Seminar Nasional Hasil Penelitian 2017, ISBN : 978-602-14917-5-1

Pemanfaatan *News Crawling* Untuk Pembangunan *Corpus* Berita Menggunakan *Scrapy* dan *Xpath*

Taufiq Rizaldi[#1], Hermawan ArifPutranto[#2]

[#] *Jurusan Teknologi Informasi, Program Studi Manajemen Informatika, Politeknik Negeri Jember*

[1]taufiq_r@polije.ac.id[1]

[2]hermawan_arief_putranto@yahoo.com[2]

Abstract

Linguistically, language corpus is a collection of written (textual) or test hypotheses about language structure. However, the existence of the language corpus, especially the Indonesian corpus today is still very less. It's caused by the use of language corpus for Natural Language Processing is rare and most of them still using the same corpus which is used by previous research. In addition, the construction of the corpus itself takes a long time and big costs. To overcome this problem, this research proposed a development of language corpus, especially Indonesian corpus, using web crawling engine Scrapy and guided X-path. So with the use of guided web crawling technology is expected to build a corpus language data in accordance with the needs of research and net of unexpected codes and links without much time and energy consuming. The result shows that the development of news corpus using Scrapy and Xpath is successfully meet the expected target. This is characterized by the resulting corpus news that has been divided into three categories of news namely, entertainment, community and culinary news. In addition, from the parameters tested it can be concluded that the use of resources on the server computer is directly proportional to the number of items obtained and the file size. This means that the more items obtained and successfully stored the greater the size of the file and resource memory used. Thus, to limit memory usage on server computers, we can limit what items will be taken at the time of the scraping process by limiting the number of links crawled by the spider or limiting the number of items to be searched.

Keywords— **Language Corpus, Natural Language Processing, Scrapy, Web Crawling, XPath**

I. PENDAHULUAN

Secara linguistik, korpus bahasa merupakan kumpulan ujaran yang tertulis (tekstual) atau lisan yang digunakan untuk menyokong atau menguji hipotesis tentang struktur bahasa (Kamus Besar Bahasa Indonesia). Akan tetapi, dengan berkembangnya teknologi dalam bidang ilmu Natural Language Processing (Pemrosesan Bahasa Alamiah), keberadaan korpus bahasa tersebut bukan hanya digunakan untuk menguji struktur bahasanya saja, melainkan termasuk analisis sentimen dan semantik yang terkandung dalam susunan bahasa tersebut. Natural Language Processing (NLP) adalah proses pembuatan model komputasi dari bahasa sehingga memungkinkan terjadinya interaksi antara manusia dan komputer dengan perantaraan bahasa alami yang dipakai oleh manusia. NLP memodelkan pengetahuan terhadap bahasa, baik dari segi kata, bagaimana kata-kata bergabung menjadi suatu kalimat dan konteks kata dalam kalimat (Tiur, 2011). Sehingga keberadaan korpus Bahasa ini bukan hanya bermanfaat bagi pengembangan ilmu Bahasa saja, melainkan juga semua bidang ilmu yang mendukung perkembangan bidang ilmu Natural Language Processing.

Akan tetapi, keberadaan korpus Bahasa, khususnya korpus Bahasa Indonesia saat ini masih sangat kurang. Hal ini disebabkan karena penggunaan korpus Bahasa untuk keperluan Natural Language Processing masih jarang dan kebanyakan masih menggunakan korpus yang sama dengan yang digunakan oleh penelitian sebelumnya. Selain itu, pembangunan korpus itu sendiri memakan waktu yang lama dan biaya yang tidak sedikit. Padahal, saat ini ada sebuah media yang memiliki jumlah data yang berlimpah dan beragam bentuknya. Media tersebut adalah internet.

Saat ini keberadaan data, khususnya data tekstual, di internet tak terhingga jumlahnya. Hal ini disebabkan banyak aplikasi berbasis website yang memudahkan pengguna menyalurkan informasi dan ide dalam bentuk teks. Mulai dari situs portal berita yang menyampaikan informsai dan berita, blog dan forum yang khusus membahas topic tertentu, sampai aplikasi microblogging dan media social yang membicarakan topic yang beragam. Data dari aplikasi berbasis web tersebut dapat digunakan sebagai sumber dari pembangunan korpus Bahasa.

