333 (Print) | ISSN: 0976 - 8491 (Online) IJCST Vol. 2, Issue 1, March 2011
4. Ambade Shruti Dinkar and S.A Shaikh, Design and Implementation Of Vehicle Tracking System Using GPS, Journal of Information Engineering and Applications, ISSN 2224-5758, Vol 1,No.3, 2011.
5. M. Ahmad Fuad and M. Drieberg, "Remote vehicle tracking system using GSM Modem and Google map," in IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (CSUDET), Selangor , 2013.
6. M. Parvez, K. Ahmed, Q. Mahfuz and M. Rahman, "A theoretical model of GSM network based vehicle tracking system," in International Conference on Electrical and Computer Engineering (ICECE), Dhaka, 2010.
7. R.Ramani,S.Valarmathy,D. N.SuthanthiraVanitha, S.Selvaraju and M.Thiruppathi.R.Thangam,"Vehicle Tracking and Locking System Based on GSM and GPS," I.J. Intelligent Systems and Applications, vol. 09, pp. 89-93, August 2013.

8. P. P. Wankhade and P. S. Dahad, "Real Time Vehicle Locking and Tracking System using GSM and GPS Technology-An Anti-theft System," International Journal of Technology And Engineering System, vol. 2,no. 3, 2011.
9. P. Verma and J. Bhatia, "Design and Development of GPSGSM based Tracking System with Googlemap based Monitoring," International Journal of Computer Science, Engineering and Applications (IJCSEA), vol. 3, no. 2,June 2013.
10. N Mangla, K Sushma, L Kumble," IPB-Implementation of Parallel Mining for Big Data", Indian Journal of Science and Technology, 2016
11. T. Le-Tien and V. Phung-The, "Routing and Tracking System for Mobile Vehicles in Large Area," Fifth IEEE International Symposium on Electronic Design, Test and Application, pp. 297-300, January 2010.
12. P. Fleischer, A. Nelson, R. Sowah and A. Bremang, "Design and development of GPS/GSM based vehicle tracking and alert system for commercial inter-city buses," IEEE 4th International Conference on Adaptive Science & Technology (ICAST), October 2012.