

**ELEC S411F (2020/21)**

**Electronic and Computer Engineering Project**

**Final Report**

|  |  |
| --- | --- |
| **Estimation of Search Engine Index Size** | |
|  |  |
| Project number: | B10 |
| Student name | Tang Ho Man |
| Supervisor: | Bruce K. B. Tong |
| Submission Date: | 13/6/2021 |

**Declaration of Originality**

I, TANG Ho Man, declare that this report and the work reported herein was composed by and originated entirely from me. This report has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the reference section.

8/1/2021

**Abstract**

Nowadays the size of the worldwide web becomes bigger and bigger. Useful information cannot find easily if we do not have a powerful search engine. The actual indexed size in the search engine is unknown due to those search engine company providers did not release the data of it, so this project is about estimating the indexed webpages in search engines.

In this project, the estimation value is calculated by the result returned by the search engine. The estimated index size of Google and Microsoft Bing (hereinafter referred to as “Bing”) on average is around 29.7 billion and around 9.3 billion respectively. Before calculating the estimated value in different search engines, many testings were done to ensure the accuracy of the estimation value, such as checking the frequency result of the Chinese characters, writing a program for requesting the return results and index item (Title, URL, Description) in the search engine, solving the problem during requesting and writing the automation for the program to let the estimation become more exactly.

Upon the completion of the project, the size of indexed web pages in search engines can be known. The quality of search engines can measure by the collected data later for choosing a better search engine and getting better information.

**Table of Contents**

[**1.** **Introduction** 4](#_Toc74474833)

[**1.1** **Problem statement** 4](#_Toc74474834)

[**1.2** **Project Objectives** 4](#_Toc74474835)

[**1.3** **Organization of the report** 5](#_Toc74474836)

[**2.** **Literature Review** 7](#_Toc74474837)

[**2.1** **Research in the pass** 7](#_Toc74474838)

[**2.2** **Chinese character** 11](#_Toc74474839)

[**2.3** **Market share of the search engine** 14](#_Toc74474840)

[**2.4** **Other related to this research** 17](#_Toc74474841)

[**3.** **Methodology** 19](#_Toc74474842)

[**3.1** **Preparation before requesting and scrapping** 19](#_Toc74474843)

[**3.2** **Data Collection and Program** 22](#_Toc74474844)

[**3.3** **Calculate, Analysis and Observer the estimation** 25](#_Toc74474845)

[**4.** **Experiment** 27](#_Toc74474846)

[**4.1** **Preparations before requesting/scrapping data** 27](#_Toc74474847)

[**4.2** **Program for scrapping the search engines** 38](#_Toc74474848)

[**4.3** **Calculating, analyzing, and observing the estimation.** 51](#_Toc74474849)

[**5.** **Results and Discussion** 55](#_Toc74474850)

[**5.1** **Estimated indexed web pages and changes** 55](#_Toc74474851)

[**5.2** **Chinese characters** 59](#_Toc74474852)

[**6.** **Conclusion and Further Work** 60](#_Toc74474853)

[**References** 62](#_Toc74474854)

[**Appendix** 65](#_Toc74474855)

## **Introduction**

In an information explosion year, the size of the worldwide web becomes bigger and bigger which has around 41 billion websites on 13th June 2021 [1] (shown in Figure 1.1). Information cannot be found easily if we do not have a powerful search engine. There are two factors to determine the search engine is powerful or not are the response time and the size of indexed web pages. As we know that the response time of search engines nowadays is quite fast so that the size of the indexed web pages becomes more important. The estimation of indexed webpages had done in English and Spanish between 2006 and 2016 [2], which has the largest and 4th largest total number of speakers [3]. Figures 1.2 and 1.3 shows the estimated indexed web pages in Google and Bing on 13th June 2021 are 47 billion and 5.7 billion respectively.

### **Problem statement**

Although Chinese having the 2nd largest total number of speakers, no one had done the estimation of indexed webpages in search engines using Chinese. Also, those search engine companies did not release how many web pages that they had been indexed in their search engine. So, this project is using Traditional Chinese for estimating the indexed web pages in search engines.

### **Project Objectives**

The project aim is to estimate how many webpages does the search engine had been indexed by Traditional Chinese. Before calculating the estimation, the frequencies of Chinese characters are required to be known and write the program to scrape the data from the search engine with the correct character frequency. So, here are the objectives of this project:

* Determine a list and find out the frequency of Chinese characters.
* Implement a python program for recording the returned result and indexed items.
* Estimate the size of indexed web pages in search engines.
* Observe the changes of estimated size in a short period of time.

### **Organization of the report**

This report is organized as follows:

Character 2 is the literature review of the research in estimating the search engine size in the past. What methodology and how is work for estimating the size of their project.

Character 3 is the methodology that I used in this project, such as how to prove the percentage of occupation in Chinese characters is up to date, what program that I wrote that is related to this research.

Character 4 is the experiment for checking the correlation of Chinese characters and getting the estimated indexed size number of the search engines, including how to get the estimation daily, how to calculate the final estimation indexed size number of the search engine, and how to save the scrapped data.

Character 5 is the results of the search engine, such as the estimated size of the search engine and the trend of the search engine’s indexed webpages.

Chapter 6 is the conclusion of this research and the future work about this project, such as what will do after this research.

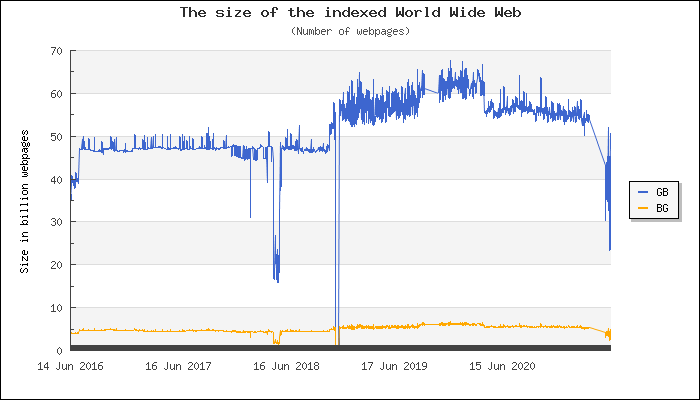
****

Figure 1.1 The size of the indexed World Wide Web

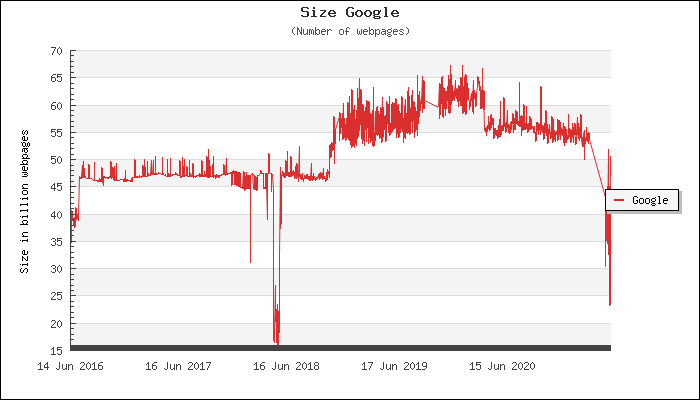
****

Figure 1.2 The estimation indexed web pages of Google.

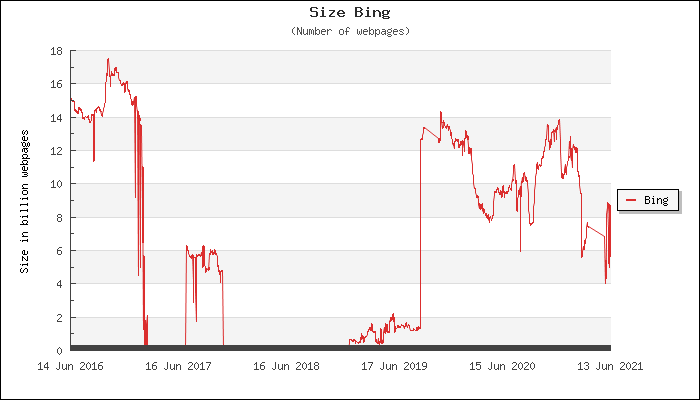
****

Figure 1.3 The estimation indexed web pages of Bing.

## **Literature Review**

In recent years, people keep finding out the actual indexed size of the search engine and the discovery are being more and more accurate comparing to the old research. However, the estimated indexed size of search engines may be different due to the year for research, the calculation method they use, and the language that they are being estimated.

### **Research in the pass**

* + 1. Estimating search engine index size variability: a 9-year longitudinal study [2]

This research is a 9-year research that starts from 2006 and end in 2016, here are some steps that they had done:

1. First, the researchers randomly select webpages with filtered the pages that are non-exist, with frames, server redirect beyond two levels, and client redirects from the DMOZ web directory, which include 531,624 webpages, contain 254,094,395 words with 4,395,017 distinct words.
2. Then, they select 28 pivot words from DMOZ data with the most frequent word and select the exponential series where they increased the low exponent viz. 1.6 into the selection rank. In the 28 pivot words that they selected, some words in the tail of the list having different languages rather than just having English although English having the dominant in the multilingual DMOZ directory.
3. Calculate the extrapolated value with all the 28 pivot words by equation 2.1 and estimate the arithmetic mean of all 28 pivot words. They will not include into arithmetic mean if that word did not occur in any documents. Although the words of selection are different, the result having a very close average number of the computed extrapolations which having the same exponential rank factor of 1.6. x is the
4. After knowing the arithmetic mean of 28 words, they compare the 28 pivot words in DMOZ document frequencies with other 4 source, which are the newspaper articles in New York Times corpus which published between 1993 and 2001, the newswire articles in Reuters RCV1 corpus, the encyclopedic articles without including the redirect page and disambiguate page in English Wikipedia and the random generate DMOZ sample webpages which having the same source with training set but not overlapping. They find out that the sample of the DMOZ just underestimated 1.3% and more accurate compared to other sources.
5. These 28 pivot words come from 3 different frequency ranges (high frequency range, mid-range frequency, and low-frequency range) and the word will be overestimated or underestimated in high-frequency words and low-frequency words but using the equation 2.2 for calculating the collection of webpages estimation of accurate mean, it will become more accurate and reasonable. Also, they mention that the estimation of pivot words may be having a large standard deviation due to the DMOZ occur frequency and Wikipedia having a weak correlation (Pearson’s R = 0.48).
6. Having the result in the upper step, they select 3 words in difference frequency (the in highest frequency range, basketball in mid-range frequency, and illini in low-frequency range was selected) for estimating the total number of web pages and recording changes from March 2006 to January 2015. Also, they record the average of the estimated number of web pages with 28 pivot word frequency changes from Google.
7. Find the estimated indexed web pages (α) with the equation 2.2 and the result returned by the search engine, then point out and explain the changes of what they can see. They estimated that the number of web pages in Google between March 2006 and January 2015, shows that Google indexed website keeps rising from not enough 30 billion to around 48 billion, and Bing dropped from around 4 billion to around 3.5 billion. They also mention that the indexed webpages having a dramatic change in Google search engine, which having around 47.5 billion difference in Google (highest estimation is around 49.4 billion in mid-December 2011 and lowest estimation is around 1.96 billion in November 2014), then raise back to around 45.7 billion in January 2015. Bing also has a dramatic change which has around 22 billion difference (highest estimation is around 1 billion in November 2010 and lowest estimation is around 23 billion in January 2014). Also, the research result points out that the keyword they selected in mid-range frequency is close to the average estimation with overall 28 words. Also, all three words and the average estimation having a similar overall trend but having some minor differences in the estimation results.

Equation 2.1 Equation for calculating the number of the document occurs in DMOZ

Equation 2.2 Equation for calculating the estimated indexed web pages in search engines.

* + 1. Estimating the size of Arabic indexed web content [6]

This research is estimating the number of web pages that are written in Arabic. Here are some steps that they have done:

1. They decide to use Arabic Wikipedia for their source in their research because Arabic Wikipedia provides a diversified from different fields and disciplines which having the largest number of Arabic articles, having a broad diversity of words and terms, and a cleaner and careful article can be found.
2. Calculate the word frequency and the document frequency which are the total number of occurrences of the word in the Arabic corpus and the total number of articles in the corpus having the word.
3. Counting the word frequency in word frequency and document frequency respectively. The top 20 words with the highest frequency are selected and sort the word frequency and document frequency into one frequency list. Calculate the word frequency with Zipf’s law using the logarithmic formula shown in equation 2.5.
4. Requesting the Google, Yahoo, and Bing search engine. Save the first 10 returned URLs from each query and calculate the cross-overlap percentage in search engines. They will search by URL if the URL had not shown in the first 10 returned URLs in another search engine. They will be marked as also indexed in another search engine if the first returned result exactly matches which search by the URL. By the above calculation, it can calculate the cross-overlap percentage of the search engines and the counting the unique URLs indexed by each search engine.
5. After calculating the cross-overlap percentage, they use Zipf’s Law distribution for selecting 36 words from word corpus and using the equation 2.4 to calculate the estimation of indexed web pages which Ep is the estimated number of webpages, DFs is the document frequency per search engine, DFc is the document frequency in the corpus and Nc is the total number of documents in the corpus. The equation that they used is following the methodology of Kunder [1] which is the document frequency per search engine divided by document frequency in the corpus and multiply the total number of documents in the corpus.
6. They point out that Yahoo having the largest indexed Arabic webpages, Google having the 2nd largest indexed Arabic webpages, and Bing having the smallest indexed Arabic webpages, which indexed around 1.9 billion, 550 million, and 248 million respectively. They mention that the indexed webpages having an upward trend between 14 November 2010 and 17 November 2010 due to some special events happen in Arabic regions, which points out that the news, special events, or activities may affect the indexed webpages in search engines. Also, the report mentions the order for testing the overlap percentage may affect the estimated indexed Arabic webpages, such as using Yahoo as the first order will having a bigger estimated result than using Google as the first order, which is around 2.1 billion and 1.3 billion respectively. At last, the estimated indexed Arabic webpages in their research is around 2.1 billion and the Arabic content on the world wide web has grown rapidly.

Equation 2.3 logarithmic formula for porting the Arabic word frequency.

Equation 2.4 Equation for calculating the estimation of Arabic web pages.

### **Chinese character**

* + 1. Wikipedia’s Chinese character list

Wikipedia [7] had done a Chinese character count project base on their data dump in 20061110 without the conversion between the Traditional Chinese and Simplified Chinese. In this project, they had contained a total of 85,902,608 words including 21,197 distinct Chinese characters in this data dump. The first 200 characters with the highest frequency are shown in Figure 2.1.

