CHAPTER 3

METHODOLOGY AND REQUIREMENTS ANALYSIS

In this chapter, there are five (5) major sections explaining the methodology used and requirements analysis of a project. They are Section 3.1 Introduction, Section 3.2 Research Framework, Section 3.3 Timeline/Milestone, Section 3.4 Data Source, 3.5 Analysis of selected tool with any other relevant tools and 3.6 Chapter Summary. In fact, sufficient preparation must be done and suitable methodology must be identified in order to produce quality works within a short period of time.

3.1 Introduction

Research methodology refers to the practical "how" of any given piece of research. To be more specific, it is all about how a researcher conducts a study more systematically to ensure valid and reliable results that achieve the research goals and objectives. Besides, a research framework should also be developed as it provides an underlying structure to support our collective research efforts (Godfrey, 2019). According to Rajasekar et. al. (2006), research is a logical and systematic search for new and useful information on a particular topic. It is an investigation of finding solutions to scientific and social problems through objective and systematic analysis. Therefore, different methodologies should be explored in order to produce quality research.

3.2 Research Framework

Figure 3.1 is an illustration of the research framework for this project.

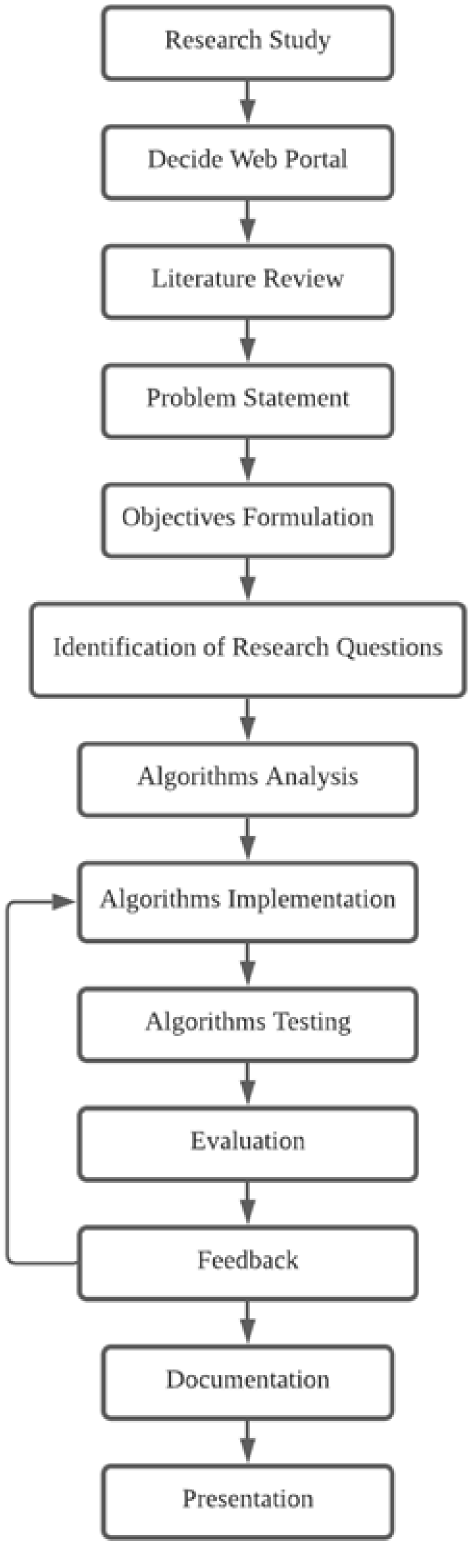

Figure 3.1: Research Framework

In this research, web structure mining and web content mining approaches are discussed. Therefore, a comprehensive study of web mining is carried out in which several relevant articles are reviewed in order to gather enough knowledge to carry out this research. Besides, web portal is decided for conducting this research and further implementation. The problem statement is determined and objectives are formulated based on the chosen journal, research questions are identified as well. Furthermore, a comprehensive analysis of chosen algorithms remains critical and need to have better understanding of each step. After that, the implementation of algorithms is carried out one at a time, once the first algorithm has finished, the second algorithm will then begin. The implemented algorithms must undergo testing phrase to ensure that the vulnerability is low and acceptable. Subsequently, the result of the algorithms is evaluated to see whether the result provides significant information regarding of this research. Receive feedback from supervisor and moderator. If any errors or enhancements need to be made, the process must be return to the algorithm implementation phase. Once both algorithms have been fully implemented, the documentation can be started and presentation can be performed.

