Research on Realization of Python Professional English Translator

Research on Realization of Python Professional English Translator

Wenge Yang, Xue Zhao

Shaanxi Institute of Technology, Xi'an city, Shaanxi Province, 710300

6246253187@dlvtc.edu.cn

Abstract—The main purpose of professional English translator is to translate more ambiguity and more professional words and sentences to meet the needs of the work. This article combines web crawler based on Python language, and analyzes Python programming and online crawler implementation. This article uses this as a basis to study how to design a professional English translator, and study the structure, business workflow, and functions of the translator. The test results prove that the designed program can effectively realize the crawler function, so that it can obtain a large amount of consistent data in a very short time, and return these related data to the user through the interface connection design. This can not only meet the needs of users for quick translation and search of English words and sentences, but also meet the needs of actual work.

1. INTRODUCTION

With the rapid development of computer technology, the amount of network information is increasing rapidly, how to efficiently and quickly move the information required from the massive data is the key. Using Python-based crawler technology can automatically obtain relevant data from web pages. This article combines the characteristics of Python crawlers to crawl webpage data and analyzes the method of using Python's third-party requests library to realize the interface connection with Youdao translation website from the perspective of English translation. In addition, we can obtain relevant data from the translation website through this interface, and the data is returned to the translator to output the translation result to the user.

2. OVERVIEW

In terms of professional English translation software, common software includes Baidu Translator, Youdao Translator, and PowerWord. Take Youdao Translation as an example, although it can basically meet the needs of work and study for some highly specialized articles, it often keeps opening web pages. Its repetitive operation is not convenient for daily work, and it is also prone to errors. In daily English teaching work, teachers need to use an English translator to assist in translation of some highly professional words. For this reason, if you want to improve teaching efficiency and ensure the quality of work, it is necessary to develop a more efficient and convenient English translation aid device.

3. PYTHON RELATED TECHNOLOGIES

3.1. Python

Python is a dynamic and object-oriented computer programming language. The difference between this language and other computer languages such as C language and C++ language is that this language is

dynamic and has the characteristics of clean syntax and tidy code. At present, the language is mainly used in the fields of artificial intelligence, web development, web crawling, data analysis and operation and maintenance. The function of the language can be used to obtain a large amount of data, and the information obtained is effective, comprehensive and efficient.

3.2. Requests Library

The Requests library is a third-party library of the Python language, which has all the features of urllibs. It not only supports http connection and cookie retention, but also supports automatic determination of response content encoding, and automatic encoding of URL addresses and POST data. Currently, the main methods of the Requests library include obtaining HTML pages, obtaining HTML page header information, submitting POST requests to HTML pages, and submitting PUT requests to HTML pages. The return object obtained by calling the get method of Requests includes the content returned by the web crawler. The returned content should include the return status of the http request, the string form of the http response content, and the binary form of the http response content [1].

3.3. Beautiful Soup Library

The Beautiful Soup library is used as an HTML or XML parsing library in the Python language. Data can be obtained from web pages using the Beautiful Soup library. Moreover, it has a powerful data analysis method. People often use this library function to parse the source code of a web page when using web crawlers, and it has the characteristic of obtaining website nodes by specifying parameters.

3.4. Web Crawler Technology

Web crawler refers to a program or script that realizes automatic crawling of network information according to established regulations. It can imitate the way the browser accesses the URL address of the webpage, and the user can automatically obtain the required data without manual operation. Among computer programming languages, Python language is an object-oriented interpreted language with clean and simple syntax and tidy code. The limit for the module is determined by the position of the first character of each line, and each space is mandatory to use as the indentation of each level of statement. It can indicate entering the module by increasing the indentation, and indicating exiting the module by decreasing the indentation. The use of Python language for web crawlers has the following advantages. First, it can grab the interface of the webpage itself, and use the dynamic scripting language to make the interface grab more convenient. At the same time, most websites do not support crawling technology. The Python language can imitate the behavior writing request in use, and imitate the behavior of user login, information storage and settings. These can be implemented in Python external libraries. Second, it can perform operations such as text information processing after crawling the web page. Using external libraries in Python can not only provide practical document resolution functions, but also quickly use shorter codes to complete a large amount of document information processing.