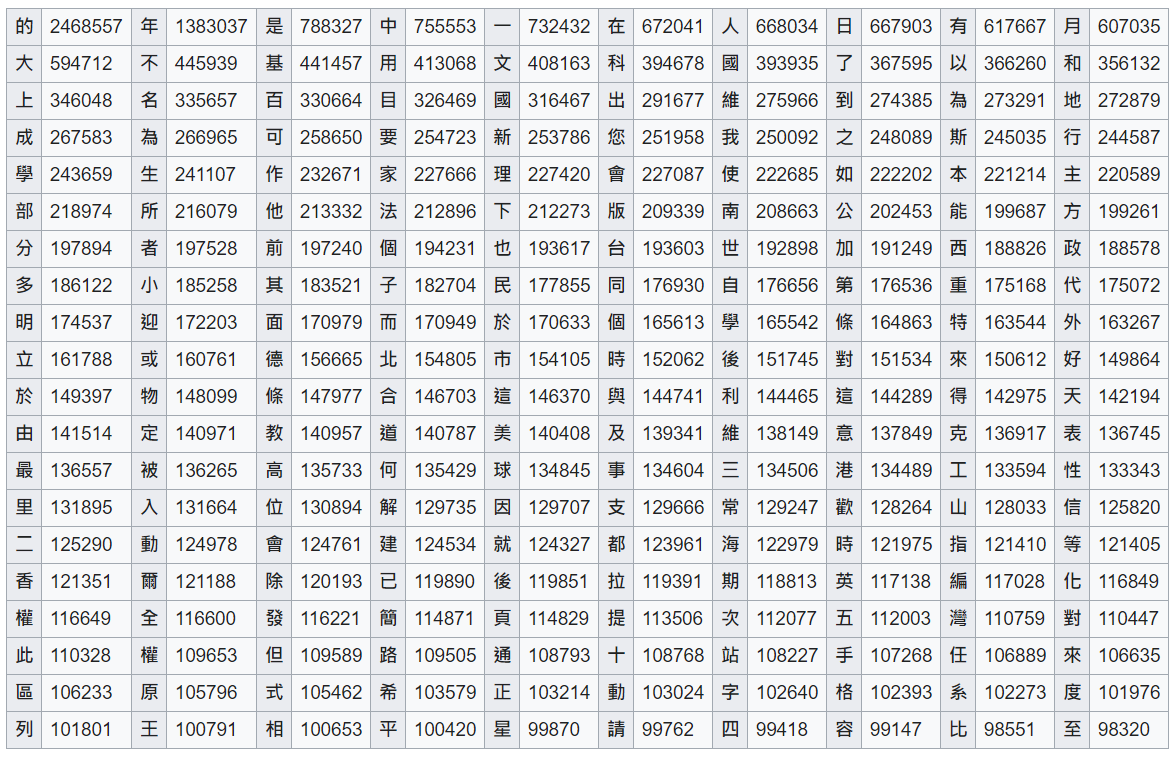


Figure 2.1 The first 200 characters with the highest frequency in Wikipedia 20061110 data dump

* + 1. Jun Da’s Chinese character count

Rather than Wikipedia has done research in Chinese character frequency, Jun Da [8] also done a Chinese character frequency research between 1998 and 2004, which had contained a total of 193,504,018 words including 9,933 distinct Chinese characters in the research. A part of Jun Da’s Chinese character research result is shown in Figure 2.2.

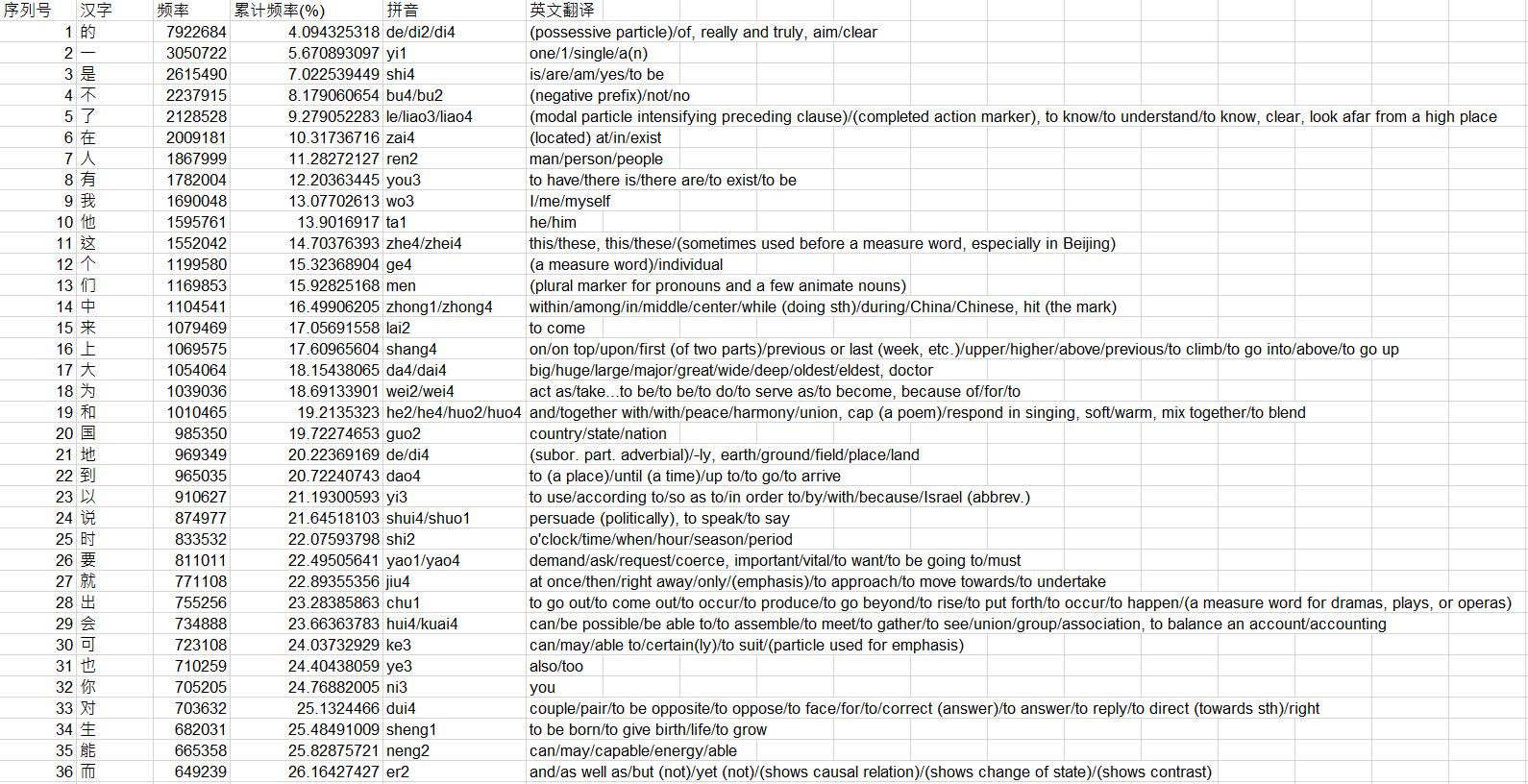


Figure 2.2 A part of Jun Da’s Chinese character research result

Here are the steps that Jun Da using in his research:

1. First, identify the corpus of the project. Jun Da mentions that his corpus will mainly focus on modern Chinese, which defined that modern Chinese is written on or after 1911 and classical Chinese is written before 1911. Jun Da’s corpus had collected the characters in different subject fields, such as computer science, economics, education… for informative way and the imaginative way he had collected general fiction, children, detective… which collect from different subject fields with a diverse range. He had collected the electronic publications with 16 online sources, such as the 中国青少年新世纪读书网 and 国学网络 which included the classical Chinese and Modern Chinese and release in printed format and the 中国科普博览 and 免费医院网 which release on the internet and providing the original text. Those online sources are released between 1997 and 2003. Also, Jun Da mentions that the imaginative texts and informative texts are different, which are related to literary works, entertainment, and related knowledge and/or information respectively.
2. After collecting the online sources, Jun Da wrote a PHP scripts for modifying the collected sources. The PHP scripts are mainly focused on removing the navigation strings, such as the 回首頁, 上一頁, 下一頁… and removing the headers and footers, such as 站長信箱, 亦凡書庫… for having an accurate frequency count result.
3. Then, is the segmentation of the characters. The method that Jun Da use is detecting the characters that are encoded in GB2312-80 or GB13000. It is because the writing technique of Chinese is totally different compared to English, which not had a whitespace between words or characters.
4. Lastly, Jun Da’s research points out the following points:
   1. Although the corpus in his research having 9,933 distinct characters, the first 1,056 character’s frequency has occupied 90% frequency of its corpus. Also, the first 2,838 characters had occupied 99% frequency and the remain around 7,100 characters just occupied 1% of its corpus.
   2. The percentage of informative texts is higher than the imaginative text, which occupies of Chinese are 55% and 45% respectively.
   3. Compared to the Changyong Zibiao list, Jun Da’s list is more comprehensive because the Changyong Zibiao list only containing around 99.23% of characters collected in Jun Da’s list.

### **Market share of the search engine**

* + 1. The search engine market share in the last 10 years by cogney.com.hk

In the search engine market share report by cogney.com.hk [9], it shows that Google having the largest market share in Hong Kong, Yahoo [10] having the 2nd largest market share in Hong Kong, Bing having the 3rd largest, and Baidu [11] having 4th largest market share in Hong Kong, which shows in Figure 2.3 respectively.

Another point that cogney.com.hk’s report had mentioned is that the global search engine market share. Google had been increased from around 50% in 2010 to around 90% in 2019, the search engine market share in global of Yahoo had been decreased from around 45% in 2010 to no more than 15% in 2019 and the gap between the largest and the 2nd largest market search in global become bigger and bigger, which had shown in Figure 2.4.

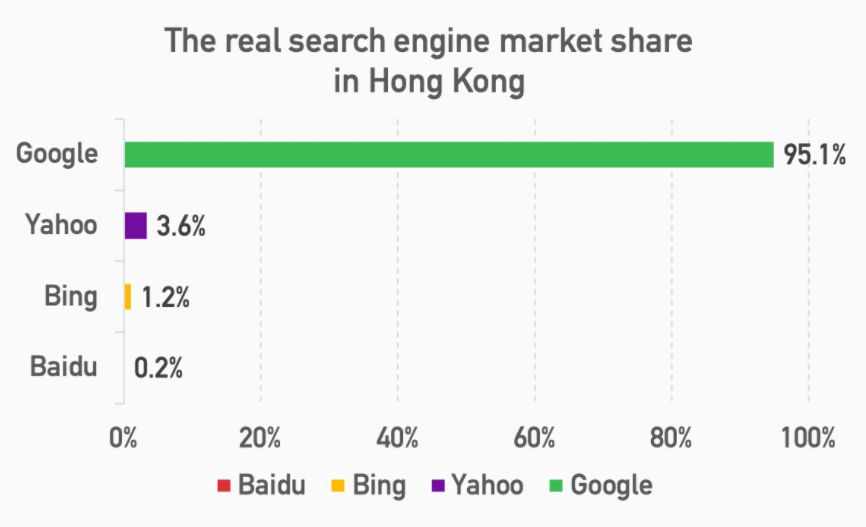


Figure 2.3 The search engine market share of Hong Kong by cogney.com.hk.

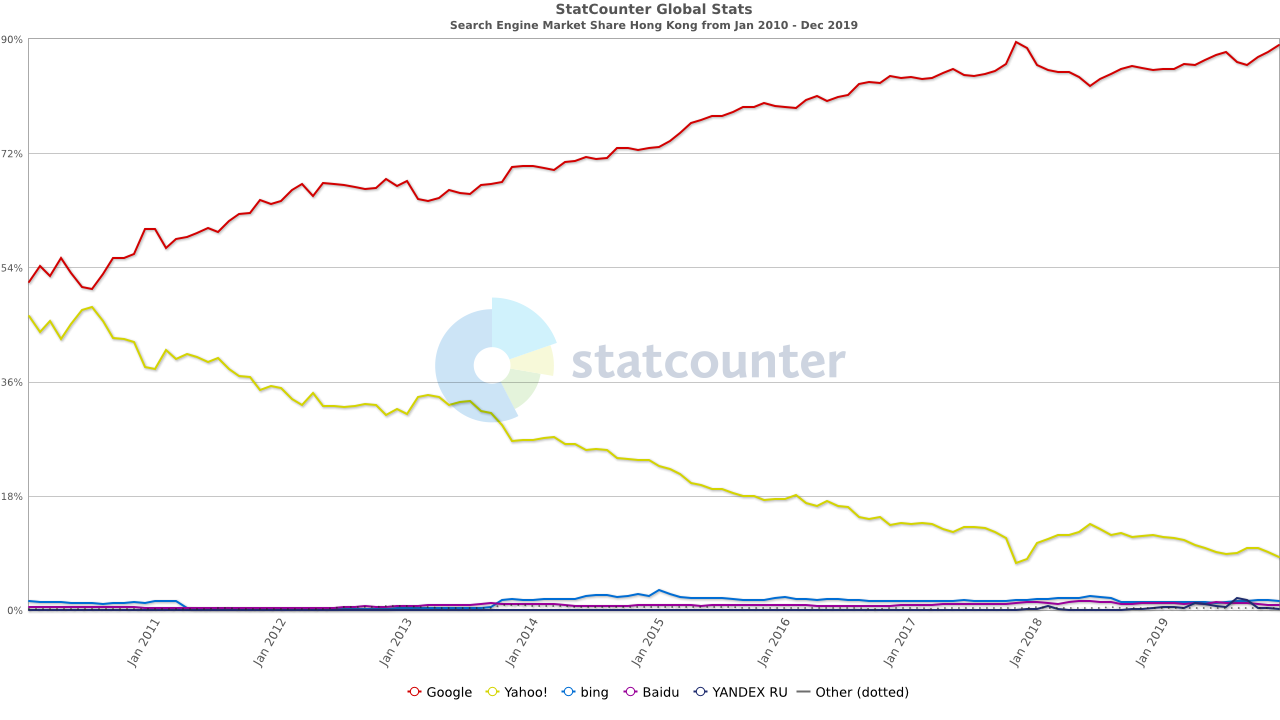
This research mention why the market share of Yahoo keeps reducing due to Google provides better search results and Google having a clear interface that is more user-friendly compared to Yahoo. They also mention about Hong Kong is a bilingual city which having Cantonese and English speakers but the users of Yahoo mostly using Cantonese for searching, so the flow searching with Yahoo becomes shrivel. Another reason that Google’s market share become bigger and bigger is that due to more than 92% of citizens lives in Hong Kong having smartphones and 96% of those people will surfing the internet daily so they can get the information easily, such as Google shut down their server in China (google.cn) 10 years before which is one of the main factors let Google being famous. 

Figure 2.4 The search engine market share global from 2010 to 2019 by cogney.com.hk.

* + 1. The search engine market shares nowadays by statista.com

The market share static by statista.com [12], shows that Google, Bing, Yahoo, and Baidu had 86.6%, 6.7%, 2.71%, and 0.54% of market sharing respectively on 21st February 2021. The statista.com shows that Google and Bing having the largest and the 2nd largest market share in the world and mention that the market share is not balanced since Google released in 1997. It shows in Figure 2.5:

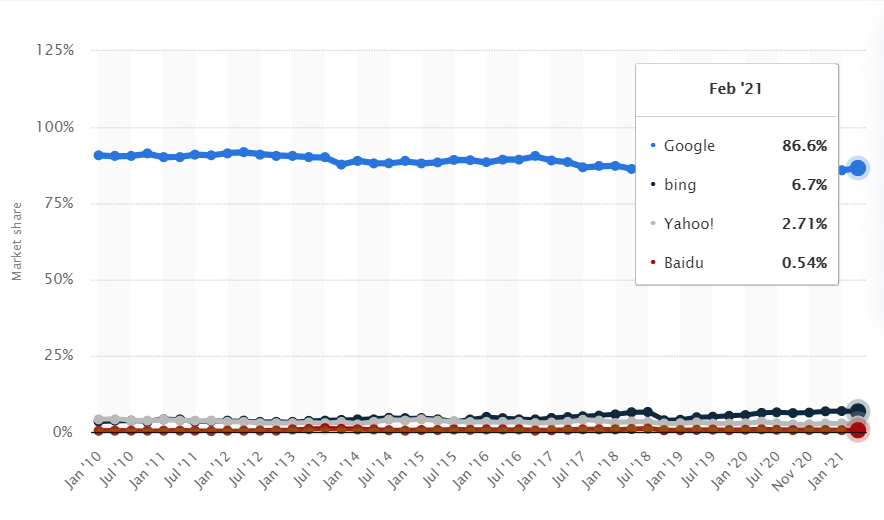


Figure 2.5 The market share in January 2021 by statista.com

### **Other related to this research**

* + 1. “Unusual Traffic Error” by Google search engine.