Namun, dengan jumlah data dan banyaknya aplikasi yang digunakan juga membawa permasalahan tersendiri bagi orang yang ingin menggunakan data aplikasi tersebut

sebagai korpus datanya. Pertama, tidak semua aplikasi menggunakan struktur halaman yang sama untuk memuat datanya, sehingga masing-masing aplikasi memiliki kode berbeda yang menandai bagian halaman yang berisi data yang akan diambil. Kedua, tidak semua aplikasi memiliki jumlah halaman dan tautan yang sama dalam memuat datanya. Hal ini menyebabkam waktu yang digunkaan untuk mengambil data dari aplikasi yang satu dengan yang lainnya menjadi berbeda. Ketiga, walaupun data yang ingin dikumpulkan sudah berada dalam satu media, masih dibutuhkan banyak waktu dan tenaga untuk menghasilkan data tekstual bila dilakukan secara manual.

Untuk mengatasi permasalahan tersebut, pada penelitian ini diusulkan sebuah pembangunan korpus bahasa, khususnya korpus Bahasa Indonesia, menggunakan teknologi web crawling yang terbimbing. Web crawling adalah sebuah kegiatan yang menggunakan script pemrograman sejenis bot yang berjalan pada internet untuk mengumpulkan data dan menyimpan ke database untuk dianalisis dan diatur lebih lanjut (Kancheria, 2014). Beberapa penelitian tentang pemanfaatan aplikasi yang berada di internet untuk pembangunan korpus bahasa sudah pernah dilakukan sebelumnya. Pada penelitiannya tentang pembangunan korpus untuk analisis sentimen secara otomatis, aplikasi berbasis web yang digunakan sebagai sumber data adalah aplikasi microblogging tweeter (Wicaksono et al, 2014). Korpus bahasa yang diambil ditujukan untuk keperluan ekstraksi opini dan polarisasi sentimen yang terkandung dalam masing-masing tweet. Metode yang digunakan dalam pengambilan datanya yaitu menggunakan tweeter streaming API. Namun penekanan pada penelitian tersebut adalah untuk mengembangkan korpus dataset yang sudah terpolarisasi untuk mengembangkan kemampuan machine learning dalam mengelompokkan dokumen, bukan optimalisasi dalam pembuatan sebuah korpus bahasa. Berikutnya adalah penelitian tentang pembuatan framework khusus untuk membangun korpus bahasa dari web yang didasarkan pada pipeline dan arsitektur modular (Adhikary, 2016). Pada penelitian tersebut digunakan Scrapy sebagai basis dari framework web crawler-nya, karena Scrapy bisa dirancang secara khusus agar bisa menjelajah situs yang diinginkan tanpa menggangu resource yang dimiliki oleh penyedia situs. Hal ini juga yang digunakan sebagai pertimbangan pemakaian web crawler Scrapy pada penelitian ini. Dengan menggunakan teknologi web crawling scrapy, tidak membutuhkan banyak waktu dan tenaga dalam menjelajah media internet untuk mengunjungi aplikasi tersebut satu persatu, karena penjelajahan tersebut bisa dilakukan secara otomatis dan paralel, sehingga dalam satu waktu ada beberapa aplikasi yang bisa dikunjungi. Selain itu, dengan menggunkan web crawling yang terbimbing, kita bisa mendapatkan data tekstual yang bersih dari kode html yang lain ataupun tautan-tautan yang tidak berhubungan dengan data yang diharapkan. Sehingga dengan digunakannya teknologi web crawling terbimbing ini diharapkan bisa membangun data korpus Bahasa yang sesuai dengan kebutuhan penelitian dan bersih dari kode dan tautan yang tidak diharapkan tanpa banyak memakan waktu dan tenaga.

