UNIVERSITI MALAYSIA SARAWAK

UNIMAS

Faculty of Computer Science and Information Technology

# *Leveraging Web Scraping in Predictive Modelling of Supply Risk Detection*

Ng Shi Ya

Bachelor of Computer Science with Honours

(Information Systems)

2022

# LEVERAGING WEB SCRAPING IN PREDICTIVE MODELLING OF SUPPLY RISK DETECTION

NG SHI YA

This project is submitted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science with Honours (Information Systems)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2022

## UNIVERSITI MALAYSIA SARAWAK

### THESIS STATUS ENDORSEMENT FORM

**TITLE** LEVERAGING WEB SCRAPING IN PREDICTIVE MODELLING OF SUPPLY RISK DETECTION

**ACADEMIC SESSION:** 2022/2023

NG SHI YA

**(CAPITAL LETTERS)**

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [ or for the purpose of interlibrary loan between HLI ]
5. ** Please tick ( √ )

□ CONFIDENTIAL (Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)

□ RESTRICTED (Contains restricted information as dictated by the body or organization where the research was conducted)

☑ UNRESTRICTED

Validated by

(AUTHOR'S SIGNATURE) (SUPERVISOR'S SIGNATURE)

Permanent Address

B-24-5,
Seputeh Permai Condo,
Taman Seputeh,
58000 Kuala Lumpur.

Date: 24 July 2023 Date: 24 July 2023

Note * Thesis refers to PhD, Master, and Bachelor Degree

** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

# DECLARATION

I hereby declare that this project is based on my original work. I have not copied from any other student’s work or from any other sources except where due to reference or acknowledgment is not made explicitly in the text, nor has any part had been written for me by another person.

------------------------------

(NG SHI YA, 70723)

# ACKNOWLEDGEMENT

I wish to express my gratitude to my Final Year Project (FYP) supervisor, Prof. Dr. Jane Labadin for guiding me in preparing this project. The process throughout this project went more smoothly because of the ideas given by her when I was having troubles with the scope of the project. She was patient and gave me the support I needed to complete this project.

I would also like to express my appreciation to the course coordinator, Prof. Dr. Wang Yin Chai and examiner, Dr. Dayang Nurfatimah for the process guide in writing the documentation and for giving feedbacks which allowed for further improvement of my project.

Lastly, I would like to thank my classmates, friends, and family for their continuous support throughout the project.

# ABSTRACT

*Supply risk is caused by interruptions to the flow of product in a supply chain whether it is economic, environmental, political, or ethical. These temporary events may cause a decrease in a supply chain's performance in terms of inventory costs, production process, flexibility, and responsiveness. Despite these events faced by companies in the past few years, the consideration of the digital implementation in supply risk strategies is still not significant. This may be due to lack of budget and needing additional guidance to transition to more advanced technologies. For these purposes, a web scraping program is developed to identify the supply risks caused by these temporary events. It is a low-cost solution for the temporary events to be evaluated so that the risks can be detected in real-time for better decision-making. The important aspects for the temporary events are collected such as the date, description, title and link. The significance of the extracted output is evaluated with the topic modelling algorithm (LDA model algorithm) for the purpose of predictive supply risk. By collecting data from news sites, this system is aimed to provide data for predictive modelling so it can be used to detect patterns in supply risks to create and predict the probability of a temporary event occurring in the future.*