In the Google search help website [13], Google has explained why the user getting the “unusual traffic error” (or call “unusual traffic block”) from their search engine. Google will consider users as automated traffic due to user acts like the computer program, automated service, software for scrapping Google (search scrapper), or robot. The block will be disappeared when the unusual traffic is gone and recover back to normal.

Another article in lifewire [14] mentions that the reason why receiving the unusual traffic error is that google thinks your computer is a malicious bot due to having a huge traffic request or response in a short period of time. Figure 2.6 shows the response unusual traffic block of Google search engine.

Some people having the same question in their research in the past, what they had done is that let their program act like a human [15]. It is because the programmer who builds the web pages or search engines does not want the content to be scrapped.

In one of the articles in stackoverflow.com [15], user CodeCaster mentions some solutions to solving or avoiding this problem:

1. Adding the request header and cookie inside the request to let your user agent acts as the browser. It is because different users may have a slightly different result due to their browsing history. The search engine can determine that your request is from the program if the request does not have the items mentioned above.
2. Having some responses to the search engine, such as the mouse event or input actions to avoid the search engine judged your request is not human.
3. The requesting frequency also is one of the main factors for judging the request is from the program or human. If the request frequency in one second is too high, it will determine the request is from the program, not from a human.



Figure 2.6 Unusual Traffic Block returned by Google search engine.

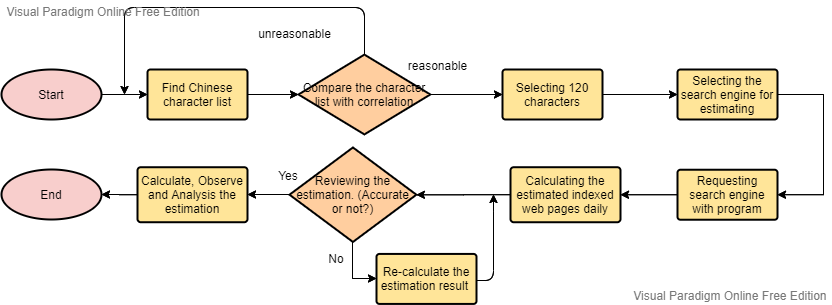
* + 1. How Google indexing webpages in their search engine

The article was written by Jesse and Nissan [16] mentions that nowadays having trillion unique website links and the size of the worldwide web growing with billion pages per day. The definition of the size of the world wide web is due to how many web pages are useful to the users and it does not have a correct answer.

The Google search will not index every meaningless website in their search engine, such as those webpages are very similar to each other or the webpages which are generated by the auto-generation program because those webpages are not useful for users and not comprehensive to the content which indexed in their search engine. They also mention that if using the size of web-link that Google had collected, it may be around 50,000 times bigger than the US Map which has 50,000 times more intersections and roads.

## **Methodology**

This chapter is about the methodology of the project for how to estimate the indexed size of the search engine. Before calculating the estimation of search engine indexed size, some activities are required to be done for the accurate estimation. This chapter will be divided into multiple parts, as 3.1 is about the preparation before requesting and scrapping, 3.2 is about the requesting and scrapping part and 3.3 is about calculating, analyzing, and observing the estimation. The whole program flow chart will be shown in Figure 3.1.

Figure 3.1 The whole project flow chart

### **Preparation before requesting and scrapping**

By reading the literature review that related to this project, here are some processes that needed to be done before doing the estimation in this project.

* + 1. Find the Chinese character list.

There are many Chinese character list in the world, such as searching “Chinese character list” in Google search, it returned result is around 504 million that had indexed in their search engine. So, the character lists are required to determine its accuracy of the list due to characters between the classical Chinese and the modern Chinese (vernacular Chinese) having a big difference for checking the character list is outdated or not and checking the list is close to the Chinese using in nowadays which will affect the estimation results.

* + 1. Find the Chinese publications for comparison.

The Chinese publications are required to collect for checking the correlation of the above character list. Different types of Chinese publications are being collected, such as the publication written in classical Chinese and traditional Chinese, different types of publications, publications that published or written in different years, and the publications that publish in different countries. Those publications will convert into text format for saving and it will generate two files. One of them will save the characters, the frequency of the character, and the occurrence rate of the character, and another file will save the characters that had been appeared in the text file in descending order and the total word count.

* + 1. Comparing the correlation with character list and publications

The correlation is one of the parameters that can use to concern the characters list is suitable using in this project or not. Another Chinese character list will be selected if the correlation between the list and those Chinese publications is too low which calculate by Microsoft Excel with the correlation function. Equation 3.1 showing how to calculate the correlation in which γ is the range of the Chinese character list and δ is the range of the Chinese publications.

Equation 3.1 The correlation in Microsoft Excel

* + 1. Select the Chinese characters

After checking the Chinese character list, it needs to determine the selection rules for selecting the Chinese characters in the list that find in above. By reviewing the research mentioned in the literature review, the character selection rule that uses in estimating by English [2] having a more accurate and comprehensive selection compared to estimating by Arabic [6]. It is because the estimation by Arabic is just selecting the top 20 words in terms of frequency and document frequency which causing the inaccurate estimation. But the method using in estimation by English is selecting 28 pivot words in DMOZ corpus with different frequency rankings, taking the arithmetic mean of all 28 pivot words, and selecting 3 pivot words from three frequency ranking which has comprehensive and accurate estimation results.

So, the selection rules of Chinese characters in this project are as follows:

1. Selected characters must obtain in different frequency rankings.

Selected characters from different rankings can have a better estimation result and will not cause overestimation or underestimation.

1. Selecting more characters and equally selected in all rankings.

In both estimation by English and Arabic, their research only selecting no more than 30 characters in the corpus which may have not detailed enough results. So, 120 characters will be selected in the list for calculating a detail and better estimation indexed webpages in the search engine.

For the efficiency of knowing the frequency of different character lists, Zipf’s law will be used for plotting the frequency of the characters, which principle is similar to the estimation by Arabic [6].

* + 1. Select the search engine.

From the above literature review, both cogney.com.hk [9] and statista.com [10] research point out that Google having the largest market sharing of search engine in the Hong Kong market and global market respectively. In the cogney.com.hk’s result, it points out Yahoo having a bigger market sharing than Bing in Hong Kong. But according to statista.com’s research, Bing having a larger market sharing than Yahoo in the global market. Also, both research mention that Baidu having the smallest search engine market share either is in Hong Kong or the global search market.

The main consideration for selecting the search engine is market sharing and the second consideration for search engine selection is the returned result of the search engine.

### **Data Collection and Program**

When the preparation part is finished, the next methodology is about the data collection and program part.

* + 1. Data that indexed in the search engine being collected.

The data in the search engine that is collected are the returned results and the indexed items which included the title, URL, and the description of the results. Figure 3.2 and 3.3 shows the data being collected, such as returned results which circle in red, and the indexed items which circle in blue.

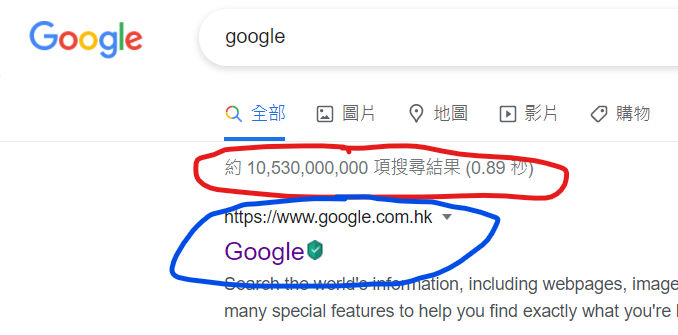


Figure 3.2 The search result by Google.



Figure 3.3 The search result by Bing.

* + 1. The programming language and the program packet are used.

For the high efficiency of scrapping the information in search engines, the program is requested to implement. After considering different program languages, Python [17] is the most suitable program language for this project because the library of python is powerful which the missions can be finished by just calling those suitable packets rather than required to write by yourselves.

* + 1. Data handling

Data is one of the main considerations in this project. The handling of data is very important due to those scrapped data in the above method will analysis later. The way that for saving the scrapped data is using two CSV/text file, one of the CSV/text file will save the characters and the returned result in search engine which named format is “{Search engine}\_{Date}\_results.csv/.txt” and another CSV/text file is saving the indexed item of the character, such as title, URL and description, which named format is “{Search engine}\_{Date}\_{Character}.csv/.txt”. The “{Search engine}” here is the name of the search engine that requesting, the “{Date}” here is the requesting date and the “{Character}” here is the character requesting the search engine which inside the 120 selected characters.

Another consideration is HTML raw data handling. To ensure that the scrapped data can be reviewed or checked in the future, the HTML raw data will be also saved during the scrapping actions and saving the HTML raw data in HTML format which file name is similar to the file with “{Character}” in above.

* + 1. Automation of the program

To ensure the program can be run daily and automatically, the consideration automation of the program. Assume that the program will run on a server or a computer that will not be shut down, the program will run at a certain time automatically and it will continue running daily until it stops by the users.

* + 1. Avoid and handle the “unusual traffic block” of Google.

As the literature review of Google [13] and lifeware.com [14], the unusual traffic block is due to huge data traffic requesting. In the discussion of stackoverflow.com [15], user CodeCaster mentions that the program needs to act like a human for requesting Google to avoid the block, so here are some methods that will be applied in the program:

1. The proxy address method.

Adding the proxy address can for requesting can misleading Google to thinks that the request is from different users and avoid meeting the unusual traffic block.

1. Break/waiting time method.

All the literature review related to unusual traffic block points out that huge network traffic will be causing this problem. So, the waiting time method is using in the program for reducing the network traffic and avoid meeting the block.

Another method for handling the blocking problem is checking the returned HTML size because the file size of a normal returned HTML must larger than a certain KB. If the size is smaller than a certain KB, which means that the program had met the block. The way to solve the block is to wait for certain minutes for the block dismissal.

### **Calculate, Analysis and Observer the estimation**

By reviewing the estimation by Arabic [6], the way that they are estimating the search engine is using the results returned by the search engine and divide the frequency of the character for getting the estimation indexed web pages of the search engine. The calculation method being used in this project is similar to the Arabic research use, but the method will slightly different due to estimating with different languages.

* + 1. Calculate the estimation indexed web pages.

The calculation method uses in this project is the number of results returned by a search engine (hereinafter referred to as “β”) dividing the frequency of the characters to get back the number of indexed web pages of a search engine that is unknown to us (hereinafter referred to as “α”). As I know that the probability of a character can be calculated by β divide by α. For example, the probability of the word “banana” is equal to 1%, which means that the β divide by α is equal to 1% (Figure 3.4 shows the β value circle in red is 413 million than the estimated indexed web page is 413 million divided 1% equal to forty-one billion three hundred million). So, the α value can be estimated by β divide by the character frequency which shows in equation 3.2. After calculated the α of each character, a daily estimated α will be calculated by taking the average value of all the character’s α values using the Excel function called “AVERAGE” which shows in equation 3.3. If the calculated daily estimated α value having a big difference from the previously calculated value, the Excel file for calculating the α value will be review and calculate all the α values again to ensure no careless mistake during the calculation.

Equation 3.2 Equation for calculating the estimation of indexed web pages use in this project.

Equation 3.3 Excel function “AVERAGE” for calculating the daily estimated α value.



Figure 3.4 The β value of the word “banana”

* + 1. Analysis and observe the scrapped data and the estimated indexed web pages.

As the data and the estimation scrapped and calculated above, here is some analysis and observe will be done in this project:

1. The rank changes and the quality of the search engine.

As Google [18] and Bing [19] mentions that their search results ranking is related to the relationship of your searching query, the last update of the web pages, user location, and the user experience to have a different searching rank in the user searches. So, this project will analyze the rank changes of the indexed item in the search engine. Another consideration related to the indexed item is the quality. The quality of the search engine can be analyzed by the repetition of the indexed item’s URL and description. The quality of the search engine is bad if the search engine indexed too many web pages having the same URL or same description.

1. The indexed size of the search engine.

The daily estimated α value calculated above can be observing the changes in the size of search engines, such as the size of indexed web pages become larger or smaller and the trend of the size in a short period of time.

## **Experiment**

This chapter is mainly focused on characters' frequency, program, and data analysis, and this chapter will be divided into multiple sections. As 4.1 is mainly focused on preparations before requesting/scrapping data, 4.2 is mainly focused on the program for scrapping the search engine, and 4.3 is mainly focused on the data analysis and indexed size estimation part.

### **Preparations before requesting/scrapping data**

Here are some actions that are required to be done before writing the program due to the accurate estimation.

* + 1. Select the search engine that being estimated.

The primary condition which search engine(s) is/are being selected for estimation. As what had proposed in sub-section 3.15 above, the main consideration for selecting the search engine is the market search of the search engine. Both cogney.com.hk [9] and statista.com [10] research point out that Google has the largest market sharing either in the Hong Kong search engine market or the global search engine market but Yahoo and Bing had the second-largest market share in the cogney.com.hk’s research and statista.com’s research respectively which those three search engines had occupied more than 96% of the search engine market share. So, the preliminary idea is to select the Google, Bing, and Yahoo search engines for estimating the indexed web pages.

Another consideration for selecting the search engine is the search engine's return results. The reason why the result that returned by the search engines is that the return results will affect the calculation of estimated indexed web pages. As shown in figures 3.2 and 3.3, Google and Bing search engine will show the return results/β which circles in red when users requesting their search engines. But Yahoo did not show the β value in their search engine which shows in figure 4.1 so Yahoo will not be estimated in this project. Maybe some people will ask that why do not estimate Baidu because Baidu’s β value is not accurate which will lead to a wrong estimation of indexed web pages. Figure 4.2 shows that Baidu will return the actual β value if the indexed web pages had not exceeded 100 million and figure 4.3 shows that Baidu will only return the β value with 100 million if the indexed web pages exceed 100 million so Baidu also will not be selected in this project.

After the above consideration, the Google and Bing search engines were selected to be estimated in this project.

Graphical user interface, application

Description automatically generated

Figure 4.1 The search result of Yahoo

Graphical user interface, text, application, email

Description automatically generated

Figure 4.2 The search result of Baidu which does not exceed 100 million indexed web pages

Graphical user interface, text, application

Description automatically generated

Figure 4.3 The search result of Baidu which exceed 100 million indexed web pages

* + 1. Considering of Chinese

The Chinese character had been classified into two types, one is traditional Chinese, and one is simplified Chinese. In this project, traditional Chinese was selected due to the following reasons:

1. The history of traditional Chinese and simplified Chinese

The period starts using traditional Chinese is much longer than simplified Chinese, which starts to use from the 5th century AD until now [20] and from 1956 until now [21]. It shows that period of using traditional Chinese is much longer than the simplified Chinese.