II. RUMUSAN MASALAH

Berdasarkan latar belakang diatas, maka permasalahan yang dapat dijabarkan pada penelitian ini yang pertama adalah bagaimana membangun aplikasi web crawler yang bisa menjelajah situs tertentu dan mengambil data yang diinginkan sekaligus mengajarkan spider (bot web crawler) agar dapat menuju dan mengambil data pada halaman yang spesifik walaupun mengunjungi aplikasi yang berlainan. Yang kedua, bagaimana mengajarkan spider agar bisa bekerja secara paralel. Yang ketiga, bagaimana mendapatkan data yang bersih dari kode html dan tautan yang tidak berhubungan dengan data yang digutuhkan.

Mengacu pada masalah yang telah dirumuskan, maka batasan masalah yang akan dibahas dalam penelitian yang pertama adalah data yang akan diambil sebagai korpus adalah data berita yang akan dikumpulkan dalam kategori sesuai isi dari berita dan situs berita yang akan dijelajahi adalah sebuah weblog yang bernama blogdetik (http://blog.detik.com/).

III. TINJAUAN PUSTAKA

A. *Web Crawler*

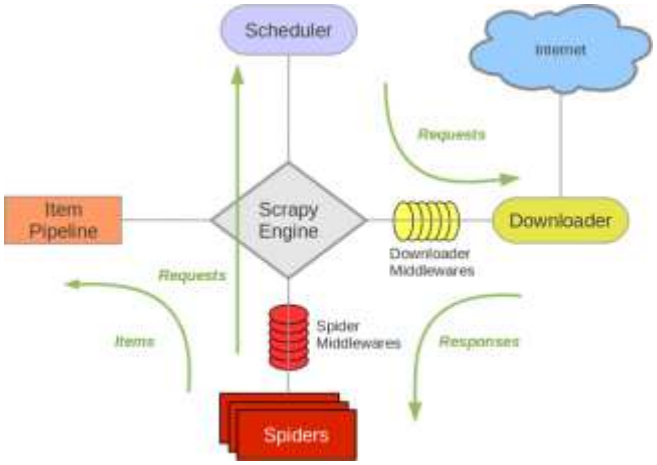
Web Crawler adalah suatu program atau script otomatis yang relatif simple, yang dengan mengunakan metode tertentu melakukan scan ke semua halaman-halaman Internet untuk membuat index dari data yang dicarinya. Pada umunya crawling diterapakn pada web yang banyak disebut *Web Crawling. Web Crawling* pada umumnya digunakan pada search engine yang dilakukan oleh sekelompok kompuer yang dikluster dimana setiap komputer menjalankan beberapa thread (Hatzi, 2014).

Prosedur dari *Web Crawling* dimulai dari memilih satu set URL yang akan dilakukan proses crawling dimana URL tersebut menyediakan banyak link ke halaman yang penting untuk proses Crawling. The crawler akan men-download konten ari halaman tersebut ke dalam penyimpanan local seperti hardisk dll. Secara bersaman thread yang ada pada *Crawler* mencari link baru dari halaman yang terhubung dengan halaman yang telah diunduh, link baru membentuk yang disebut dengan crawler's frontier. Tujuan utama dari crawler's frontier adalah mendapatkan halaman baru sebanyak mungkin dan mengupdate tingkat kebaruan halaman yag telah diunduh.

B. *Scrapy*

Scrapy adalah sebuah framework yang digunakan untuk melakukan proses crawling dan mengextract data yang

tersruktur. Scrapy digunakan pada proses data mining, pemrosesan informasi dan pengarsipan history. Scrapy dibangun dengan menggunakan pyhton yang disupport dengan twisted (Jing Wang, 2012). Terdapat tujuh komponen utama pada scrappy seperti yang ditunjukan pada gambar 1, yaitu Scheduler, Item Pipeline, Downloader, Downloade Middleware, Spiders, Spiders Middleware.



Gambar 1 . Arsitektur Scrapy

Scrapy Engine bertanggung jawab untuk mengendalikan arus data antar semua komponen sistem. Downloader bertangung jawab untuk mengambil halaman web yang diminta dan memasukananya kedalam engine. Spiders adalah sebuah class yang dibuat oleh user untuk memindah respon yang didapat dari engine dan mengektrak item dari respon tersebut. Pipeline bertanggung jawab untuk memproses item setelah item tersebut terekstrak oleh spiders. Downloader middlewares adalah perantara atau jembatan yang berada diantara engine dan downloader yang bertugas memproses request dari engine ke downloader dan memberikan respon dari downloader ke engine. Downloader middlewares menyediakan mekanisme yang sesuai untuk memperluas fungsi Scrapy dengan memasukkan kode yang dapat dirubah sesuai dengan kebutuhan.