# ABSTRAK

Risiko pembekalan disebabkan oleh gangguan kepada aliran produk dalam rantaian pembekalan sama ada ekonomi, alam sekitar, politik atau etika. Peristiwa sementara ini boleh menyebabkan penurunan dalam prestasi rantaian pembekalan dari segi kos inventori, proses pengeluaran, fleksibiliti dan responsif. Walaupun peristiwa ini dihadapi oleh syarikat dalam beberapa tahun kebelakangan ini, pertimbangan pelaksanaan digital dalam strategi risiko pembekalan masih tidak digunakan secara meluas. Ini mungkin disebabkan oleh kekurangan belanjawan dan syarikat memerlukan panduan tambahan untuk beralih kepada teknologi yang lebih maju. Untuk tujuan ini, program web scraping dibangunkan untuk mengenal pasti risiko pembekalan yang disebabkan oleh peristiwa sementara ini. Ia adalah penyelesaian kos rendah untuk peristiwa sementara untuk dinilai supaya risiko dapat dikesan dalam masa nyata untuk membuat keputusan yang lebih baik. Aspek penting bagi acara sementara dikumpul seperti tarikh, penghuraian, tajuk dan pautan. Kepentingan data yang diekstrak dinilai dengan algoritma pemodelan topik (LDA model algorithm) untuk tujuan risiko bekalan ramalan. Dengan mengumpul data daripada laman berita, sistem ini bertujuan untuk menyediakan data untuk pemodelan ramalan supaya ia boleh dilakukan untuk mengesan corak dalam risiko pembekalan untuk mencipta dan meramalkan kebarangkalian peristiwa sementara yang akan berlaku pada masa hadapan.

## TABLE OF CONTENTS

**DECLARATION** ........ i

**ACKNOWLEDGEMENT** ........ ii

**ABSTRACT** ........ iii

**ABSTRAK** ........ iv

**LIST OF FIGURES** ........ viii

**LIST OF TABLES** ........ x

**CHAPTER 1 : INTRODUCTION** ........ 1

**1.1 Background** ........ 1

**1.2 Problem statement** ........ 2

**1.3 Scope** ........ 2

**1.4 Objectives** ........ 3

**1.5 Methodology** ........ 3

**1.6 Significance of project** ........ 5

**1.7 Expected outcome** ........ 6

**1.8 Project Schedule** ........ 6

**1.9 Project Outline** ........ 6

**CHAPTER 2 : LITERATURE REVIEW** ........ 8

**2.1 Introduction** ........ 8

**2.2 Review of Existing Data Collection Methods** ........ 8

**2.2.1 Primary Data Collection** ........ 8

**2.2.2 Secondary Data Collection** ........ 9

**2.2.3 Findings of the Review of Data Collection Methods** ........ 9

**2.3 Review of Web Scraping Techniques** ........ 10

**2.3.1 Manual Web Scraping** ........ 10

**2.3.2 Automated Web Scraping** ........ 10

**2.4 Comparison of Web Scraping Tools and Libraries** ........ 11

**2.4.1 OctoParse** ........ 11

**2.4.2 BeautifulSoup** ........ 12

**2.4.3 Scrapy** ........ 12

**2.4.4 Selenium** ........ 13

2.4.5 Comparison Between Web Scraping Tools and Libraries .................................................. 13
2.5 Predictive Modelling of Supply Risk Detection .......................................................... 14
2.6 Review of Predictive Modelling Techniques to Assess the Scraped Output for Predictive Supply Risk Detection .......................................................... 15
2.6.1 Name Entity Recognition (NER) .......................................................... 15
2.6.2 Text Classification (Naïve Bayes or SVM) .......................................................... 16
2.6.3 Topic Modelling (LDA) .......................................................... 16
2.7 Proposed System .......................................................... 17
2.8 Summary .......................................................... 17
CHAPTER 3 : REQUIREMENT ANALYSIS AND DESIGN .......................................................... 19
3.1 Introduction .......................................................... 19
3.2 Requirements Analysis .......................................................... 19
3.2.1 Hardware Requirements .......................................................... 21
3.2.2 Software Requirements .......................................................... 21
3.3 System Architecture Design .......................................................... 22
3.4 User Interface Design .......................................................... 25
3.5 Significance Testing through Topic Modelling with LDA Model Algorithm .......................... 28
3.6 Evaluation on the effectiveness of the developed system for predictive modelling of supply chain risks .......................................................... 29
3.7 Conclusion .......................................................... 29
CHAPTER 4 IMPLEMENTATION AND TESTING .......................................................... 30
4.1 Introduction .......................................................... 30
4.2 Development and Testing Tools .......................................................... 30
4.3 Web Scraping .......................................................... 31
4.3.1 Spider Implementation .......................................................... 31
4.4 User Interface .......................................................... 38
4.4.1 Main Menu .......................................................... 38
4.4.2 Web Scraper .......................................................... 39
4.4.3 Topic Modelling .......................................................... 40
4.4.4 Web Scraper Execution .......................................................... 42
4.5 Web Scraping Algorithm Output .......................................................... 46
4.6 Usability Testing .......................................................... 48
4.6.1 Correctness of Scraped Data .......................................................... 48