4. IMPLEMENTATION OF ONLINE TRANSLATION CRAWLER BASED ON PYTHON

Youdao Translation website uses http protocol to realize the communication between client browser and translation website, and uses Python language to write crawler programs. Youdao translation website can disguise the program as a browser to realize the communication with the online translation website and realize the online translation crawler. We can use the third-party requests library in Python to crawl website data, that is, use the requests library to initiate a post request to the Youdao translation website, and process the translation results feedback from the website to implement a Python-based online translation crawler.

4..1. The Principle of Web Crawler

If you want to get the detailed data of a target, you can use the tag selector and the attribute selector. According to this feature, we can design a web page structure that allows users to find it by themselves, and provide tag selectors and attribute selectors. In this way, the URL of the target webpage can be

specified and the relevant data can be obtained using web crawler technology. Web online crawlers are divided into web applications and crawler programs. Among them, the former mainly acts as an intermediary between the crawler program and the user, sending the user's request to call the crawler to the program, and returning the crawler program to process future data. The crawler uses Superagent to request the initial page URL to get the entire web page. It can use Cheerio to combine the tag selector and attribute selector in the user configuration to analyze the page and obtain the target data. In order to meet the user's demand for invoking crawler data, the web online crawler generates a data interface as a URL for returning crawler data. The web application is connected to the database, and the database is used to save the user crawler configuration. The system can write the crawler configuration into the database when the data interface is formed. Afterwards, the system can obtain the relevant required configuration from the database when sending the data interface, and set it in the crawler for data crawling. In order to improve the efficiency of crawling, the results of crawling by the crawler are directly saved in the database and the timing function is set. Not only can the data obtained by the crawler program be called regularly, but also the database can be updated. In this way, the data can be obtained directly from the database when the user requests the data interface [2].

4.2. Crawler Crawling Strategy

In the crawler system, the URL queue that it needs to crawl is an important part, and the order of the URLs in the queue is very important. There are four crawling strategies. The first is the depth-first traversal strategy. That is, the web crawler starts from the start page, processes each link one by one, and then processes the links behind another start page after processing one line. The second, breadth first strategy. That is to search according to the tree level, first of all, it must complete a level search. If there is no match, go to the next level. This strategy is a blind and extensive search with lower efficiency, but it can cover more web pages. The third, part of PageRank strategy. That is, the downloaded web pages and the URLs of the URL queue to be crawled form a collection on the web page, and the PageRank value of each page is calculated. In this way, all URLs in the queue to be crawled can be sorted according to the level value of the web page. The fourth, the online page importance calculation crawling strategy. This method can crawl according to the priority priority in the queue without the iterative calculation process, and the calculation speed is faster.

4.3. Web Online Crawler Workflow

When using the web crawler to design a professional English translator for Python, the workflow of the crawler part is as follows. Firstly, the user initiates a URL request, the web responds to the web online crawler configuration panel, the user submits the crawler configuration, and the web submits this part of the configuration requirements to the crawler program. Secondly, the crawler program completes the data request and processing analysis according to these configurations, and transmits the processed results back to the web, and the web returns the results to the user. If the data meets the user's search needs, the user initiates a data interface generation request through the Python professional translator. The web application determines whether the user is logged in, and saves the crawler configuration and data processing results submitted by the user in the database while logged in. The system can obtain the crawler configuration from the database according to the program timing function, obtain the crawler program regularly, and update the data in the database. The user then calls the data interface and takes out the data in the database to complete the web response.

4.4. Web Online Crawler Implementation

First of all, the back end of the web crawler uses the node.js platform, the front end uses the Vue framework for page construction, and the front end completes the construction of the web development framework. Secondly, the system uses the superagent request library as the client request proxy module, and cheerio grabs the page module to improve the server's rapid response capability, flexibility and ease of implementation. Besides, the system will use the Async process control toolkit to provide asynchronous functions, and complete the node.js design based on the Python language. Meanwhile,

the system will use MongoDB database based on distributed files. Its requirements for document structure are relatively loose, and this storage method improves the convenience for users to obtain data.