1. The misunderstanding of simplified Chinese.

The simplified Chinese characters may be causing misunderstandings in the meaning. It is because the simplified Chinese characters may have more than one meaning in a single word due to simplified Chinese using the homophone replacing other characters but traditional Chinese having a clear definition of their characters [22].

1. The size of the corpus and the written characters in simplified Chinese.

The comparison of simplified Chinese and traditional Chinese article [23] mentions that the work for simplifying Chinese is not enough. It is because the characters that people need to be learned are around 6,700 characters but the corpus of simplified Chinese only included 2,238 characters. Also, simplified Chinese only occupy around 33% and traditional Chinese occupy around 67% for every 100 written characters which also direct to not enough characters in the corpus of simplified Chinese.

So, traditional Chinese is selected in this project.

* + 1. Collecting the Chinese publications

Before comparing the correlation of the character list, collecting the Chinese publications are required. The collecting action is due to some considerations:

1. The types of Chinese publications.

The Chinese publication is the combination of different types of articles, books, or blogs, such as fiction, non-fiction, government press release, or articles written in colloquial Chinese. For the comprehensive correlation, different types of Chinese publications had been collected in this project.

1. The publish region, district, and country of Chinese publications.

Traditional Chinese is using in multiple areas in the world. Not only Hong Kong is using traditional Chinese, but also some other countries using traditional Chinese, such as Taiwan, Macau, Singapore, and Malaysia…. So, the publication from different countries is collected in this project.

Due to the above consideration, here are the Chinese publication that collected in the project shows at table 4.1. All publications are collected from the internet in PDF/text format.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Author/Editor | Publish area | Year of publication | Types |
| Golden Brother | Sit Ho Ching | Hong Kong | 2012 | Fiction release in the forum |
| Hong Kong Yearbook | Ng Wai Yee | Hong Kong | 2019 | Government publications |
| Chinese Language Learning Reference Chapter in Secondary School | Ho Man Sing | Hong Kong | 2007 | Teaching Materials |
| Lost on a Red Minibus to Taipo | Mr. Pizza | Hong Kong | 2012 (Part I),  2013 (Part II) | Fiction release in the forum |
| You are the apple of my eye | Giddens Ko | Taiwan | 2006 | Fiction |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| The Heaven Sword and Dragon Saber | Louis Cha Leung-Yung | Hong Kong | 1961 | Fiction |
| Romance of the Three Kingdoms | Luo Guanzhong | China | 14th century (Late Yuan Dynasty and Early Ming Dynasty) | Fiction |
| Classic Chinese Teaching Material | Education Bureau | Hong Kong | 2014 | Teaching Materials |
| Lost rainy season | You Jin (Tham Yew Chin) | Singapore | 1991 | Fiction |
| Call to Arms | Lu Xun (Zhou Shuren) | China | 1923 | Fiction |

Table 4.1 Chinese publications collected in this project.

In the above collected Chinese publications, some of the publications written in formal Chinese, such as teaching materials and government publications, some of them are written in spoken Chinese, such as Golden Brother which is a fiction that written by the forum member and published in a Hong Kong famous forum, and some of them are written in classical Chinese. The reason why collected publications are written in different types of Chinese because Chinese use in our daily life is the combination of different types of Chinese and knowing the changes of Chinese.

* + 1. Handling and counting the frequency of the collected publications.

As what had mentioned in sub-section 4.1.3, the collected publications were saved in PDF/text format. The way for handling the PDF file is by implementing a python program to converting the PDF file to text format. The python program with a packet called “pdfplumber [24]” was used for converting the PDF file to a text file. Figure 4.4 shows the original PDF file and the text file converted by pdfplumber.

After converting the PDF file, the content of the convert text file may have some repeated headers and footers which will affect the frequency counting. So, the text files are required to review and modify by the manpower before counting the word frequency. Figure 4.5 shows a text file example that having the repeated headers and footers.

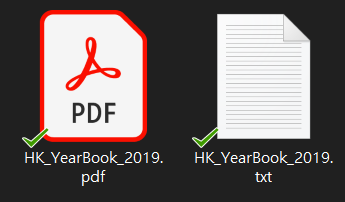


Figure 4.4 The original PDF file and the converted text file.

Text, letter

Description automatically generated

Figure 4.5 Text file having the repeated headers and footers.

For knowing the frequency of the characters inside the publications, the python program with method “counter [25]” was used to counting the frequency of each character. This program will be returned two files, one file containing the characters, frequency of the characters, and occurrence rate in CSV format, and one file contained the character which occurred in the converted text file and the total number of words of the converted text file in text format. Both CSV and text files are saving the frequency results in descending order. Figures 4.6, 4.7, and 4.8 show the name and content of the above CSV and text file.

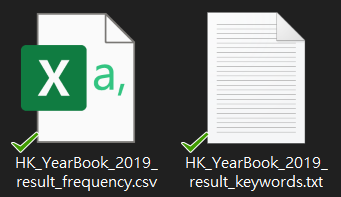


Figure 4.6 The name of the counted CSV and text file.

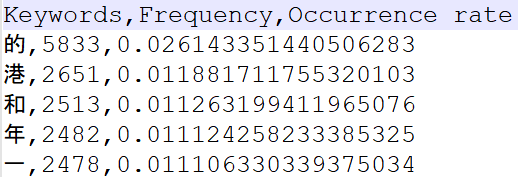
.

Figure 4.7 The content of the CSV file generated by the “counter” program.

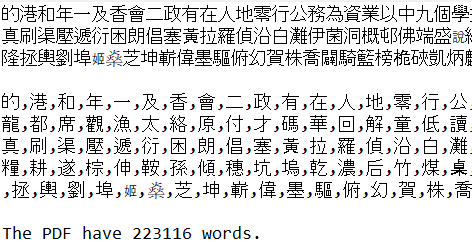


Figure 4.8 The content of text file generated by “counter” program.

* + 1. Modifying character frequency of the list.

Some modified actions are required:

1. Calculate the character frequency.

In this project, the character frequency is required for calculating the α value. But Jun Da’s character frequency list of Modern Chinese (hereinafter referred to as “Jun Da’s list”) only provided the cumulative frequency of the characters. The way for calculating the character frequency is the current cumulative frequency minus the previous cumulative frequency.

1. Translation of Jun Da’s list.

Jun Da’s list collected the characters in simplified Chinese but traditional Chinese is selected for this project, so translation is required. The way for translating the list is using the translate feature inside Microsoft office.

1. Combination of Jun Da’s list and Wiki list.

There were some repeated characters after the translation action. What had done is combine the repeated character’s frequency and re-rank the character list.

After modifying the list, the distinct characters inside the list become fewer, which dropped to 8,955 distinct characters from 9,933 distinct characters for Jun Da’s list and dropped from 200 distinct characters to 185 distinct characters for Wikipedia’s list.

* + 1. Selecting the Chinese character list

The character list was selected is character list by Jun Da [8]. The reasons why selecting Jun Da’s list are:

1. The long reference period and a certain degree of acceptance of Jun Da’s list and its research result.

Jun Da’s list and the research result were released in 2004. By searching Jun Da’s research in Google scholar, Jun Da’s research is still being quoted in research released in recent days. Jun Da’s list is also referenced in Chinese learning websites and Chinese frequency websites, such as studycli.org [25] and hanzidb.org [26], which means that Jun Da’s list and research result having a certain degree of acceptance.

1. The source and size of Jun Da’s list.

Jun Da’s research, mentions that the list included 16 online sources from different subject fields which had a comprehensive result of Chinese characters. Another reason is the size of the list. The size of Jun Da’s list is a combined 193,504,018 characters which having 9,933 distinct characters.

1. The academic record of Jun Da.

Jun Da is the commissioner in Linguistics and Chinese [27] which means the research done by him is having a certain degree of credibility.

Comparing the Wikipedia list [7] and the Jun Da’s list, Jun Da’s list counted 2.25 times more word frequency than the Wikipedia list, which counted 193,504,018 and 85,902,608 words respectively. Wikipedia’s list only listed out the first 200 highest frequency characters which will cause the wrong estimation in this project if selecting the characters from Wikipedia’s list.

Also, the correlation between character lists and the publications is another consideration. The way for comparing the correlation of Jun Da’s list, Wikipedia’s list, and Chinese publications are using Microsoft Excel and compare it one by one. In the Excel file, the characters that occur in both lists will be selected with their occur rate and using the Excel function called “CORREL” for calculating the correlation between 2 lists, which shows in equation 3.1. The correlation is between 0-1, the correlation if close to 0 means that two lists having low correlate, and close to 1 means that two lists having high correlate. In figure 4.9, it shows the Excel file for comparing the correlation of the 2 lists, which character highlight in yellow means that the character occurred in both lists.

Jun Da’s list correlation is higher than Wikipedia’s list in vernacular Chinese, which means that is more suitable for this project. Figures 4.10 and 4.11 show the correlation of Jun Da’s and Wikipedia compared to the Chinese publications.

So, Jun Da’s list was selected for this project.

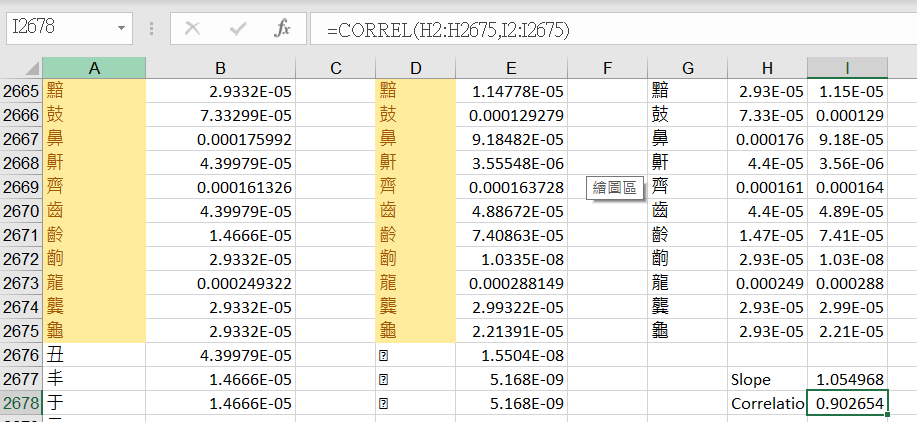


Figure 4.9 The Excel file for calculating the correlation between 2 lists.

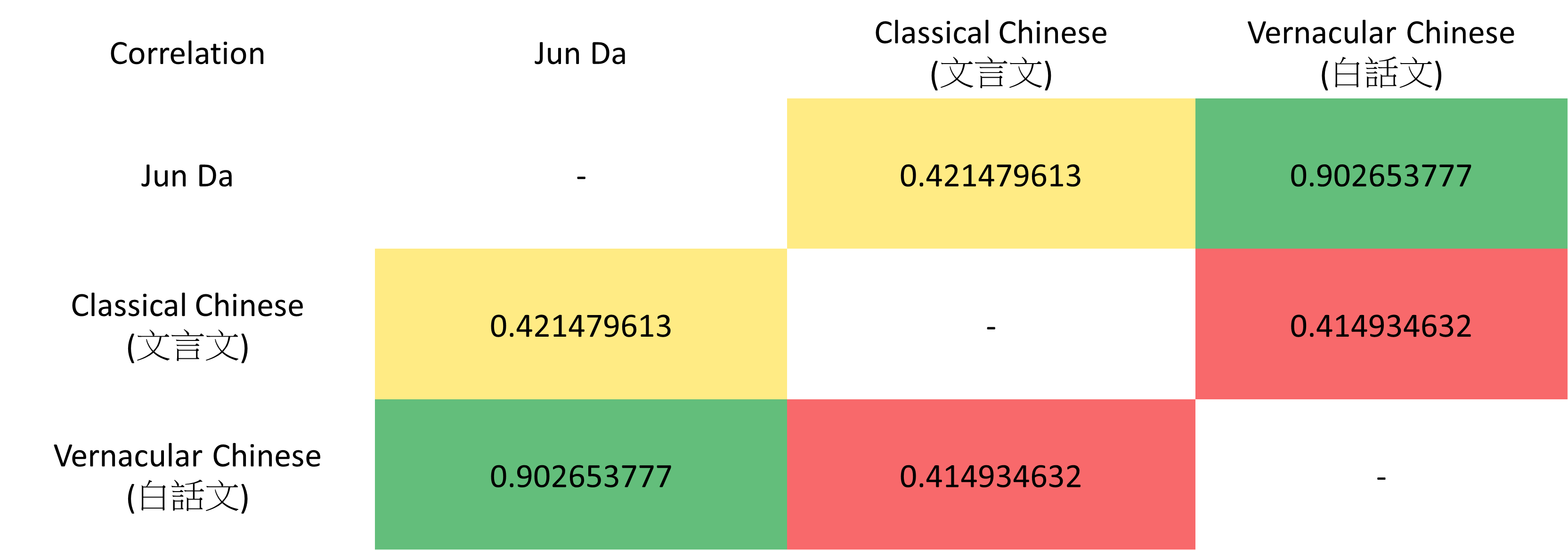


Figure 4.10 The correlation of Jun Da and Chinese publications

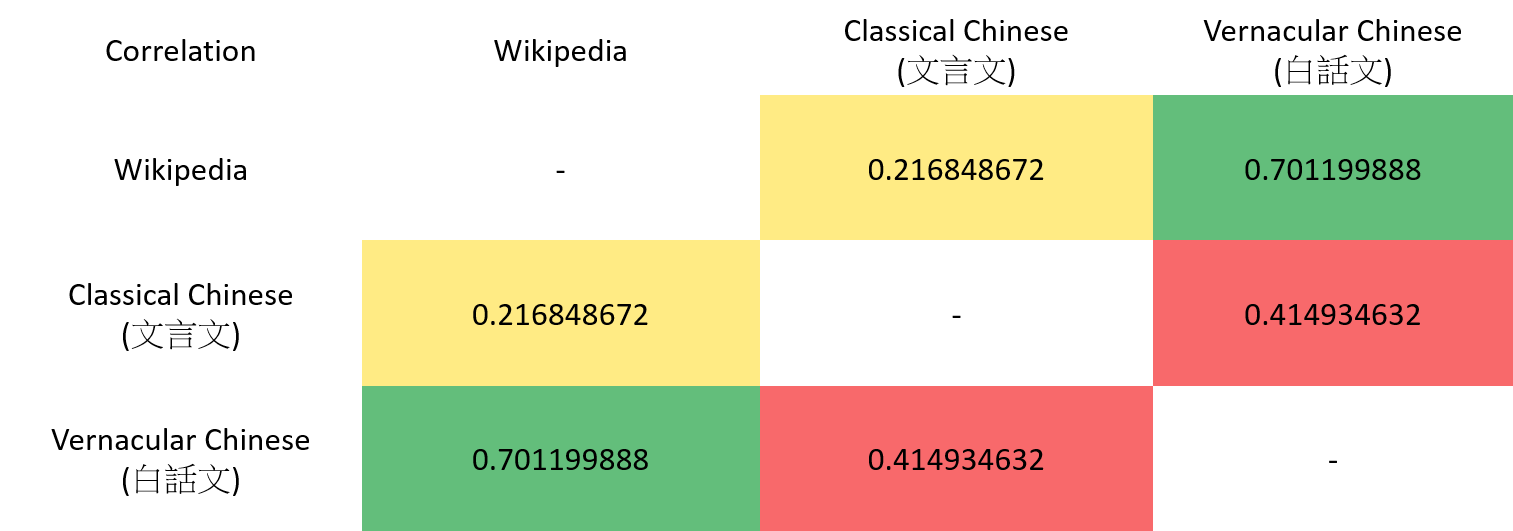


Figure 4.11 The correlation of Wikipedia and Chinese publications

* + 1. Selecting the characters for scrapping.