**4.6.2 Topic Modelling with Latent Dirichlet Allocation (LDA) algorithm** .................................49

**4.6.3 User Focused Testing**.............................................................................................55

**4.7 Testing Evaluation** ...................................................................................................55

**4.7.1 Coherence Score Measure for Topic Quality** ......................................................56

**4.7.2 Interpreting the Topics Modelled**..........................................................................58

**4.8 Summary**..................................................................................................................63

**CHAPTER 5 : CONCLUSION AND FUTURE WORK**......................................................64

**5.1 Introduction**..............................................................................................................64

**5.2 Contributions**............................................................................................................64

**5.3 Limitations**...............................................................................................................64

**5.4 Future Works** ...........................................................................................................65

**5.5 Conclusion** ...............................................................................................................66

**REFERENCES**...................................................................................................................67

**APPENDIX A** ....................................................................................................................69

**APPENDIX B** ....................................................................................................................70

## LIST OF FIGURES

Figure 3.1 Examples of temporary events that cause supply risk. ........ 20
Figure 3.2 Workflow overview of the project ........ 20
Figure 3.3 Structure of web scraping with Scrapy spiders ........ 23
Figure 3.4 System architecture of Scrapy (Scrapy, n.d.) ........ 24
Figure 3.5 Flowchart of proposed web scraping algorithm ........ 25
Figure 3.6 Proposed system Main Menu ........ 26
Figure 3.7 Web Scraping Algorithm Menu ........ 27
Figure 3.8 Topic Modelling Examples Menu ........ 28
Figure 3.9 Topic Modelling Workflow ........ 29
Figure 4.1 Import Libraries for web scraping. ........ 31
Figure 4.2 Initialize Scrapy Items and Spider ........ 32
Figure 4.3 Set scope output for spider ........ 33
Figure 4.4 Load and extract page content ........ 34
Figure 4.5 Extract article elements by css selector ........ 34
Figure 4.6 Logics to extract article title and time ........ 35
Figure 4.7 Conditions to stop scraping. ........ 35
Figure 4.8 Append and store the scraped article data ........ 36
Figure 4.9 Export sorted data to csv file ........ 36
Figure 4.10 Inspect elements action (Reuters, 2023) ........ 37
Figure 4.11 Example of CSS selector ........ 37
Figure 4.12 Example of Xpath selector ........ 38
Figure 4.13 Main Menu Window ........ 38
Figure 4.14 Web Scraper Window ........ 39
Figure 4.15 The temporary events category to scrape. ........ 40
Figure 4.16 The time scope of the output ........ 40
Figure 4.17 Topic Modelling Window ........ 41
Figure 4.18 Topic Model Example in HTML ........ 41
Figure 4.19 Load data ........ 42
Figure 4.20 Scraped output in tabular format. ........ 42
Figure 4.21 Import libraries for GUI and define user input variables ........ 43
Figure 4.22 Logics for ‘Web Scraper’’ GUI window ........ 44
Figure 4.23 Execute crawling function. ........ 44
Figure 4.24 Web Interface of Reuters (Natural Disaster Headlines) (Reuters, 2023) ........ 45
Figure 4.25 Web Interface of Container News (Port News) (Container News, 2023) ........ 46
Figure 4.26 Temporary event data (Natural Disaster) ........ 47
Figure 4.27 Temporary event data (Geopolitical Activity) ........ 47
Figure 4.28 Temporary event data (Port News) ........ 48
Figure 4.29 Example showing Conditional Formatting for Duplicated Data ........ 49
Figure 4.30 Import libraries for LDA topic modeling ........ 50
Figure 4.31 Code for tokenization, lemmatizing, and removing stop words ........ 50
Figure 4.32 Output after tokenization, lemmatizing, and removing stop words ........ 51