Professional English translators based on Python language will use online crawler programs to obtain relevant data. The writing of the web front-end page is very important. The function of this page is to fill in and submit the crawler configuration. Firstly, the basic configuration includes the initial page URL, each level of tag selector from the initial page to the intermediate page, and the target data tag selector and attribute selector of the target page. Secondly, writing a crawler program will request the initial page URL based on the superagent request library. Then, it will use Cheerio analysis to obtain the next-level page URL based on the tag selector provided by the user. This will push all the obtained URLs into the queue until the tag selector queue, and all tags will be analyzed. Thirdly, the system will store the data obtained by the crawler into the database and send it back to the user by the web application. In order to enhance the efficiency of crawling data, the system will use the async asynchronous request library to request concurrent requests. In order to avoid being found by the target server to crawl data, the system will use http proxy pool proxy to request access. Finally, the system should look for the user configuration in the database when calling the data interface before calling it, and then return the result. The system can store the data which is obtained by the crawler in the database and set the timing function to periodically call the crawler program and update the database. When requesting a data interface, the system can directly obtain relevant data from the database and return the result to the user.

5. PYTHON PROFESSIONAL ENGLISH TRANSLATOR

5.1. The Structure of Python Professional English Translator

The role of a professional English translator is to translate some highly professional words and sentences. In the meantime, it should also have the advantages of simple operation, no blocking, and automatic translation. The professional English translator is divided into three parts: input, completed translation and output. The input and result output are the parts displayed on the translator interface. The translation software must be easy for users to use, and the operation interface should be as simple as possible. The operation of translation requires the translator to have extremely strong and extremely flexible translation capabilities. In this way, it can not only perceive the input words and sentences sharply, but also retrieve relevant data from translation software such as Youdao Translation. Then, the system will complete the translation conversion and realize the result output.

5.2. Python Professional English Translator Business Process

Before developing the translator, we must analyze the business process of the program. The staff can give feedback based on daily teaching experience and students' learning. This requires professional English translators to have interface operations and functions for inputting translation content and outputting translation results. It has the function of automatically switching translation between Chinese and English. Researchers combine the characteristics of translation requirements and functional structure of the translator to design the system business translation process as shown in Figure 1.

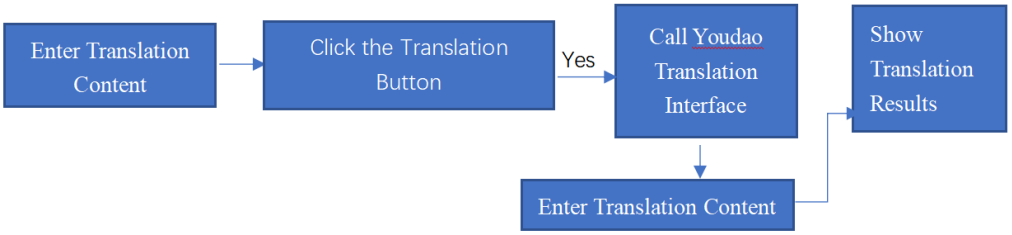

Figure 1. Translator Process

5.3. Python Professional English Translator Function Realization

The software uses the Requests library and Tkinter window design module, and calls the Youdao translation interface to realize the translation function. Requests need to be sent to third parties when developing in the Requests library in the background. The Requests library in Python meets this requirement. The system uses the module data to complete function installation in a third-party library [3].

5.3.1. Requests Library

When developing in the background, the system uses the Requests library in Python to well meet the requirements of sending http request scenarios to third parties. The Requests module data will be installed using pip in a third-party library.

5.3.2. Tkinter Module

This module is the standard library of Python. Python uses Tkinter to quickly create applications. The Tkinter module is embedded in the Python installation package. Therefore, after installing Python, you can start the Tkinter library, and use the Tkinter module for IDLE writing and simple graphical interface processing.

5.3.3. Youdao Translation Interface

Take Youdao translation interface as an example. This software interface provides developers with an open excuse. The website can establish various functions and applications through the connection with Youdao translation interface to provide users with accurate, fast and convenient translation services. The open interface does not have high requirements for language applications, which helps reduce the difficulty of language understanding, and is especially suitable for beginners.

5.3.4. Programming

First of all, the Requests library is a module for http requests. Importing the Requests library can facilitate web crawling in the program. All expression functions of the Python language included in the re library can be easily checked whether the string matches a certain pattern after importing the Requests library. Secondly, the system will request the URL address entered externally to understand the encoding method of the web page. Subsequently, the system will use regular expressions in its source program to perform string matching on the defined data format, and convert the matched URL into a string form and store it in the dictionary. Thirdly, the system will make a URL request for the externally input words and sentences, and use the relevant function to open the file at the specified location. As a result, we have to store a large number of words and sentences, and the specified file location should be defined as a variable mode.