Then, plotting the character list’s frequency with logarithm base 10. In Figure 4.12, it shows that the ranking between 500 and 5,600 having a straight line compared to other rankings so characters for scrapping the search engine will be selected in this range.

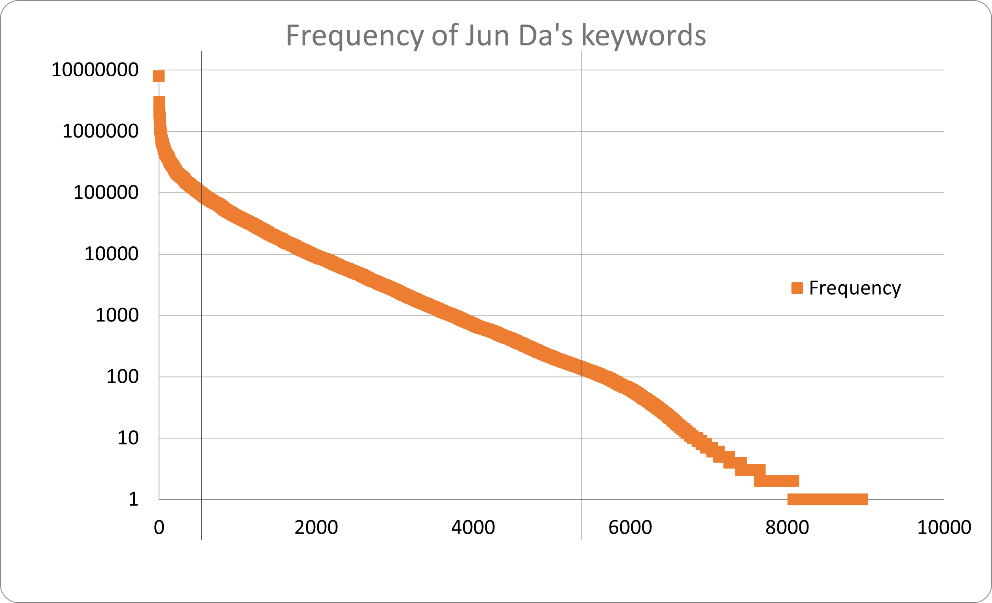


Figure 4.12 The chart for plotting the character frequency.

The “random shuffle [29]” packet with “for loop” was used for randomly selecting 120 characters in ranking between 500 and 5,600.

### **Program for scrapping the search engines**

This part is mainly focused on the program for scrapping the search engines. Figure 4.13 shows the flow chart of the program.



Figure 4.13 The program flow chart of the scrapping program.

* + 1. Scheduling and automation of the program.

Before explaining the program, here are some assumptions:

* The program is running on a server or a computer that will not be shut down.
* The program will continue to run until the user stop it.

The “schedule [30]” packet and “while” loop are used for handling the program automation to keep the program continue to run. The program will keep waiting until the computer time is equal to a certain time value and the format of time here is “HH: MM: SS”. The certain time value is set to 11 am for this project and it will run daily. Another main function of this program is for calling another program to run. Figure 4.14 show some code for the program automation.

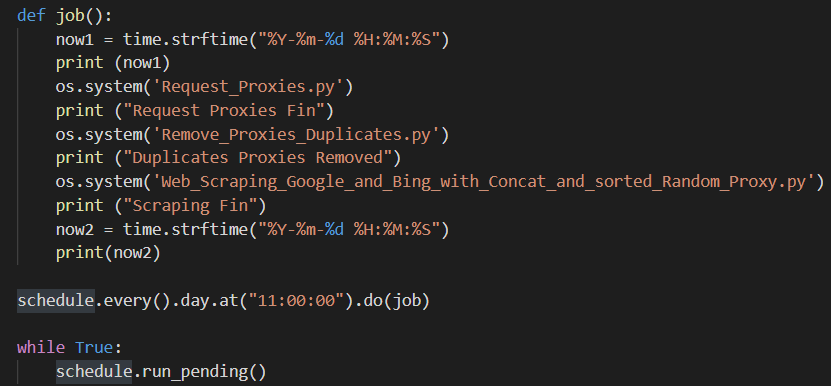


Figure 4.14 Code for program automation.

* + 1. Requesting the proxy list and removing duplicate proxies

The second program being introduced is the program related to the proxy. As what had mentioned in sub-section 3.2.5, one way for avoiding the unusual traffic block of Google is using the proxy.

The methods being used for renewing the proxy list daily were called “Beautiful Soup [31]” and “request [32]”. The reason why using Beautiful Soup and request is that the packet just required the programmer to encode few lines of code for parsing the HTML web pages and getting the information easily, such as the proxy address and port number from proxynova.com is save in the “script” tag inside 1st “td” tag the “and inside the 2nd “td” tag with align equal to left respectively. Another website used for renewing the proxy list daily is the free-proxy-list.net. The structure for saving the proxy address and port number is much simple than proxynova.com, which just saves all the proxy address and port number with the “div” tag and the class called “modal-body”. The proxy list was saved by named “online\_proxies\_of\_{Date}” in CSV format which using the “datetime [33]” packet for getting the date. Figures 4.15, 4.16, 4.17, and 4.18 show the HTML code and the program code for getting the proxy address and proxy port from proxynova.com and free-proxy-list.net respectively. The reason why saving the file in CSV format is that CSV format can be recalled back to the program easily due to the CSV using a comma to separate its data and can be read into the Excel easily due to its method of data handling.

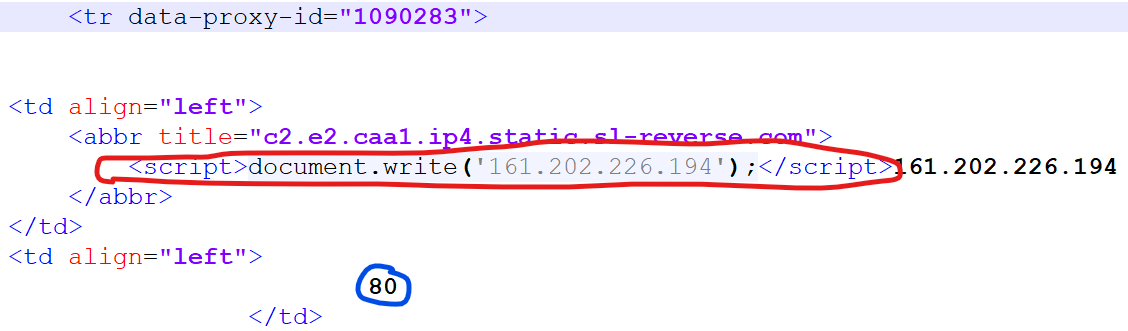


Figure 4.15 The HTML code of ProxyNova.com.

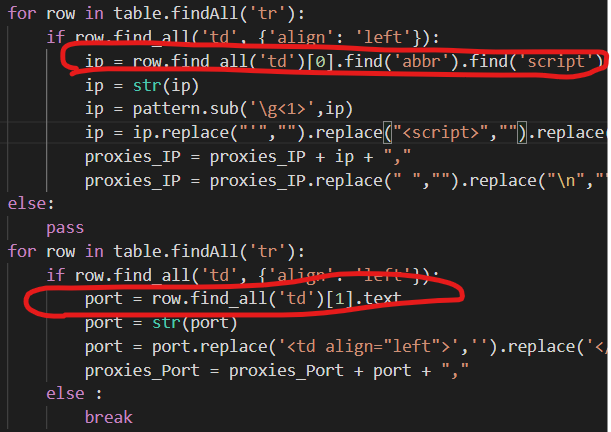


Figure 4.16 Code for scrapping the proxy address and port in ProxyNova.com.

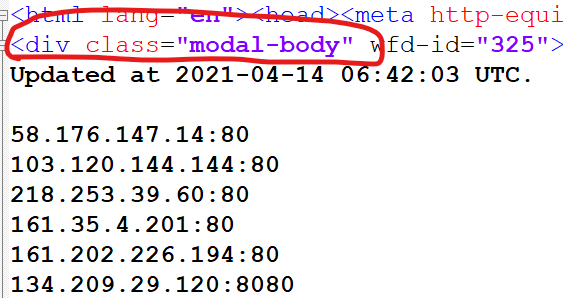
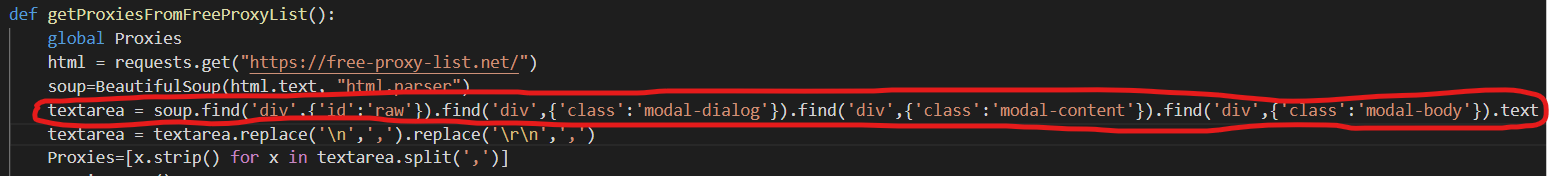


Figure 4.17 The HTML code of free-proxy-list.net.

Figure 4.18 Code for scrapping the proxy address and port in free-proxy-list.net.

Another program for removing the duplicate proxies was written due to the proxy may recurring in different free proxy websites. Removing those duplicates is to avoid requesting search engines with the same proxy address in a short period of time, meeting the unusual traffic error, and ensure having equal chances of using every proxy. The program used the “for loop” method was written to removing the duplicates and saving a new CSV file for proxies without and duplicates name “online\_proxies\_of\_{Date}\_without\_Repeat.csv”. Figure 4.19 shows the code for removing duplicates in the proxies’ file.

Both programs will be called by the automation program mentioned in 4.2.1.

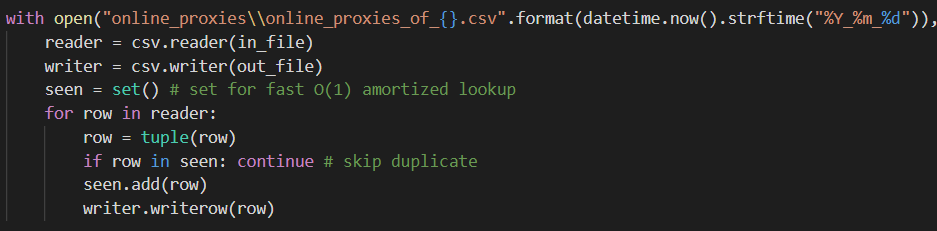


Figure 4.19 Code for removing duplicate proxies.

* + 1. Validating the proxies and scraping the data from the search engines.

In the scraping program sub-section, it will divide into multiple division for clearly explain:

1. The proxy availability checking.

Before the scraping program scrapes the data from the search engine, it required adding a proxy address into its request. But the problem is the proxy address scrapped in sub-section 4.2.2 can be use or not and the request action will be causing the unusual traffic problem again if the proxy address cannot be work perfectly. So, the way to determine the proxy address is available or not is by sending a request with the python packet called “urllib.request [34]” and “urllib.error[35]”.

The proxy list CSV file will be opened at the beginning of this program and using an array for holding all the proxy addresses and port numbers. After passing all the proxy addresses and port numbers into the array, the random shuffle action will appear to re-rank the ranking of the proxy address and passing the 1st item into a function called “is\_bad\_proxy” for checking the availability of the proxy, which shows in figure 4.20. It will print the error message or the error code if the proxy is not validated at this moment and change the proxy address to the next one for checking the availability until a validated proxy was found. Also, the program will print a message with “Bad Proxy: {proxy address}” after the error message or the error code is printed.

On the other hand, the program will print a message if the proxy is validated, which structure is “{proxy address} is working”, and the program will be scraping the data from the search engine with that success proxy address. Figure 4.21 shows the code for checking the proxy’s availability.

The printed results show in figures 4.22 and 4.23 for unsuccess and success proxy.



Figure 4.20 Code for random shuffle the proxies and pass the proxy address to check.

一張含有 文字 的圖片

自動產生的描述

Figure 4.21 Code for validating proxies.

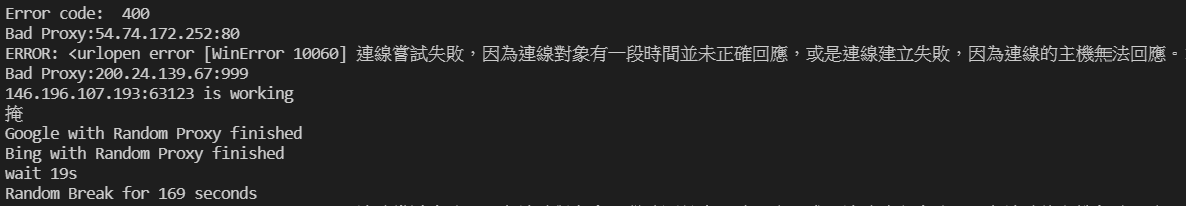


Figure 4.22 The error message for unsuccess proxy

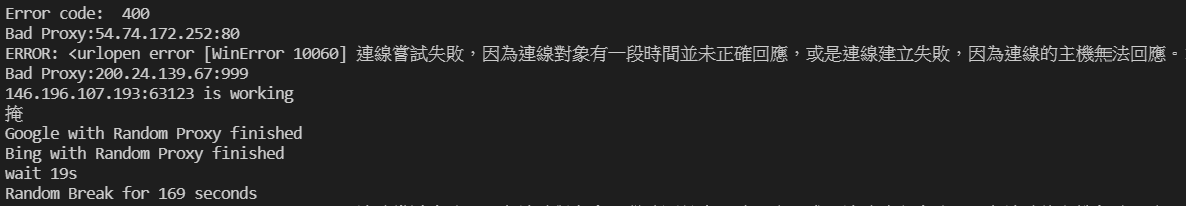


Figure 4.23 The message for success proxy

After checked the proxy can be used, the success proxy will pass to another function called “Web\_Scraping”. The below actions are done in the “Web\_Scraping” function.

1. Converting the character format.

The character that requesting the search engine needs to be converted into Unicode first. Why converting the character into Unicode is that the Bing search did not accept the request query that is not in Unicode format when the request is sent by the program.

1. Modifying the request URL and saving the raw data.

After checking the availability of the proxy address, the scraping action can be started. The method that is used in this project is beautiful soup and request, which had used for updating the proxy list in sub-section 4.2.2. But the way that for saving the scrapped data are slightly different, not only saving in CSV format but also saving in HTML format. The reason why the HTML format file is also saved is that the raw data may need to review if the estimation is not accurate. The returned raw data requested by the beautiful soup is called “soup” and the soup will copy itself into other soup for avoiding confusion inside the program. Figure 4.24 and 4.25 shows how to save the soup that was scrapped from the search engines.