Figure 4.33 Create dictionary and corpus from tokens and train LDA topic model.................................51
Figure 4.34 Generate coherence scores using C_v coherence measure....................................................52
Figure 4.35 C_v output of coherence score over number of topics in ‘Natural Disaster’ for the recent 2 Years. .........................................................................................................................................52
Figure 4.36 Code to train the final model with the optimal number of topics. ..........................................53
Figure 4.37 Code to generate baseline coherence score. ............................................................................53
Figure 4.38 Topics printed for 'Natural Disaster’. ......................................................................................53
Figure 4.39 To display the topics using pyLDAvis in visualization. .........................................................53
Figure 4.40 Screenshot of the Interactive Topic Modelling Visualization (Natural Disaster Recent 2 Years). ........................................................................................................................................54
Figure 4.41 Topics modelled (Natural Disaster).........................................................................................58
Figure 4.42 Top 30 terms on lower Lambda value .....................................................................................59
Figure 4.43 Top 30 terms on higher Lambda value ....................................................................................60
Figure 4.44 Intertopic Distance Map with Topics Modelled ......................................................................61
Figure 4.45 Frequency of each topic...........................................................................................................62
Figure 4.46 Scatterplot for Topic Coordinates............................................................................................62
Figure 4.47 Topic Distribution over Time ..................................................................................................63

# LIST OF TABLES

Table 2.1 Comparison between web scraping tools and libraries ........ 14
Table 3.1 Hardware requirements ........ 21
Table 3.2 Software requirements ........ 22
Table 4.1 Number of topics chosen and coherence measures (C_v) of the temporary event categories for 'Recent 1 Year'. ........ 56
Table 4.2 Number of topics chosen and coherence measures (C-umass) and (C_v) of the temporary event categories for 'Recent 2 Years' ........ 57
Table 4.3 Baseline topic coherence score for each temporary event category based on the number of data input. ........ 57

# CHAPTER 1 : INTRODUCTION

## 1.1 Background

Supply risk is caused by interruptions to the flow of product in a supply chain whether it is economic, environmental, political, or ethical. These interruptions are temporary events which may cause a decrease in a supply chain's performance in terms of inventory costs, production process, flexibility, and responsiveness (Avelar-Sosa et al., 2014). According to Jabil, disruptive events such as Covid-19 has affected 78% of the supply chains surveyed, 46% for Geopolitical and Trade Instability and 35% for Natural Disasters (Jabil, 2022).

Despite the temporary events faced in the past few years, only less than 27% of company manufacturers have implemented digital transformation plans such as automation, artificial intelligence, and robotics to face supply chain challenges. The consideration of the digital implementation in supply risk strategies is still less than 37%. This may be due to lack of budget and needing additional guidance to transition to more advanced technologies (Jabil, 2022).

Meanwhile, data collection techniques such as web scraping help to identify supply risks caused by these temporary events. It is a low-cost solution for these temporary events to be evaluated so that the risks can be detected in real-time for better decision-making. By collecting data from websites, predictive modelling can be done to detect patterns in supply risks to create and predict the probability of a disruptive event occurring in the future (Fan et al., 2015).

### 1.2 Problem statement

Along with the rise in supply demands, the probability where supply risk occurring due to external factors such as economic, environmental, political, ethical factors and more is significant. According to Jabil, disruptive events such as Covid-19 have affected 78% of the supply chains surveyed, 46% for Geopolitical and Trade Instability and 35% for Natural Disasters (Jabil, 2022).