C. XPATH

Xpath, XML *Path Language*, adalah bahasa *query* untuk memilih node dari dokumen XML. *Xpath* dapat digunakan untuk menghitung nilai (misalnya, string, angka, atau nilai Boolean) dari isi dokumen XML. Xpath didefinisikan oleh *World Wide Web Consortium*(W3C)(Abdilah, 2015). Bahasa *Xpath* didasarkan pada representasi pohon dokumen XML, dan menyediakan kemampuan untuk menavigasi di sekitar pohon XML, memilih node dengan berbagai kriteria. Dalam penggunaannya yang populer, meskipun tidak dalam spesifikasi resmi, ekspresi *Xpath* sering disebut hanya sebagai *Xpath*.

Xpath awalnya didorong oleh keinginan untuk menyediakan sintaks yang umum dan model dari perilaku Antara XPointer dan XSLT, yang merupakan subset dari bahasa query *Xpath* digunakan dalam spesifikasi W3C lainnya seperti XML Schema, XForms dan Tag Set Internasionalisasi (ITS). Xpath telah diadopsi oleh sejumlah library dan alat-alat pemrosesan XML, banyak yang juga menawarkan Selectors CSS, W3C memiliki standar yang berbeda, sebagai alternative sederhana untuk *Xpath*.

IV. METODOLOGI PENELITIAN

A. Objek Penelitian

Dalam penelitian ini, objek penelitiannya adalah sebuah weblog yang bernama blogdetik (http://blog.detik.com/). Blog ini disediakan oleh salah satu situs berita polpuler di Indonesia detik.com (http://www.detik.com) sebagai wadah untuk menampung karya tulis seluruh blogger di Indonesia, baik yang sudah memiliki blog sendiri maupun yang belum.

Ada tiga kategori dalam blog ini yang digunakan sebagai objek scraping yaitu, komunitas, hiburan dan kuliner. Hanya artikel yang berada dibawah kategori tersebut yang akan diekstrak. Hal ini untuk mengetahui apakah scrapy juga bisa digunakan sebagai web scrapng terbimbing.

B. Variable Penelitian

1. URL situs berita

URL situs berita yang merupakan singkatan dari “Unifrom Resource Locator” adalah rangkaian karakter menurut format standar tertentu, digunakan untuk menunjukan alamat dari suatu sumber dokumen berita yang terdapat di internet. Hanya beberapa alamat dari situs-situs berita tertentu yang akan dijelajahi oleh spider yang akan digunkan dalam penelitian ini. Hal ini juga digunkaan untuk melihat apakah spider yang dibuat mampu menjelajah satu alamat situs berita dalam derajat tertentu atau tidak.

2. Kategori berita

Dalam penelitian ini, korpus bahasa yang akan dibangun merupakan kumpulan dokumen berbentuk teks yang nantinya akan dikelompokkan dalam kategori yang berbeda. Kategori ini didasarkan pada isi dari dokumen yang akan disimpan.

3 Jumlah spider

Jumlah spider yang digunakan dalam proses web crawling juga bervariasi, hal ini untuk melihat performansi aplikasi apabila berada dalam skenario dimana ada lebih dari satu spider yang bekerja secara paralel, selain itu juga dilihat waktu komputasi yang dibutuhkan dalam menyelesaikan sebuah tugas dengan jumlah spider yang berbeda.

C. Parameter Penelitian

Parameter penelitian adalah beberapa aspek yang diamati sesuai dengan prosedur pengamatan melalui pengukuran yang telah ditentukan dalam kerangka metode penelitian yang digunakan. Pada penelitian ini, parameter yang digunakan untuk menilai kinerja system ada empat, yaitu;

1 Jumlah item

Jumlah item adalah parameter yang menunjukkan banyaknya item yang dapat disimpan oleh bot spider dalam satu proses crawling. Item-item yang disimpan adalah link, judul, deskripsi dan post. Link berisi tautan atau hyperlink dari artikel yang sudah dikunjungi, judul berisi judul artikel atau berita, deskripsi berisi penjelasan singkat tentang berita, dan post berisi konten berita secara lengkap.