一張含有 文字 的圖片

自動產生的描述

Figure 4.24 The saved HTML name and content of Google.

一張含有 文字 的圖片

自動產生的描述

Figure 4.25 The saved HTML name and content of Google.

Considering acting like a human or a browser, not only the proxy address being added into the request but also the header was added into the request. The header included the “Accept”, “Accept-Language”, “User-Agent”, and the “cookie” for camouflaging the request is sent by the browser. Also, another consideration is the number of indexed items being returned, the language, and the region in the search engine because the β value and indexed item will be different if the requested URL is different. To ensure getting the largest indexed items, the region is Hong Kong and the result is in traditional Chinese, those parameters “num=100”, “lr=lang\_zh-TW” and “google.com.hk” replaced “google.com” are added into URL for requesting Google and “count=50”, “setLang=zh-hant” and “cc=hk” are added into URL for requesting Bing. Why the result of Google is equal to 100 but Bing is equal to 50 only is that the maximum result that can be shown in Google is 100 and Bing is 50.

1. Scraping the data from search engines.

The scraping order is Google first and Bing at last. The reason why having this order is that it may meet the unusual traffic block when requesting the Google and ensure the time of requesting Google and Bing will not have a big difference to affect the estimation value. The way for checking the program meets the unusual traffic block or not is by checking the HTML file size. If the HTML file size is smaller than 150 KB, it means that the program had met the unusual traffic error and the program will request Google with the same character and same URL after 5 minutes until the returned HTML size is larger than 150KB.

After requesting the soup, the program will pass the soup to another function called “Scraping\_Google” and “Scraping\_Bing” for parsing the data inside Google and Bing. The way that Google saving the β value is using the “div” tag with the id called “result-stats”, so the program is finding the id called “result-stats” for getting the β value and saving into a CSV file named “{Search engine}\_{Date}\_result” which will save all the β value of the characters returned by the same search engine. Handling the Bing β value is similar to Google but the Bing saving the β value with the “span” class called “sb\_count”, so the program is written to find the “sb\_count”. Figures 4.26-4.30 show the name and content of the CSV file, the HTML code, and the program code for scrapping the β value inside Google and Bing respectively.

一張含有 桌 的圖片

自動產生的描述

Figure 4.26 The name and content of CSV file which save the β value.

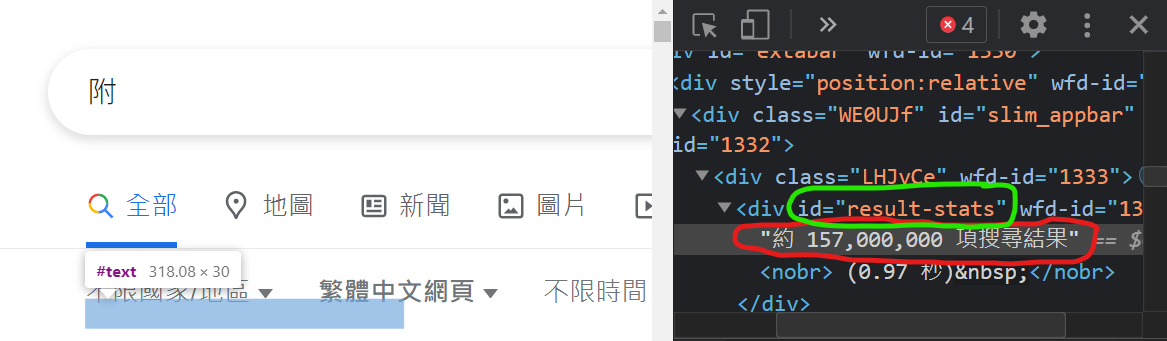


Figure 4.27 HTML code of Google

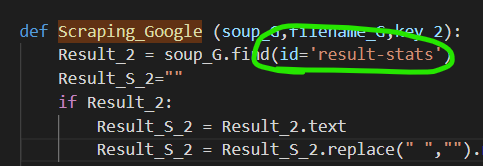


Figure 4.28 Program code for scraping Google β value.

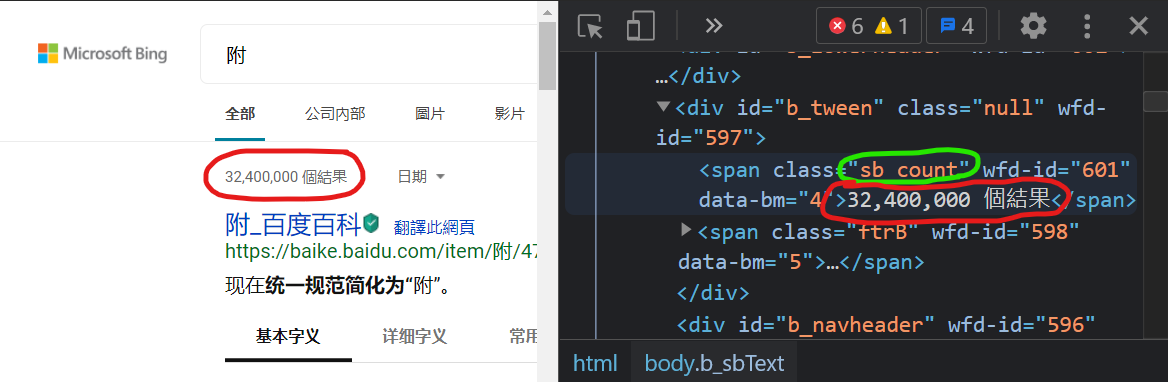


Figure 4.29 HTML code of Bing

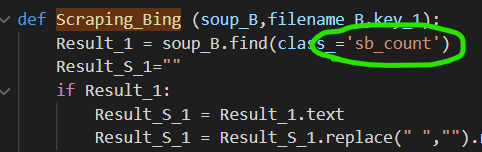
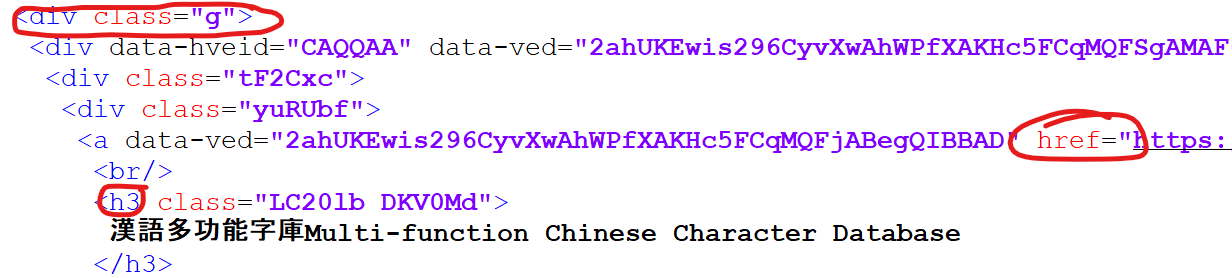
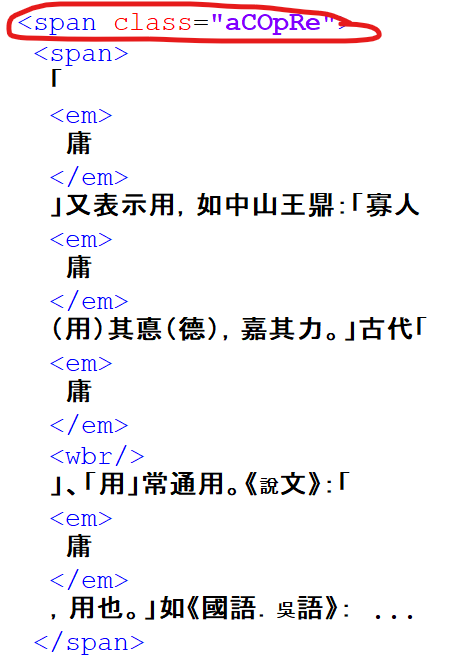


Figure 4.30 Program code for scraping Bing β value.

The program starts to scrape the items that are indexed in the search engine after the β value being scrapped. The indexed items included title, URL, and description that were returned by the search engine. The way that Google showing their indexed item in HTML is using a div tag with a class named “g” and the title is inside the “h3” tag, the URL is inside the “a” starts with “href” and the description is inside the class “aCOpRe” or “qkunPe”. The name, URL, and description will be saved inside another CSV file called “{Search engine}\_{Date}-{Character}\_sorted” until all 100 results had been saved.

For Bing, the method for scrapping Bing is similar but the difficulty is much higher due to the HTML structure of Bing is more complex compared to Google. Bing uses a class called “b\_algo” for showing their results, the title and URL are saving in the same “a” tag which inside the “h2” tag, the description is saved inside multiple tags but all inside the “p” tag under the “b\_caption” div class but the description of Bing were separated into multiple parts so the concatenation actions are required before saving. When the above items were found, they will save into a CSV file which similar to Google on above until all 50 results had been saved. Also, it will replace with “N/A” if the description cannot be found for handling the exception cases. Figures 4.31-4.36 show the HTML code and the program code for scrapping the data for Google and Bing respectively. Figure 4.37 shows the name and the content of the CSV file.



  
Figure 4.31 and 4.32 The HTML code of Google

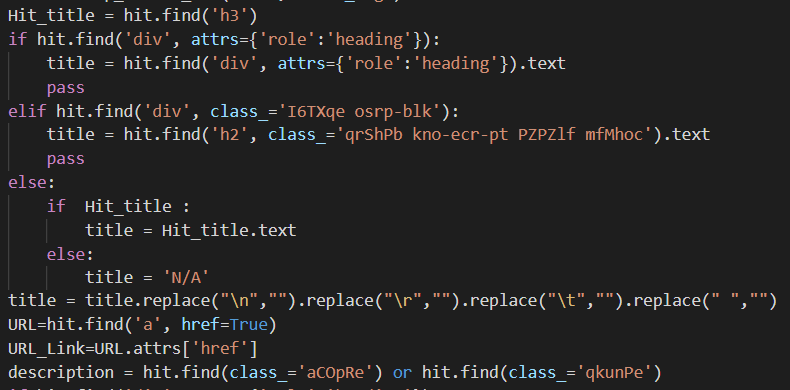


Figure 4.33 The code for scrapping the title, URL, and description in Google.

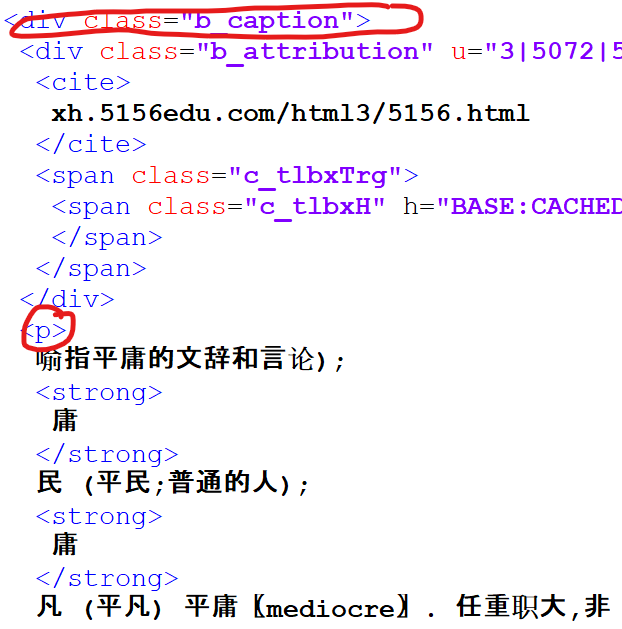
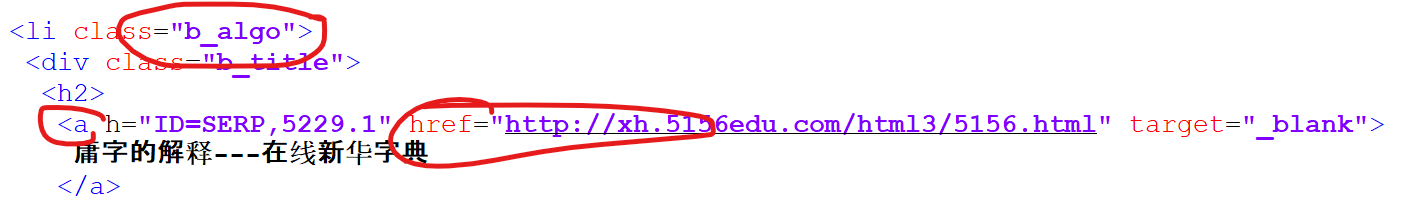


Figure 4.34 and 4.35 The HTML code of Bing

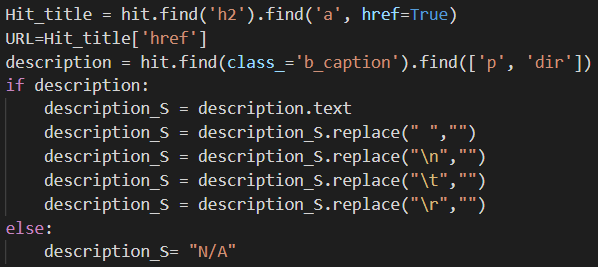


Figure 4.36 The code for scrapping the title, URL, and description in Bing.

一張含有 文字 的圖片

自動產生的描述

Figure 4.37 The name and content of the CSV file.

The program will change to the next proxy if the above A-D actions are finished once, and above A-D actions will repeat until all 120 characters finish scrapping.

* + 1. Other things related to the scrapping program.

The literature review mentions that the program needs to act like a human being to avoid the unusual traffic block of Google. Two methods had been mentioned above, which are the adding proxy method and adding headers into the request. It is not enough to just using these two methods for avoiding the unusual traffic error in Google due to Google's search engine is so clever.

In light of that, the third method for avoiding this problem is required. The third method being used in the project is that the waiting time/break of the program. It is because CodeCaster [15] mentions that the frequency for requesting the search engine cannot be too fast. To slow down the request frequency of the program, the “time.sleep” method is used for having a short break between 15 and 25 seconds after each request and having a long break between 90 and 250 seconds after 4 to 7 requests. The time of the short break, time of the long break, and the time for having a long break is generated by the “random.randint [29]” method randomly for avoiding the program being estimated by Google and trying to shorten the program run time. Figure 4.38 shows the success of scrapping action and program break with the printed message. Figure 4.39 shows the code of waiting time.

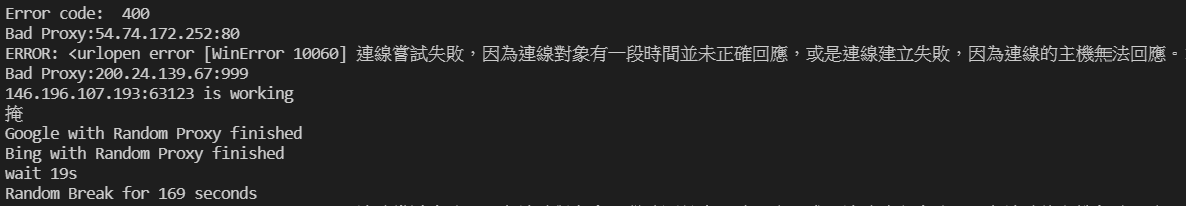


Figure 4.38 The message for success scrapping and program breaking.

