CHAPTER III
RESEARCH METHODS

3.1 Research Design

Research design that is used to developed web scraping program that scrap job vacancy from different websites using web services that use Java programming language and Jsoup to parse the scraped HTML is applied research where the research can be used directly. The process of web scraping will be explained by the flowchart shown in Fig. 3.1.

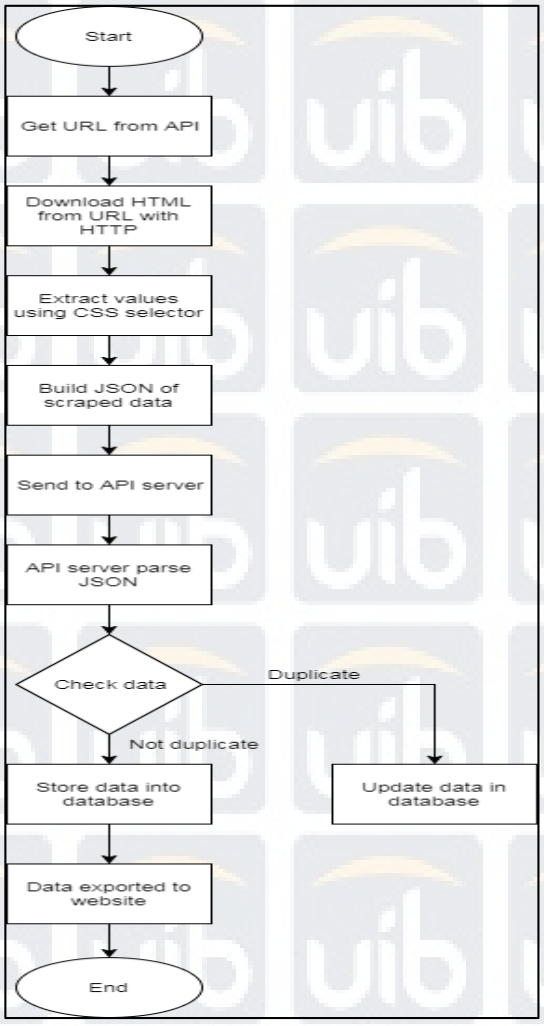

Figure 3.1 *Flowchart of Web Scraping Process*

Web scraping process start from getting URL from API and then download the HTML from the retrieved URL using HTTP. After getting the HTML, the values in the HTML will be extracted using CSS selector. Then JSON will be built from the extracted values and will be sent to API server to be process. If there are duplicates data, it will update the existing data and if there aren't any duplicates data, it will store the data as a new data and export it to the website.

3.2 Web Scrap Method

3.2.1 Rotating Proxy

Rotating proxy is a method often used in web scraping. The main purpose of using this method is to prevent real IP blocked by a website doing many concurrent requests. Rotating proxy usually is a pool of proxy that is being switched after request to emulate request being done from different location by different device. Rotating proxy can be seen in Fig. 3.2.

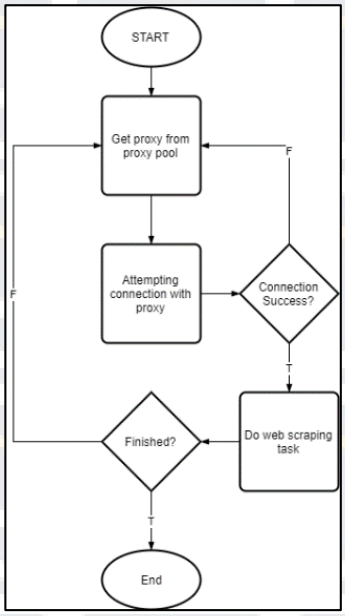

Figure 3.2 *Rotating Proxy Process*

Application will first get a proxy from the proxy pool and try to connect to website with the proxy. If the connection fails, the proxy will attempt to get another proxy and retry the connection. If the connection success, the scraper will do the parsing job and if there are another task, it will repeat the above steps again.

3.2.2 DOM Parsing

DOM parsing is often used when scraping web pages. The purpose of using this method is because reading HTML DOM is hard to do without making a mistake, and by parsing the DOM, we can convert the DOM into DOM tree object. DOM tree object value can be retrieved by using method provided by DOM parser library. In Fig. 3.3 on the left side we can see that HTML from website is often not well formatted. JSOUP will format the bad formatted DOM into structured format as shown in Fig. 3.3 on the right side and application can use CSS selector to retrieve wanted data from the java object.