2 Ukuran file

Ukuran file adalah nilai satuan yang menunjukkan pemakaian space media penyimpanan dari file json yang berada dalam computer client. File json yang didapatkan dari proses crawling adalah hasil penggabungan dari semua item yang berhasil disimpan.

3 Penggunaan memory

Pada saat aplikasi dijalankan, maka secara otomatis computer server akan menyediakan sebagian dari memory-nya untuk digunakan. Penggunaan memory ini mempengaruhi kinerja dari computer server. Semakin besar penggunaan memory ini, maka semakin berat juga kinerja dari computer pada saat menjalankan aplikasi tersebut.

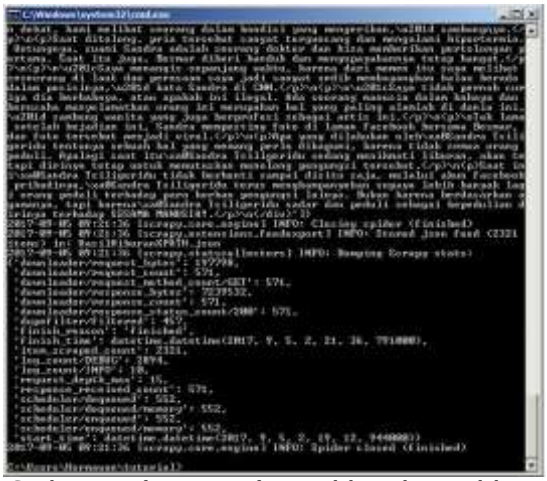
4 Waktu

Parameter waktu disini adalah untuk mengetahui berapa jumlah waktu yang digunakan untuk masing-masing proses web crawling pada tiap kategori berita. Semakin lama proses crawling dilakukan, maka semakin banyak pula resource yang digunakan.

V. HASIL DAN LUARAN YANG DICAPAI

Hasil luaran dari proses web crawling terhadap tiga kategori berita dalam penelitian ini disimpan kedalam file dengan ekstensi .JSON, yang kemudian disebut sebagai korpus berita komunitas, korpus berita hiburan dan korpus berita kuliner. Masing-masing korpus berita tersebut berisi hasil web scraping blogdetik menggunakan dua metode yaitu XPATH *Selecto*r yang disimpan secara terpisah berdasarkan kategori artikelnya.

Pada Gambar 1 ditunjukkan hasil dari proses crawling dan scraping berita pada kategori hiburan. Seluruh parameter yang dibutuhkan dalam penelitian sudah diatur agar dimunculkan pada setiap akhir proses sebagai system log, sehingga analysis terhadap parameter tersebut bisa dilakukan pada tiap akhir proses crawling dan scraping. Untuk proses crawling dan scrapping pada berita dengan kategori komunitas dan kuliner masing-masing ditunjukkan pada Ganbar 2 dan Gambar 3



Gambar 1. Hasil Proses crawling untuk berita kategori hiburan

Gambar 2. Hasil proses crawling untuk berita kategori komunitas

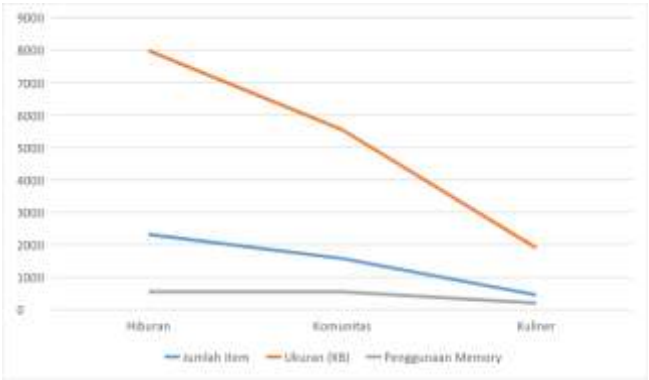
Gambar 3. Hasil proses crawling untuk berita kategori kuliner

Untuk memudahkan anaslis hasil perolehan proses crawling dan scraping data yang sudah didapatkan untuk tiap kategori berita dimasukkan kedalam Tabel 1.