When designing an application, first define the translate class, and complete the initial window name and size design, input box and text display box settings. The input box is used to input the user's translation of words and sentences, and the display box is used to output the translation result. In addition, the system must complete the reasonable design and layout of the page elements, so that the positions of the various components are placed in the appropriate positions in order to improve the efficiency and convenience of page operations. The design of the translation function is the key to the program design, and the translation function can be triggered instantly by adding steps through buttons. Finally, we should also pay attention to the functional design of clearing the output box and clearing the content of the input box, as well as the design and interface calls of the examples of the translate class.

6. REALIZATION OF TRANSLATION CRAWLER FOR PROFESSIONAL TRANSLATOR BASED ON PYTHON

6.1. Preparation for Translation Crawler

Firstly, we need to enter the website address of Youdao translated in the address bar of the browser. Afterwards, we should open the fiddler software to capture the communication packets, and enter the text to be translated in the dialog box of the website translated by Youdao. In this way, the relevant translation results can be output through translation, and the results are used for later crawler verification. Secondly, we can use fiddler software to analyze browser and Youdao translation website communication data packets to analyze the communication protocol and related parameters, and use fiddler software to obtain data packet analysis. Youdao Translation uses the post method in the tttp protocol, which submits the text to be translated to the server and analyzes the communication post request parameters of the Youdao translation website. According to the analysis results of the fiddler software, the translation results returned by Youdao Translation website use the json format, which is a simple data exchange format. It is easy to read and write and uses small memory, easy to parse and supports multiple languages [4].

6.2. Translator Encryption Verification Analysis

After using the fiddler software to analyze the protocol of Youdao's translation http website, in order to avoid the acquisition of information by web crawlers, the system adopts an encryption verification method. In this way, you can analyze the files on Youdao Translation, generate the salt timestamp, and use the message digest algorithm to generate the hash value. The translated text, the encrypted string on Youdao Translation, etc. will change. The hash value of the specific encrypted string calculated by MD5 is the value of the post request. The parameters in the timestamp parameter algorithm indicate that the timestamp of the current time is obtained. The timestamp returned by Python is a floating point number with a 10-digit integer. In order to obtain the 13-bit time value by multiplying the time stamp by 1000, we add the 13-bit time stamp value to a positive integer within 10 to obtain the integer part of the time stamp parameter salt.

6.3. Translation Crawler Implementation and Result Analysis

After analyzing the request method, parameter meaning, returned result, and encryption verification of the professional translator website, the system will construct the request header through http and construct the data parameter in the post request. At the same time, the system will send a post request to the translator website to obtain the returned json data to obtain the translation result [5].

First of all, the HTTP request header is created to disguise it as a browser, which can first translate the website's communication with Youdao. Secondly, the constructed post request data parameters should include the main parameters such as the text to be translated, salt, and sign. The specific way to construct the parameters in the post request is to first define the data parameters in the post. Among them, headers is the address header of the http construction website, and src_word is the text to be translated. Thirdly, the system will send a post request to the translation website. That is, after constructing the post request data parameter, use the post method in the Requests library in Python to send a post request to Youdao Translation. Finally, the system will obtain the json parameters from the response and obtain the translation results through related processing and print them.

We will compare and analyze the translation results through the design of Youdao Translation and the Python-based professional translator crawler program. It proves that the translation results are consistent, which shows that the Python-based professional translator crawler has been implemented.

7. CONCLUSION

To sum up, the use of Python language crawler technology for professional English translator program design not only can effectively realizes the function of web crawler, but also can obtain the translation result in a short time. Furthermore, this allows for limited optimization of program conditions and

real-time data acquisition. Meanwhile, this also provides users with more convenient and efficient translation services, saving a lot of time for data searching and repetitive operations such as translation input.

REFERENCES

[1] Han Qianjin. Design and Implementation of Web Online Crawler [J]. Software, 2018 (9): 111.
[2] Qiu Ming. Program design of image crawler based on Python [J]. Industrial Technology and Vocational Education, 2019 (3): 2-4.
[3] Bai Xueli. Analysis on the characteristics and application of Python crawler technology [J]. Shanxi Science and Technology, 2018 (2): 53-55.
[4] Guo Lirong. Web crawler program design based on Python [J]. Electronic Technology and Software Engineering, 2017 (23): 248-249.
[5] Xiong Chang. Webpage data capture and analysis based on Python crawler technology [J]. Digital Technology and Application, 2017 (9): 35-36.