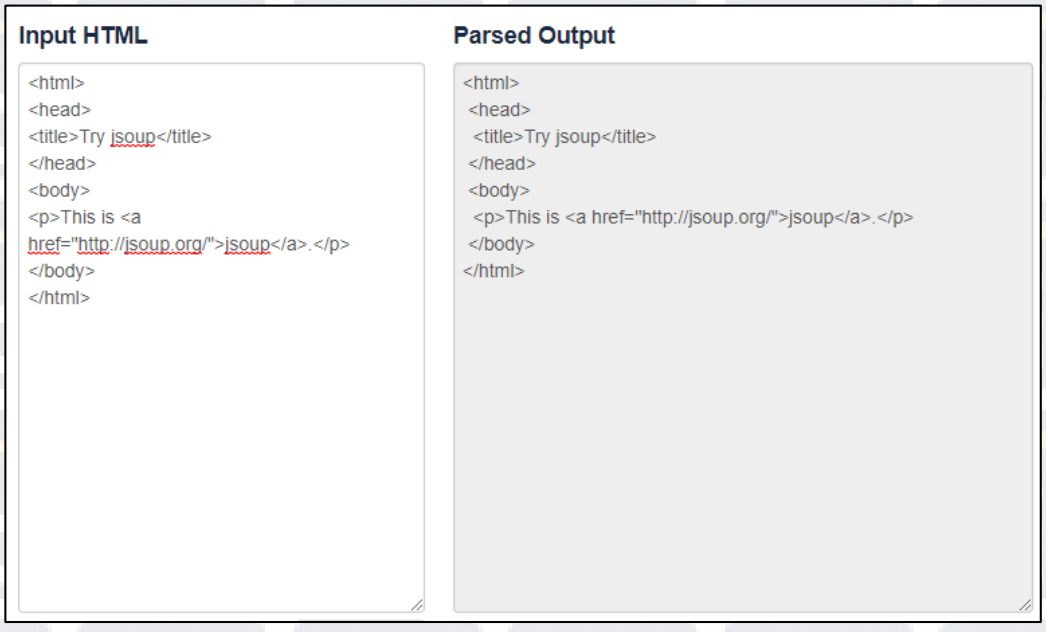

Figure 3.3 *Well Formatted and Bad Formatted HTML*

3.2.3 Web Service

Web service is a software system designed to support interoperable machine interaction over network. Main purpose of using web service is to import scraped data into job search application database. Because scraper and the job search application is a different application which is explained in the Fig. 3.4 as shown below.

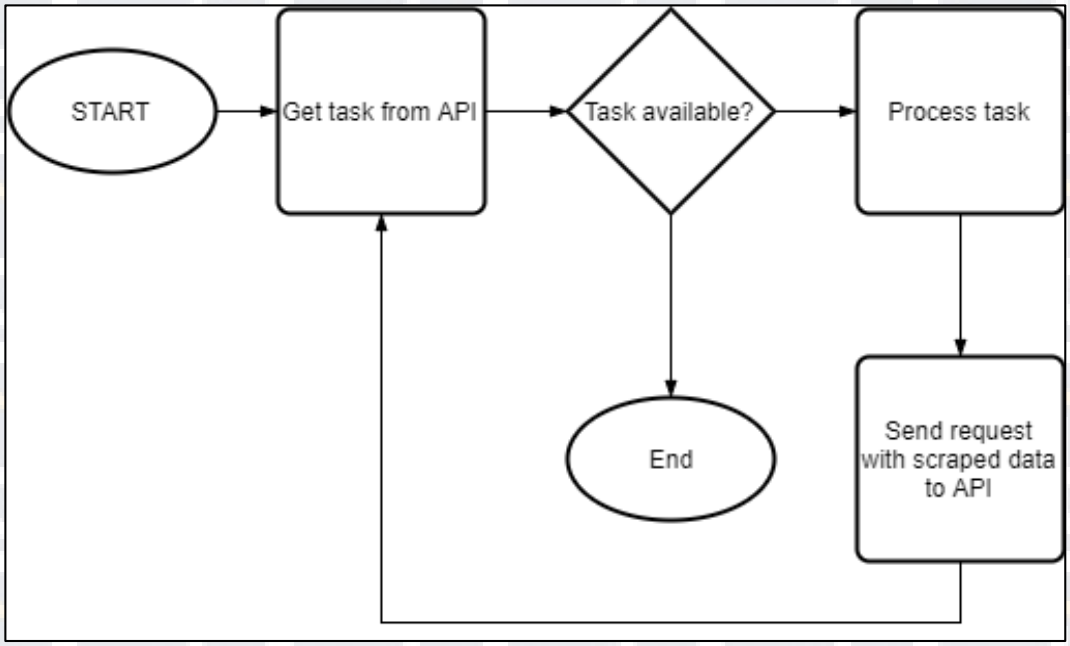

Figure 3.4 *Web Service Process*

The scraper will get task from API server when started. This task will consist of what URL scraper will work on and what kind of data will scraper get from the URL. If API server responded with a task, then the scraper will start to work on the task and send a POST request with scraped data as body to API server. API server will process the data and respond with another task is available.

3.3 Required Tools

In order to develop the scraping application, there are several tools that are needed to develop which are Okhttp3, JSoup, Maven, Laravel, Visual Studio Code. Okhttp3 is an efficient HTTP client for java applications, it is able to recover from common connection problems and connection failure, if a service has multiple IP addresses, it can retry the request to alternate addresses which is what we needed for our rotating proxy process.

JSoup is an open source java library which is mainly used for extracting data from HTML. It loads the page HTML and builds the corresponding DOM tree. This tree works the same way as the DOM in a browser. Jsoup also guarantees the parsing of any HTML, from the most invalid to the totally valid ones.

Maven is a popular build tool for Java which gives a default build process which can be customized where necessary. This has the advantage that the build configuration (POM) is very simple for ordinary builds. Laravel is also popular tool in writing PHP, and we will use Laravel to write our application API. Laravel is also known quick, functional, clean and simple routing. It has effective Object Relational Mapping (ORM) and database layer, it is also easy to integrate third party libraries.

Lastly which is Visual Studio Code, it has the simplicity of a source code editor with powerful developer tooling like intelliSense code completion and debugging. It also supports hundreds of programming languages.