TABEL I REKAPITULASI HASIL PROSES CRAWALING DAN SCRAPING TIAP KATEGORI BERITA

Parameter	Kategori		
	Hiburan	Komunitas	Kuliner
Jumlah Item	2321	1581	460
Ukuran (KB)	7991	5549	1905
Penggunaan Memory	552	553	202
Waktu	0:02:24	0:00:18	0:00:08

Dari Tabel 1 dapat diketahui bahwa jumlah item terbesar didapatkan dari berita dengan kategori Hiburan yaitu 2321 item. Hal ini disebabkan oleh dua kemungkinan, yang pertama jumlah artikel yang berada pada berita dengan kategori hiburan lebih banyak dari dua kategori yang lain, yang kedua, jumlah artikel yang memiliki tautan yang aktif atau bisa dikunjungi lebih banyak dari pada dua kategori yang lain, sehingga jumlah item yang berhasil disimpan setelah proses crawling dan scraping jauh lebih banyak dari dua kategori yang lain.



Gambar 4. Grafik Rekapitulasi Hasil Proses Crawaling Dan Scraping Tiap Kategori Berita

Gambar 4 meenunjukkan visualisasi dari jumlah item, ukuran dan penggunaan memori yang didapat dari Tabel 1. Dari grafik tersebut dapat dilihat penggunaan memory pada computer server berbanding lurus dengan jumlah item yang didapatkan dan ukuran file. Hal ini berarti semakin banyak item yang diperoleh dan berhasil disimpan maka semakin besar pula ukuran file dan resource memory yang digunakan. Dengan demikian, untuk membatasi penggunaan memory pada computer server, kita dapat memberi batasan item apa saja yang akan diambil pada saat proses scraping dengan membatasi jumlah tautan yang dijelajahi oleh spider atau membatasi jumlah item yang akan dismpan.

VI. KESIMPULAN DAN SARAN

Dari penelitian yang sudah dilakukan dapat disimpulkan bahwa pembangunan korpus berita menggunakan Scrapy dan Xpath berhasil memenuhi target yang diharapkan. Hal ini ditandai dengan dihasilkannya korpus berita yang sudah terbagi menjadi 3 kategori berita yaitu, berita hiburan, komunitas dan kuliner.

Selain itu, dari parameter yang diuji dapat diperoleh kesimpulan bahwa penggunaan resource pada computer server berbanding lurus dengan jumlah item yang didapatkan dan ukuran file. Hal ini berarti semakin banyak item yang diperoleh dan berhasil disimpan maka semakin besar pula ukuran file dan resource memory yang digunakan. Dengan demikian, untuk membatasi penggunaan memory pada computer server, kita dapat memberi batasan item apa saja yang akan diambil pada saat proses scraping dengan membatasi jumlah tautan yang dijelajahi oleh spider atau membatasi jumlah item yang akan dismpan.

DAFTAR PUSTAKA

[1] Akbar, Subhan Agus, Eko Sediyonob, Oky Dwi Nurhayati. 2015. *Analisis Sentimen Berbasis Ontologi di Level Kalimat untuk Mengukur Persepsi Produk. Jurnal Informasi Bisnis*. 02. 2015. Universitas Diponegoro. Semarang

[2] Kouzis-Loukas, Dimitrios. 2016. *Learning Scrapy*. Packt Publishing: Birmingham-Mumbai.

[3] Hatzi, Vassiliki, dkk. 2014. *Web Page Download Scheduling Policies for Green Web Crawling*. 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM).

[4] Jing Wang, Yuchun Guo. 2012. *Scrapy-based Crawling and User-behavior Characteristics Analysis on Taobao*. 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discover.

[5] Wijaya, Akhmad Pandu, Heru Agus Santoso. 2016. *Naive Bayes Classification* pada Klasifikasi Dokumen Untuk Identifikasi Konten *E-Government*. Journal of Applied Intelligent System. 2016;1(1):48-55