Supply disruption, price fluctuations and incidents from these factors create quality problems, accountability, and reputational issues for businesses. The increase in supply risk faced by these companies is the reason why frequent update of important information that affects their supply flow is needed. Meanwhile, the supply chains' decision-makers that rated their predictive supply chain "excellent" facing Covid-19 are of only 17% (Ng, n.d.). This is because most companies are not ready to have supply risk management as a long-term implementation, due to budget and lack of support to incorporate advanced technologies.

Features like automated web data scraping help to identify risks caused by temporary events and in turn enhances supply operations. The aim of this project is to provide a low-cost solution for these disruptive events to be evaluated so that the risks can be detected early for companies to potentially make better decisions.

### 1.3 Scope

This project aims to propose a suitable web scraping program to collect data on temporary events and categorize the data for supply risk that could interrupt a company's supply management. The temporary events include natural disasters, dangerous incidents, geopolitical activity, port disruption and more.

Research will be done to choose a web scraping tool suitable to create a web scraping program for this project. Then, the scraped data from the web scraping program output will be tested to see the significance of use in predictive modelling for supply risk detection. By the end of the project, the web scraping program would be done to ensure the related temporary events can be extracted.

The main data source for this project would be reliable sources such as news and media sites. The program would have a simple interface and the selected data scraping attributes are for the use of cataloguing the temporary events in corresponding to running data analytics for the use of predictive supply risk.

### 1.4 Objectives

1. Develop a robust web scraping program capable of efficiently collecting data from selected news sites to identify temporary events that may cause supply chain risk.
2. To apply the Latent Dirichlet Allocation (LDA) model algorithm to evaluate the significance of the extracted data.
3. To evaluate the effectiveness of the developed system for predictive modelling of supply chain risks.

### 1.5 Methodology

Objective 1 involves in designing and implementing a scalable and automated system that can extract relevant information specifically date, description, title, and link. The methodology used for this project is agile software development because the project size is not considered large and the focus is to constantly optimize, predict possible errors and do testing to meet the specified needs of the project. The benefit of this method is the flexibility to change the system requirements as they change throughout the development process. The processes are

Requirements Analysis, Design, Coding and Testing. Next, objective 2 entails in implementing the LDA algorithm, conducting topic modelling analysis, and assessing the significance of the scraping output. Lastly, objective 3 involves evaluating the effectiveness of the developed system for predictive modelling of supply chain risks. For this objective, a user focused review is done to evaluate the developed system. Both objective 2 and 3 are involved during the Testing process.

### i. Set requirements and analysis

Research is done on the system requirements needed, strength and weaknesses of the web scraping tool (Scrapy) chosen for the project and what can be improved from it to meet the project goals. After the information and resources needed for the project are determined, analysis is done on what needs to be categorized for the attributes of the temporary events that are to be extracted from the web for the project, the system architecture, and usability testing method.

### ii. Design

In this phase, the design requirements needed are decided, such as the programming language, tools and libraries that will be needed to support the web scraping activities. The design scope such as the main features of the web scraping algorithm are specified, and system requirements are added depending on the additional functionalities needed to assess the developed system.

### iii. Coding

After the technical and design requirements are set, the tools and libraries installed will be used to conduct the coding. The web scraping algorithm will be developed in units and tested to ensure it works according to the design planned.

### iv. Testing

After the coding phase is completed, further review and testing are done on the algorithm before the deploy to ensure it extracts the required web data for the project and that there are no unexpected errors. The testing will be done on the correctness of the data extracted, to ensure that there are no missing data, or duplicate entry. The testing method will be manual testing on smaller samples among the large amounts of web data extracted on the temporary events that cause supply risk. Then, the assessment of the scraped output from the developed web scraping system is done by using the LDA model algorithm to conduct topic modelling analysis and test the significance of the scraped output. Lastly, the effectiveness of the developed system is evaluated objectively through a focused review by a user with the related expertise.