3.3 Timeline/Milestone

Milestone	Milestone	Deadline
Research Study	Conduct feasibility studies and literature review in the field of web mining. Two journals have been chosen and approved by the supervisor.	28/05/2021
Implementation of Selenium Web Crawler	Study on Selenium Python and implement a web crawler for a chosen web forum are approved by the supervisor	09/06/2021
Introduction	Introduction to project background based on the chosen algorithms are approved by the supervisor.	11/06/2021
Methodology	The methodology and requirements analysis are completed, corrected and approved by the supervisor.	23/07/2021

Implementation of Journal 1 (coding and testing)	Implementation of the first journal based on the algorithm proposed is approved by the supervisor.	23/07/2021
Source Code Walkthrough	Walkthrough the sourcecode with supervisor for him to review.	3/09/2021
Documentation of Journal 1	Prepare the documentation report for the implementation	19/09/2021
Final System Testing	Conduct final algorithms testing with supervisor and moderator.	18/11/2021
Project Submission	Submission of Final FYP Report and all associated deliverables.	5/12/2021

Table 3.1: Research Timeline

3.4 Data Sources

This project is based on multiple data sources on different web forums. The example of web forums are as follow:

Web Forums
SixCrazyMinutes
UK of Equestria
GardenStew

Table 3.2: Data Sources

3.5 Analysis of selected tool with any other relevant tools

Tools comparison	Jupyter Notebook	PyCharm	Visual Studio Code
Type of license and open source license	Modified BSD license	Community edition: Apache License 2.0 Professional edition: Trialware	MIT License and Proprietary Software
Year founded	2014	2010	2015
Founding company	Fernando Perez and Brian Granger	JetBrains	Microsoft
License Pricing	100% open source software, free of charge	Free of charge (Community version) USD 199.00 (Professional version)	Free of charge (Visual Studio Code) $45-$250 (Visual Studio Professional)
Supported features	It supports over 40 programming languages including Python, R, Julia and Scala. It allows the display of the result of computation using rich media representations. In-browser editing for code, with automatic syntax highlighting, indentation, and tab completion/introspection.	Intelligent coding assistance Web development Build-in developer tools Scientific tools Customisable and cross-platform IDE The terminal is also present in the bottom-left corner which can be useful in installing new packages for python	Allow the users to choose from thousands of extensions to customize the IDE It supports hundreds of programming languages such as Python, C++ and PHP Extension Package Easy integration with git It comes with built-in support for Web applications
Common applications	● Data Science ● Machine Learning ● Natural Language Processing ● Image Processing	● Data Science ● Machine Learning ● Natural Language Processing ● Image Processing ● Web mining	It is a streamlined code editor with support for development operations like debugging, task running, and version control

Customer support	The customer support is given through online community	The customer support is given through online community	The customer support is given through online community
Limitations	● Less secure ● Run cell out of order ● Lack of IDE integration, linting and code-style correction ● Slow performance if the processor is poor	It takes more space than other text editors which degrade the functionality of code. The community version is idle for python development only and does not support programming languages. It can be complicated for beginners to set a virtual environment variable and thus they might not prefer to use it.	Some workflows are not very intuitive Some extensions show internal conflict due to accessing the same part of API

Table 3.3: Comparison of Tools

PyCharm has been selected as the tool to complete this project. It is because it is free of charge and easy to use. Besides, it can help in the analysis of syntax errors even before you compile your code to reduce any overhead. Furthermore, the PyCharm tool has the import assistance where the missing librarian from another part of the project will be detected and imported automatically. Lastly, the PyCharm comes with many plug-ins that can be used to enhance the quality of the project.

3.6 Chapter Summary

In conclusion, the research framework, timeline, data source, and analysis of tools are covered in this chapter. In fact, a good research methodology and framework must be identified before the development stage so that the program can be developed accordingly. Besides, it is also very crucial as the research framework determines how far the project goes. Thus, different methodologies and resources must be explored in order to produce a quality web crawler.