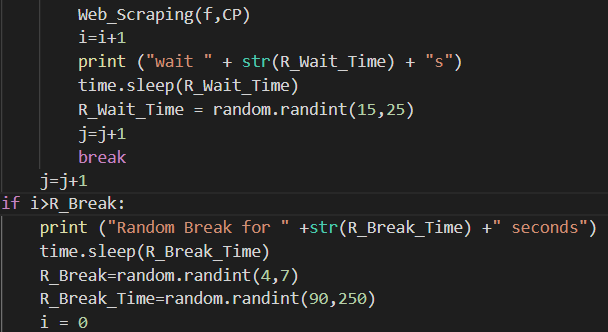


Figure 4.39 Code for waiting time.

Rather than the unusual traffic error, it is also required to handle the run out of proxy list problem. As what mentioned in 4.2.3, all the proxy needed to be checked before requesting the search engine but the proxy inside the proxy list can be used up and causing the error or bug in the program. To avoid the above problem is using the “mod” to get back the first proxy address to use because the program is using a counter for counting the number of proxies to be used. The counter will be mod if its value is bigger than the length of the proxy list file. The program is required to wait if the previous time minus the current time is no longer than 10 minutes for avoiding the unusual traffic error. The program does not need to wait, vice versa.

Last but not least, is about the saving method. All the files are saved or generated by the program are encoding in “utf-8” format. It is because the language for doing the estimation is traditional Chinese and it can be handling Chinese easily if encoding change to “utf-8” format rather than saving the garbled data or useless data. So, all the file either in CSV format or in text format, it will be encoded in “utf-8” format before saving.

### **Calculating, analyzing, and observing the estimation.**

When the scrapping action is finished, the calculating, analyzing, and observing actions will be started. To clearly explain, this section will be divided into two parts. Sub-section 4.3.1 explains the calculation method and the tools for calculating the estimation and sub-section 4.3.2 is explain how to analyze and observe the estimation results.

* + 1. Calculation method and tool for calculating the estimation.

The estimation by English [2] and by Arabic [6] was used in a different way for calculating the estimation value. For this project, the estimation method used by Arabic is more suitable because the method used by English required knowing the words that occurred in how many documents and the real versus estimated numbers which is unknown. So, the method for calculating the estimated indexed web page (α) is by the returned results (β) and the frequency of the characters, which is equal to equation 3.2 mentioned in sub-section 3.3.1. After knowing the estimated indexed web pages of each character, the daily average estimated indexed web pages will be calculated. The equation for calculating the daily average estimated indexed web pages is equal to equation 3.3 which is mentioned in sub-section 3.3.1.

The tool used for calculating the values mentioned above is Microsoft Excel. Before the calculation, the data are required to be modified. As what had shown in figure 4.26, the returned results were saved in the CSV file which included meaningless and useless information. The find and replace function inside the Excel is used to modifying the β value before calculating the α value, such as removing the “約”, “項搜尋結果(0.65秒)” in Google’s returned results and the “個結果” in Bing’s returned results. Figures 4.40 and 4.41 show the way to modify the β before calculating α and the content of the Excel file after modification.

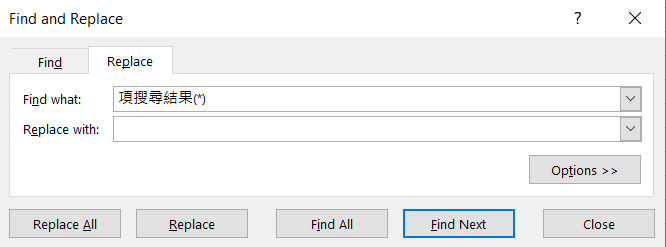


Figure 4.40 The find and replace the function of Excel.

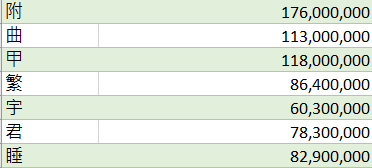


Figure 4.41 The content of the Excel file after modification.

Not only calculating the estimation of indexed web pages above but also the daily standard deviation, the daily standard error, the peak value, the valley value, the average value, and the difference of all estimation value had been calculated. Equation 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6 show the calculating the values in Excel that mentioned above. The μ here refers to the range for all daily estimations.

Equation 4.1 The equation for calculating the daily standard deviation.

Equation 4.2 The equation for calculating the daily standard error.

Equation 4.3 The equation for finding the peak value of daily estimation.

Equation 4.4 The equation for finding the valley value of daily estimation.

Equation 4.5 The equation for calculating the average value of daily estimation.

Equation 4.6 The equation for calculating the difference of daily estimation.

The method for saving the above data values is mainly divided into 2 Excel files, one of it is saving the daily results, which included the returned results CSV files generated by the program mentioned in division D of subsection 4.2.3, the estimated indexed web pages of each character, the daily standard deviation, the daily standard error, and the daily estimated indexed web pages. Another Excel file is saving the data in the period, which included the daily estimated indexed web pages, the daily standard deviation, the daily standard error, the peak value, the valley value, the average value, the daily percentage error, the average percentage error and the difference (with percentage) during this estimation period. Figures 4.42 and 4.43 show the content of those Excel files which name “Daily estimation.xlsx” and “Estimation results.xlsx”. Equations 4.7 and 4.8 show how to calculate the percentage error and average percentage error.

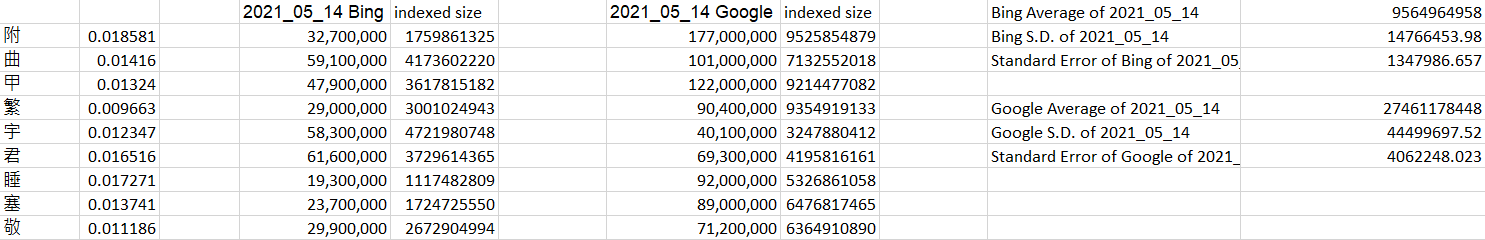


Figure 4.42 The content of the Excel file contains the daily data.

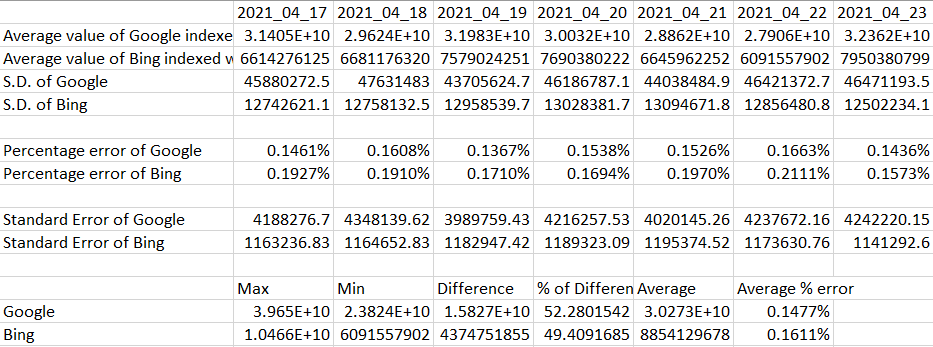


Figure 4.43 The content of the Excel file contains the data in the period.

Equation 4.7 The equation for calculating the daily percentage error.

Equation 4.8 The equation for calculating the average percentage error.

All the calculation actions mentioned above will be happening daily after the program is finished.

* + 1. Analyze and observe the estimation results.

After calculated those data, is about how to analyze and observe the calculation. The way to analyze and observe the data is using 3-line charts for reviewing the trend in a short period of time, which included reviewing the average indexed web page changes, the standard deviation changes, and the percentage error changes.

## **Results and Discussion**

After the above experiment, here are some results after the above experiment had done. This chapter will be divided into 2 parts, the main part is about the estimation of indexed web pages in the search engine, another part is about the Chinese characters.

### **Estimated indexed web pages and changes**

After running the program for multiple days, here are some results and changes that can be found. The average estimated indexed web pages of Google are around 29.7 billion and Bing is around 9.3 billion respectively.

The trend of indexed web pages in Google is around 31 billion in mid of April, start drop to around 24 billion in late April, risk back to around 39 billion, and start to drop and rise again in recent days. On the other hand, the Bing indexed web pages is around 6.6 billion in mid of April, keep increasing slightly to around 11 billion in the resent days. By comparing the indexed web page trend of the search engines, the indexed size of Google is much unstable compared to Bing. It is because the trend of Google shows that it is having more peak, or valley value compared to the Bing trend which having a smoother trend. Figure 5.1 shows the trend of Google and Bing and the changes of the estimated indexed web pages. Also, the gap between the two search engine trends had been reduced from 24.8 billion to 18.6 billion.

Comparing the results in this project with worldwidewebsize.com’s [1] result, Google indexed more than 100% of worldwide web pages and Bing indexed around 14% of worldwide web pages on 13th June 2021. The size of the indexed web page of Google is much bigger than Bing, which Google indexed web page size around 8.25 times more than Bing indexed. Another result is that the average Google result that estimated in this project is smaller than the result that estimated by worldwidewebsize.com but the average Bing results are bigger than the result estimated by worldwidewebsize.com, which is around 63% and 162% respectively. The reason why the results estimated in this project and estimated by worldwidewebsize.com is that the language used for estimating is different from this project, the language of this project used is Chinese, but the language used in worldwidewebsize.com is English. Another reason that may affect the estimation results is the calculation method. The method used in worldwidewebsize.com is calculating by document frequency per search engine (returned results) divided by document frequency in the corpus and multiply the total number of documents. But the method used in this project is the returned results divided by the character’s frequency results which may affect the estimated results.

Another comparison is compared to the 9-year longitudinal study [2]. The estimated size in their research is around 46 billion of Google and around 2.5 billion of Bing respectively. Compared to this project average estimated size, the estimated size of the 9-year longitudinal study (hereinafter referred to as “9-year study” or “9-year research”) is around 1.55 times and 0.27 times of this project estimation. Here are some reasons why having a big estimation difference between the 9-year study and this project:

1. The different languages used for estimation.

The language used for the estimation is different which had mentioned above.

1. The search engine indexing rule was changed.

The final point of the 9-year study is 20th January 2015 and the rule for indexing web pages in search engine Bing may be changed, so the indexed size of Bing having 3.5 times bigger than the estimation in 2015. Another result related to the search engine indexing rule is the difference/gap between Google and Bing. In the last estimation point of the 9-year study, the indexed web page size gap between Google and Bing is around 44.5 billion in 2015, the gap in 2021 is around 18.6 billion only, which had reduced more than a half.

Also, in the 9-year research, they point out that search engines having a dramatic change of their indexed web pages. From the trend of estimation in this project, its partially support the dramatic changes because the trend of Google having a big difference in 3-4 days but the trend of Bing just having a big difference at the being part and rising steadily starts from early May to late May, having a few days fluctuation and start to drop steadily from early June. So, the trends of Bing show that their indexed size is more stable than Google’s trend and did not have a dramatic change. The third reason is that the period for doing the estimation, the period for doing the 9-year research is more than 9 years but this project only contains around 2 months estimation so it may cause the dramatic changes of Bing cannot be discovered easily. Figures 5.2 and 5.3 show the percentage error and the standard deviation of the search engines.

So, what points out in the 9-year study that search engines having a dramatic change is partially correct only for this project.

Figure 5.1 The Google and Bing estimated indexed web page size from 17th April 2021.

Figure 5.2 The percentage error of Google and Bing.

Figure 5.3 The standard deviation of Google and Bing.

### **Chinese characters**

By the correlation of Jun Da’s list and the Chinese publications, I found that the characters use in modern Chinese having a huge difference compared to traditional Chinese. By the correlation compared in above, the characters in Jun Da’s list are similar to the modern Chinese because is having a high correlation with the publications written by vernacular Chinese and having a low correlation with the classical Chinese publications which means that the characters used in vernacular Chinese and classical Chinese are different and Jun Da’s list is written by vernacular Chinese. The correlation of Jun Da’s list, classical Chinese, and vernacular Chinese had been shown in figure 4.10.

Another result that I found is the modern Chinese characters changing. By comparing the Lost on a Red Minibus to Taipo and the Golden Brother which both are fictions written by forum members and released in a public forum, the correlation of these two publications is not very high which is around 0.61 only. Also, the correlation between Jun Da’s list and two publications mentioned above, the correlation between Jun Da and the Lost on a Red Minibus to Taipo is much higher than the correlation between Jun Da and Golden Brother, which having nearly 0.5 difference with their correlations. It shows that the modern Chinese characters using nowadays are still changing. Figure 5.4 shows the correlation between Jun Da’s list, Lost on a Red Minibus to Taipo, and Golden Brother.

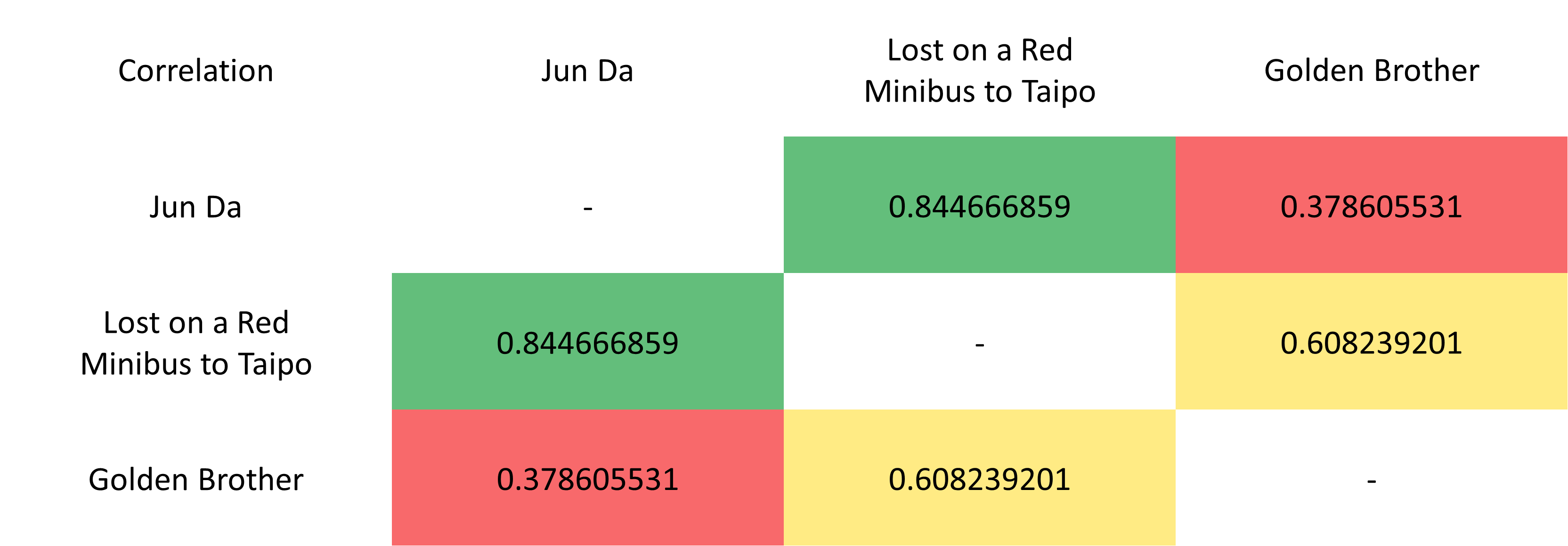


Figure 5.4 The correlation between Jun Da’s list and fiction release in a public forum.