## 1.6 Significance of project

With the increase of business opportunities along with the demand for global competition, companies are more dependent on their inbound and outbound logistics, sourcing large volumes of supplies for profit improvement. The rising supply demands increases risk and vulnerabilities in supply management and prevention measures are needed to prepare businesses for these potential risks. Therefore, this project aims to potentially apply data analytics on detecting supply risk, that is, to study the data extraction step of data analytics by collecting web data of temporary events that might cause supply risk.

This project is expected to contribute for data analysts seeking to analyse data on supply risk management and needs the related categorised information for it. With this, they can easily acquire the data they need from the web without going through extra steps of requesting personally or formally from organizations. The efficiency of the web scraping process saves time and resources and may indirectly benefit businesses eventually through the web scraping algorithm aimed towards predictive supply risk detection.

## 1.7 Expected outcome

The web scraping system is robust and developed to efficiently collect data on the temporary events that cause supply risk from selected news sites. The data extracted is categorized for easier analysis and tested to be significant for the use in predictive supply risk detection. The developed system is effective for predictive modelling of supply chain risk.

## 1.8 Project Schedule

The project schedule of FYP 1 and FYP 2 is shown in Gantt chart in Appendix A.

## 1.9 Project Outline

In Chapter 1: Introduction, the overview of the general details of the project is listed. The details included are the background, problem statement, project scope, objectives, methodology, significance of the project, and expected outcome.

In Chapter 2: Literature Review, research is done on previously published works on the project topic. The research is done to analyze the most suitable project solution and determine the proposed system requirements.

In Chapter 3: Requirement Analysis and Design, the requirements determined in the previous chapter will be further discussed in this chapter. Analysis of the requirements is done to design the development of the proposed system.

In Chapter 4: Implementation and Testing, the execution, software, selected tools, and testing methods are discussed for the development of the proposed system.

In Chapter 5: Conclusion and Future Works, discussion of the challenges faced and result findings during the project development is done and what is expected for future improvements.

# CHAPTER 2 : LITERATURE REVIEW

## 2.1 Introduction

Data collection is the most essential process in predictive modeling for risk prediction in supply chain management. It is the process to collect information that is reliable to analyze, create and build more persuasive and effective analytics or to find solutions to research problems. There are multiple ways to collect data, so the data collection method and the web scraping techniques are studied. In the case of this project, where the most recent online data is required, the existing web scraping tools will be compared and studied by their nature, strengths, and weaknesses to select the most suitable data collecting strategy for the project. For the evaluation of the scraped output, a suitable predictive modeling technique is needed to assess the significance of the data for predictive supply risk detection. Hence, predictive modelling techniques that are suitable to assess web scraped news data are selected and compared to choose the most suitable one for the project.

## 2.2 Review of Existing Data Collection Methods

The two main data collection methods are primary data collection methods and secondary data collection. The main goal of data collection is to collect information from available resources for research to solve problems or make predictions for future trends and evaluate possible outcomes. In this project, data collection is needed to make informed supply risk decisions for supply chains to ensure quality and integrity is retained for their businesses.

### 2.2.1 Primary Data Collection

Primary data is the raw or unstructured data collected directly from the source. The sources can be surveys, interviews, focus groups, and observations. The strength of this data collection method is that it is more accurate compared to secondary data and more detailed information that considers actual real-life situations can be obtained. The data is also not subjected to personal bias, the user has more control over the information gathered and the data is usually more recent. Meanwhile, the weakness of this method Is that it is more expensive, time consuming and has higher complexity in collecting (Valcheva, 2021).

