INTISARI

Internet dapat menjadi sumber data *public* yang tersedia di berbagai *website*. Proses pengambilan data dari sebuah *website* memerlukan teknik tertentu karena data-data yang tedapat pada *website* merupakan data yang tidak terstruktur. Teknik pengambilan atau ekstraksi data dikenal dengan proses *scraping*. Sebuah *website* juga mempunyai banyak halaman-halaman web yang saling terhubung sehingga juga diperlukan teknik untuk dapat memeriksa seluruh halaman web dimana data akan diambil. Teknik untuk mengakses halaman web yang terhubung disebut dengan *crawling*. Dalam proses pengolahan data hasil ekstraksi dibutuhkan data yang terstruktur, oleh karena itu dibutuhkan sistem *scraping* dan *crawling* yang dapat menghasilkan data yang terstruktur dari sebuah *website*. Pada tugas akhir ini dipaparkan tentang teknik *scraping* dan *crawling* untuk mengekstrasi data dari sebuah *website*. Data yang ekstrasi adalah data *review* hotel pada *website* Traveloka.

Penggunaan *javascript* dan *ajax* pada sebuah *website* membuat akses data pada sebuah *website* tidak memerlukan *refresh* keseluruhan halaman web. Data pada *website* dapat ditampilkan dengan lebih interaktif. Untuk melakukan *crawling* pada *website* yang menggukan *javascript* dan *ajax* diperlukan teknik tertentu sehingga sistem *crawling* dapat berinteraksi dengan *ajax* dan proses *scraping* dapat mengambil semua data yang ada pada sebuah halamam web. Teknik *scraping* dan *crawling* yang dikembangkan menggunakan dan mengintegrasikan berbagai teknologi yang ada. *Scrapy* yang merupakan sebuah *framework scraping* dan *crawling* menjadi pilihan dalam pengembangan teknik ini. *Selenium* dan *chrome driver* digunakan untuk dapat berinterasi dengan web berbasis *ajax*. *Elasticsearh* digunkan sebagai tempat penyimpanan data hasil *scarping* melalui proses *pipeline item.*

Pengembangan teknik *scraping* dan *scrawling* dilakukan melalui beberapa tahapan. Tahap dimulai dari evaluasi *website* yang akan menjadi sumber data untuk mendapatkan elemen-elemen dimana data berada. Pemilihan elemen dilakukan dengan mengguakan *xpath selector*. *Xpath* digunakan dalam proses *scraling* dan *crawling* yang dikembangkan dalam *spider* pada *framework Scrapy*. Semua teknik ini dikembangkan mengunakan bahasa pemrograman *Python*. Hasil dari pengembangan teknik ini adalah sebuah sistem *scraping* dan *crawling* untuk mengekstrasi data *review* hotel dari web Traveloka. Sistem dapat berjalan dengan stabil mengambil jutaan *review* hotel yang ada. Data-data *review* juga dapat disimpan dan ditampilkan dengan baik pada *elasticsearch*.

Keyword: *scraping, crawling, scrapy, selenium, ajax, xpath,* traveloka

ABSTRACT

The internet can be a source of public data available on various websites. The process of retrieving data from a website requires certain techniques because the data found on the website is unstructured data. Data retrieval or extraction techniques are known as scraping processes. A website also has many web pages that are interconnected so that techniques are also needed to be able to check all web pages where data will be taken. The technique for accessing linked web pages is called crawling. In the process of processing data from extraction, structured data is needed, therefore we need a scraping and crawling system that can produce structured data from a website. In this final project, it is explained about scraping and crawling techniques for extracting data from a website. Extracted data is hotel review data from the traveloka website.

The use of javascript and ajax on a website makes accessing data on a website does not require refesh the entire web page. Data on the website can be displayed more interactively. To perform crawling on websites that use javascript and ajax, certain techniques are needed so that the scrawling system can interact with ajax and the scraping process can retrieve all the data on a web page. Scraping and crawling techniques are developed using and integrating various existing technologies. Scrapy which is a scraping and scrawling framework is an option in developing this technique. Selenium and chrome drivers are used to interact with ajax-based web. Elasticsearch are used as a place to store data from scarping through the item pipeline process.

The development of scraping and scrawling techniques is carried out through several stages. The stage starts from evaluating the website that will be the source of the data to get the elements where the data is. The element selection is done by using the xpath selector. Xpath is used in scraling and crawling processes that are developed in spider in Scrapy framework. All of these techniques were developed using the Python programming language. The result of developing this technique is a scraping and crawling system to extract hotel review data from the traveloka web. The system can run steadily taking millions of hotel reviews. Data review data can also be stored and displayed properly in elasticsearch.

Keyword: scraping, crawling, scrapy, selenium, ajax, xpath, traveloka