## **Conclusion and Further Work**

In this project, the number of indexed web pages in the search engines was estimated, which is around 29.7 billion for Google and around 9.3 billion for Bing respectively. The general public can know which search engine can search for a better result and know how many web pages had been indexed in the search engine. Also, the trade of search engine indexed web pages had been shown that the Google’s indexed size is much unstable than Bing’s indexed size in the last two months, which keeps steady increased from 6.6 billion to around 11 billion for Bing and slightly decreased from 31.4 billion to 29.4 billion for Google respectively.

The code part of this project had been finished and the program can be run automatically and robustly, which can send the request to search engine daily and automatically. It had obtained one of the objectives of the project to record the returned result and the title, the URL, and the description inside the search engine.

The Jun Da’s list and the character’s frequency in this list had been determined which is still accurate nowadays by comparing different kinds of Chinese publications. The correlation of Jun Da’s list and other Chinese publications shows that the characters use in modern Chinese do not have a big difference between different distract and different years which has a high correlation between those publications written in modern Chinese.

As the quality of the search engines and the occupation rate of types of modern Chinese still did not mention in this project, here are some suggestions for future work:

* Compare the content in the same search engine. To determine the quality of the search engines, the way that can be done is to know the repetition rate of the description and the simulation rate of the URL. If the description’s repetition rate or the URL simulation rate is too high, it means that the indexed web pages most of the web page has the same content or come from similar sources which will affect the search results being incomprehensive and the information that users know is just the tip of the iceberg.
* Estimate the occupation rate of different types of Chinese. It is because the Chinese that we use today is a combination of different types of Chinese, such as fiction, non-fiction, written language, spoken language…. If the occupation rate of Chinese can be known, the estimation of search engine’s indexed web pages can be more accurate.

In addition, the estimation of the search engine indexed size is required to request more data for accurate and comprehensive estimated results.

## **References**

1. Kunder, Maurice de. “The Size of the World Wide Web (The Internet).” WorldWideWebSize.com | The Size of the World Wide Web (The Internet), www.worldwidewebsize.com/. [Accessed 03.06.2021]
2. Antal van den Bosch, Tonie Bogers, Maurice de Kuder, "Estimating search engine index size variability: a 9-year longitudinal study", Scientometrics (2016) 107 DOI 10.1007/s11192-016-1863-z, On page(s): 839-856
3. “List of Languages by Total Number of Speakers.” Wikipedia, Wikimedia Foundation, 14 May 2021, en.wikipedia.org/wiki/List\_of\_languages\_by\_total\_number\_of\_speakers [Accessed 03.10.2020]

1. “Google.” Wikipedia, Wikimedia Foundation, 25 May 2021, en.wikipedia.org/wiki/Google. [Accessed 03.10.2020]
2. “Microsoft Bing.” Wikipedia, Wikimedia Foundation, 17 Apr. 2021, en.wikipedia.org/wiki/Microsoft\_Bing. [Accessed 03.10.2020]
3. A. Alari and M. Alghamdi and M. Zarour and B. Aloqail and H. Alraqibah and K. Alsadhan and L. Alkwai, "Estimating the size of Arabic indexed web content", Scientic Research and Essays, 7(28):2472-2483, 2012.
4. “字數統計.” Wikipedia, Wikimedia Foundation, 5 Dec. 2006, zh.wikipedia.org/wiki/Wikipedia:%E5%AD%97%E6%95%B8%E7%B5%B1%E8%A8%88. [Accessed 03.10.2020]
5. Da, Jun. “A Corpus-Based Study of Character and Bigram Frequencies in Chinese e-Texts and Its Implications for Chinese Language Instruction.” 2004, doi:online publication.
6. PIERCE, DOUG. “The Real Search Engine Market Share in Hong Kong.” Cogney, 21 May 2020, www.cogney.com.hk/blog/the-real-search-engine-market-share-in-hong-kong. [Accessed 05.10.2020]
7. “Yahoo!” Wikipedia, Wikimedia Foundation, 15 May 2021, en.wikipedia.org/wiki/Yahoo! [Accessed 03.10.2020]
8. “Baidu.” Wikipedia, Wikimedia Foundation, 24 May 2021, en.wikipedia.org/wiki/Baidu. [Accessed 03.10.2020]
9. Johnson, Joseph. “Search Engine Market Share Worldwide.” Statista, StatCounter, 12 Mar. 2021, www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/. [Accessed 05.10.2020]
10. “‘Unusual Traffic from Your Computer Network.’” Google Search Help, Google, support.google.com/websearch/answer/86640?hl=en#zippy=%2Ci-dont-see-a-recaptcha%2Ci-shouldnt-be-getting-blocked. [Accessed 20.01.2021]
11. Mitchell, Bradley. “Why Does Google Report 'Unusual Traffic' From Your Computer Network?” Lifewire, 27 Jan. 2020, www.lifewire.com/unusual-traffic-from-computer-network-errors-3968165. [Accessed 20.01.2021]
12. ElektroStudios. “How to Bypass This ‘Unusual Traffic’ Detection on Website?” Edited by CodeCaster, Stack Overflow, 29 Dec. 2015, stackoverflow.com/questions/34510532/how-to-bypass-this-unusual-traffic-detection-on-website. [Accessed 27.05.2021]
13. Alpert, Jesse, and Nissan Hajaj. “We Knew the Web Was Big...” Official Google Blog, 25 July 2008, googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.
14. “Python (Programming Language).” Wikipedia, Wikimedia Foundation, 28 May 2021, en.wikipedia.org/wiki/Python\_(programming\_language). [Accessed 20.09.2020]
15. “How Google Search Works | Search Algorithms.” Google, Google, www.google.com/intl/en\_uk/search/howsearchworks/algorithms/. [Accessed 10.10.2020]
16. “Bing Webmaster Guidelines.” Bing, Microsoft, www.bing.com/webmasters/help/webmaster-guidelines-30fba23a. [Accessed 10.10.2020]
17. “繁体字.” Wikipedia, Wikimedia Foundation, zh.wikipedia.org/wiki/%E7%B9%81%E4%BD%93%E5%AD%97#%E6%AD%B7%E5%8F%B2%E8%AA%9E%E8%A8%80%E5%AD%B8%E7%9A%84%E8%A7%92%E5%BA%A6. [Accessed 27.10.2020]
18. “简化字.” Wikipedia, Wikimedia Foundation, 11 May 2021, zh.wikipedia.org/wiki/%E7%AE%80%E5%8C%96%E5%AD%97. [Accessed 27.10.2020]
19. “中文「繁簡之爭」系列報導(中).” 老醫之家－華語健康資訊入口網站－台灣健康資訊網, olddoc.tmu.edu.tw/Taiwanese-3/Tai-go-0006.htm. [Accessed 10.03.2021]
20. “中文「繁簡之爭」系列報導(下).” 老醫之家－華語健康資訊入口網站－台灣健康資訊網, olddoc.tmu.edu.tw/Taiwanese-3/Tai-go-0007.htm. [Accessed 10.03.2021]
21. “Pdfplumber.” PyPI, 20 Oct. 2020, pypi.org/project/pdfplumber/0.5.24/. [Accessed 27.10.2020]
22. “Collections — Container Datatypes.” Collections - Container Datatypes - Python 3.9.5 Documentation, docs.python.org/3/library/collections.html#collections.Counter. [Accessed 27.05.2021]
23. CLI Team. “The 100 Most Common Chinese Characters: Chinese Language Institute.” CLI, 12 Sept. 2019, studycli.org/chinese-characters/the-100-most-common-chinese-characters/. [Accessed 10.03.2021]
24. hanziDB. Most Common Chinese Characters - Ordered by Frequency, hanzidb.org/character-list/by-frequency. [Accessed 10.03.2021]
25. “Dr. Jun Da: Faculty: Middle Tennessee State University.” Dr. Jun Da | Faculty | Middle Tennessee State University, www.mtsu.edu/faculty/jun-da. [Accessed 10.03.2021]
26. “Random - Generate Pseudo-Random Numbers.” Random - Generate Pseudo-Random Numbers - Python 3.9.5 Documentation, docs.python.org/3/library/random.html#random.shuffle. [Accessed 10.01.2021]
27. “Schedule.” PyPI, 10 Apr. 2021, pypi.org/project/schedule/. [Accessed 20.03.2021]
28. “Beautiful Soup Documentation.” Beautiful Soup Documentation - Beautiful Soup 4.9.0 Documentation, www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed 15.11.2020]
29. “Requests: HTTP for Humans™.” Requests 2.25.1 Documentation, docs.python-requests.org/en/master/. [Accessed 15.11.2020]
30. “Datetime - Basic Date and Time Types.” Datetime - Basic Date and Time Types - Python 3.9.5 Documentation, docs.python.org/3/library/datetime.html. [Accessed 15.11.2020]
31. “Urllib.request - Extensible Library for Opening URLs.” Urllib.request - Extensible Library for Opening URLs - Python 3.9.5 Documentation, docs.python.org/3/library/urllib.request.html#module-urllib.request. [Accessed 05.04.2021]
32. “Urllib.error - Exception Classes Raised by Urllib.request.” Urllib.error - Exception Classes Raised by Urllib.request - Python 3.9.5 Documentation, docs.python.org/3/library/urllib.error.html#module-urllib.error. [Accessed 05.04.2021]

## **Appendix A – correlation heatmap between character list and Chinese publications**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Correlation | Wikipedia | Jun Da | Golden Brother | Hong Kong Year Book | Chinese Teaching Material | Lost on a Red Minibus to Taipo | You are the apple of my eye | The Heaven Sword and Dragon Saber | Romance of the Three Kingdoms | Classic Chinese Teaching Material | Lost rainy season | Call To Arms |
| Wikipedia | - | 0.812717645 | 0.115145665 | 0.698328094 | 0.777341947 | 0.650740916 | 0.649222883 | 0.525578095 | 0.216848672 | 0.197394354 | 0.763709052 | 0.701199888 |
| Jun Da | 0.812717645 | - | 0.378605531 | 0.676976387 | 0.950210299 | 0.844666859 | 0.837810545 | 0.776900387 | 0.421479613 | 0.418548859 | 0.942727113 | 0.902653777 |
| Golden Brother | 0.115145665 | 0.378605531 | - | 0.202124644 | 0.385042592 | 0.608239201 | 0.514780018 | 0.419905519 | 0.340342018 | 0.233504116 | 0.372671256 | 0.374519432 |
| Hong Kong Year Book | 0.698328094 | 0.676976387 | 0.202124644 | - | 0.607915672 | 0.490877423 | 0.477224074 | 0.41376192 | 0.27651089 | 0.216970187 | 0.588713935 | 0.382640702 |
| Chinese Teaching Material | 0.777341947 | 0.950210299 | 0.385042592 | 0.607915672 | - | 0.869392236 | 0.879876823 | 0.796978725 | 0.416367789 | 0.428275344 | 0.966090672 | 0.942391534 |
| Lost on a Red Minibus to Taipo | 0.650740916 | 0.844666859 | 0.608239201 | 0.490877423 | 0.869392236 | - | 0.892985295 | 0.768083469 | 0.384570489 | 0.292525568 | 0.873072952 | 0.844629121 |
| Correlation | Wikipedia | Jun Da | Golden Brother | Hong Kong Year Book | Chinese Teaching Material | Lost on a Red Minibus to Taipo | You are the apple of my eye | The Heaven Sword and Dragon Saber | Romance of the Three Kingdoms | Classic Chinese Teaching Material | Lost rainy season | Call To Arms |
| You are the apple of my eye | 0.649222883 | 0.837810545 | 0.514780018 | 0.477224074 | 0.879876823 | 0.892985295 | - | 0.717497141 | 0.323305606 | 0.307491899 | 0.881298894 | 0.839784786 |
| The Heaven Sword and Dragon Saber | 0.525578095 | 0.776900387 | 0.419905519 | 0.41376192 | 0.796978725 | 0.768083469 | 0.717497141 | - | 0.557053493 | 0.435269404 | 0.768708638 | 0.829807993 |
| Romance of the Three Kingdoms | 0.216848672 | 0.421479613 | 0.340342018 | 0.27651089 | 0.416367789 | 0.384570489 | 0.323305606 | 0.557053493 | - | 0.661653508 | 0.398332636 | 0.414934632 |
| Classic Chinese Teaching Material | 0.197394354 | 0.418548859 | 0.233504116 | 0.216970187 | 0.428275344 | 0.292525568 | 0.307491899 | 0.435269404 | 0.661653508 | - | 0.376820784 | 0.382640702 |
| Lost rainy season | 0.763709052 | 0.942727113 | 0.372671256 | 0.588713935 | 0.966090672 | 0.873072952 | 0.881298894 | 0.768708638 | 0.398332636 | 0.376820784 | - | 0.933164214 |
| Call To Arms | 0.701199888 | 0.902653777 | 0.374519432 | 0.382640702 | 0.942391534 | 0.844629121 | 0.839784786 | 0.829807993 | 0.414934632 | 0.382640702 | 0.933164214 | - |

**Appendix B – Program code**

All program code is saved in OpenDocument Text format.

#python program for converting PDF file to test file

#pdfp.py



#python program for counting the word frequency

#word\_frequency\_count.py



#python program for random selecting characters

#Random\_Keywords.py



#python program for automation

#daily\_Run.py



#python program for requesting proxy

#Request\_Proxies.py



#python program for removing duplicate proxy

#Remove\_Proxies\_Duplicates.py



#python program for scrapping Google and Bing

#Web\_Scraping\_Google\_and\_Bing\_with\_Concat\_and\_sorted\_Random\_Proxy.py

#The program included the scrapping action and checking the validity of the proxy address before scrapping.



**Appendix C – Estimated result before modifying the characters.**

At the beginning of the project, the methodology of the 9-year study had been referenced for selecting the characters with 5 different frequency, which are highest frequency (characters ranking #1-#1791), high frequency (characters ranking #1792-#3582), middle frequency (characters ranking #3583-#5373), low frequency (characters ranking #5374-#7164) and lowest frequency (characters ranking #7165-#8955) and randomly select 100 characters is each frequency, total selected 500 characters.

After running the program and calculating the estimation for around 20 days (from 29th March 2021 to 16th April 2021), the estimation result shows that the estimation with the highest frequency and lowest frequency having the underestimation and overestimation problem. Another problem is that the estimated indexed web page size of Bing is much bigger than Google has, which is unreasonable. So, modification of character selection was required. Figure A1 shows the estimation result before modifying the characters.

Figure A1 The Google and Bing estimated indexed web page size from 29th March 2021.