### 2.2.2 Secondary Data Collection

Secondary data is data that has been gathered for another purpose but is relevant for use. It is usually easily obtained for the public and provides some type of analysis to support specific research. Some of the examples are government reports, official statistics, previous research, historical data, Google Analytics and Web information. The strength of secondary data collection is that is very affordable, time spent in collecting is very little, and helps create new findings and patterns from primary data. On the other hand, its weaknesses are that the data might not be authentic, further verification is needed to ensure its reliability from other sources and the data might be outdated (Formplus Blog, 2022). With this, although the data is more readily available, there is a main problem of the data being snapshot for certain points in time and the information gathered is very much subject to interpretation which is present with a certain bias (Ellram & Tate, 2016).

### 2.2.3 Findings of the Review of Data Collection Methods

In the review of data collection methods, primary data collection method is not as efficient for use in this project that requires large amount of data to be extracted. This is because the supply risk information that needs to be collected for supply chain management in ideal needs a faster and organized option over time for frequent updates of the recent news. To deal with this problem of primary data collection, secondary data collection would be a more practical method in gathering data for this project. However, there is a certain limit to secondary data collection in terms of reliability in the time the data is collected and its authenticity. To address the shortcomings of this method, collecting data through official news and alert sites could reduce the unreliability of the data collected. Hence, web scraping is needed to collect large amounts of data efficiently from news websites.

### 2.3 Review of Web Scraping Techniques

Web scraping is the extraction of data from websites, it can be manually done by a user or through automated process using a web crawler or bot. The data extracted is normally imported into files or spreadsheets. The main goal of web scraping is to transform specific data from non-tabular form into a structured and usable format, such as a database or a spreadsheet for archival of data and to track its changes.

#### 2.3.1 Manual Web Scraping

Web scraping through manually copy-and-pasting data from websites can be done with a spreadsheet to keep track of the extracted data. The advantage of this technique is that it is cheaper in terms of resources and spendings, there is no need to acquire new sets of skills to perform the scraping and there are less restrictions on the websites we can access to collect the data from. However, if a human operator must be hired for the manual task, then it will be more expensive in the long run. On the other hand, the fact that there are human errors and that it is the slowest method is a constant downside for this technique. With this project, the human operator will have to manually collect large amounts of the related data for a more accurate data analysis for the predictive supply risk model.

#### 2.3.2 Automated Web Scraping

Automated web scraping typically involves the use of software or scripts to retrieve unstructured data from the web and save it as structured data for further analysis. After identifying the website to scrape, web scraping tools or libraries are chosen to run the scraping process. For this technique, it is essential to inspect the HTML structure of the website to identify the specific HTML elements such as the tags, classes and id of the data that is to be extracted. This is for the web scraping tool or code to locate the specific HTML element and extract the data. Depending on the website layout, the automated scraping would have to be able to handle the dynamic content of the website, anti-scraping measures or paginations. Besides that, there is a need to monitor and maintain the web scraping code if the website were to change their structure or implement updates. With this, automated web scraping would be the more suitable choice to meet the project's needs.

## 2.4 Comparison of Web Scraping Tools and Libraries

There are several web scraping tools and Python libraries for web scraping that simplifies the web scraping process. In this section, the existing web scraping tools and libraries will be compared to evaluate the suitable solution for this project.

### 2.4.1 OctoParse

An automated web scraping tool that cuts down time and costs would be an existing web scraping software like OctoParse which is one of the best web scraping tools on the market. The strength of this tool is that it is easy to learn, has a simple interface, and can fulfill the needs of the shortcomings of the manual copy-and-paste method. With no coding needed, inexperienced users would be able to easily run and extract data from websites by using Octoparse. It has a fast-scraping speed due to its ability to schedule scraping tasks and the data can be saved in cloud platform (Octoparse, 2022). On the other hand, the weakness of this tool is that it can be very pricy according to the features and plans offered to meet the scalability of the project. The free Octoparse plan has limited functions where it does not include the task scheduling feature and only allows up to 2 concurrent local runs with 10,000 records per export (Octoparse, n.d.). To meet the project's requirements, the professional plan would be the minimum bar of the subscription to run the project efficiently which is $249 monthly. Adding