Rancang Bangun Aplikasi Web Scraping untuk Korpus Paralel Indonesia - Inggris dengan Metode HTML DOM

Vivensius Mitra[#1], Herry Sujaini[#2], Arif Bijaksana Putra Negara[#3]
[#]*Program Studi Teknik Informatika Universitas Tanjungpura*
Jl. Prof Dr H. Hadari Nawawi, Kota Pontianak, 78115
[1]vivensiusmitra@gmail.com,[2]herry_sujaini@yahoo.com,[3]arifbpn@gmail.com

Abstrak **- Korpus paralel merupakan dua dokumen text yang saling berhubungan dimana dokumen text pertama berisi kumpulan kalimat sumber dan dokumen kedua berisi kumpulan kalimat terjemahan. Korpus paralel berfungsi sebagai sumber utama dalam mengembangkan mesin penerjemah statistik. Pengumpulan korpus paralel secara manual memerlukan waktu yang lama dan biaya yang tidak sedikit. *Web scraping* adalah suatu teknik penggalian informasi dari situs web. Pembuatan aplikasi web scraping dapat dikombinasikan dengan berbagai metode, dalam penelitian ini metode yang digunakan adalah HTML DOM. Sistem ini dibangun untuk mengumpulkan korpus paralel Bahasa Indonesia dan Inggris. Pengujian dari aplikasi ini adalah menggunakan metode *blackbox*, serta beberapa rangkaian pengujian secara manual untuk mengetahui tingkat keberhasilan aplikasi ini dalam mengumpulkan data korpus paralel dan kecepatan sistem dalam mengumpulkan korpus paralel. Hasil implementasi dan pengujian akhir dari aplikasi *web scraping* dengan metode HTML DOM adalah proses yang berjalan dalam aplikasi *web scraping* dengan metode HTML DOM adalah proses *scraping*, tokenisasi, *cleaning*, dan *lowercased*, semua proses tersebut berjalan secara otomatis sehingga sangat menghemat waktu dan biaya dan menghasilkan korpus paralel Bahasa Indonesia dan Inggris.**

Kata kunci : Web Scraping, HTML DOM, Korpus Paralel, Tokenisasi, Mesin Penerjemah Statistik

I. PENDAHULUAN

Korpus paralel adalah dua buah kumpulan dokumen yang memiliki isi sama dan ditulis dalam bahasa yang berbeda. Korpus paralel berguna bagi para peneliti khususnya dalam bidang pemrosesan bahasa alami untuk pengembangan mesin penerjemah statistik [1]. Hasil terjemahan dari sebuah sistem penerjemah statistik tergantung pada jumlah dari paralel korpus yang tersedia, karena jika semakin banyak jumlah paralel korpus yang dimiliki, maka akan semakin baik pula hasil terjemahannya. Korpus paralel banyak tersimpan dalam "hard disk" para peneliti, tapi sangat sedikit yang di-share [2]. Korpus paralel yang dipublikasikan saat ini oleh Septina Dian Larasati IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus berjumlah 27.326 kalimat (Korpus Identic sudah menggabungkan isi dari korpus Lembaga Pengkajian dan Penelitian) [3].

Secara umum tahap pertama untuk membuat korpus paralel adalah mengumpulkan data berupa text atau dokumen dari berbagai sumber yang telah ada seperti dari situs berita, kamus, subtitle film dan masih banyak sumber-sumber lainnya. Setelah data mentah korpus paralel telah terkumpul maka data tersebut akan melalui beberapa proses yaitu *cleaning*, tokenisasi, dan *lowercased*. Fungsi dari proses *cleaning* yaitu menyaring data, seperti menghilangkan kata yang terlalu panjang sesuai dengan batas yang ditentukan. Selanjutnya proses tokenisasi adalah proses pemisahan suatu karakter spasi dan mungkin pada waktu yang bersamaan dilakukan proses penghapusan karakter tertentu, seperti tanda baca. *lowercase* adalah untuk mengubah huruf dalam dokumen menjadi huruf kecil. Setelah menyelesaikan ketiga proses tersebut maka korpus paralel yang dihasilkan sudah siap untuk digunakan dalam mesin penerjemah.

Setelah menganalisis beberapa tahapan di atas maka pembuatan korpus paralel khususnya bahasa Indonesia – Inggris secara manual sangat memakan waktu yang lama, biaya yang tidak sedikit dan terbatasnya jumlah korpus paralel yang didapatkan. Oleh karena itu diperlukan suatu sistem otomatis yang dapat membuat sebuah paralel korpus dengan cepat dan efektif agar dapat menambah perbendaharaan korpus yang sudah ada. *Web Scraping* adalah aplikasi yang secara khusus dikembangkan untuk mengekstraksi informasi dari berbagai situs. Aplikasi ini berguna untuk mengumpulkan beberapa bentuk data dari internet. Dalam penelitian ini web scraping dapat digunakan untuk mengambil dokumen text dari sebuah website yang berisi dokumen dua bahasa.

Berdasarkan permasalahan di atas, penelitian ini melakukan analisis, perancangan dan pembuatan aplikasi web scraping dengan metode HTML DOM yang difokuskan untuk menghasilkan korpus paralel bahasa Indonesia - Inggris. Situs yang akan diambil data kalimatnya adalah sebuah situs berita dua bahasa (http://berita2bahasa.com/). Adanya aplikasi ini diharapkan dapat memperkaya perbendaharaan korpus dalam bahasa Indonesia - Inggris.

II. TIJAUAN PUSTAKA

A. Web Scraping

Web Scraping adalah proses pengambilan sebuah dokumen semi - terstruktur dari internet, umumnya berupa halaman - halaman *web* dalam bahasa *markup* seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain [4]. *Web scraping* sering dikenal sebagai screen scraping . *Web Scraping* tidak dapat dimasukkan dalam bidang *data mining* karena data mining menyiratkan upaya untuk memahami pola semantik atau tren dari sejumlah besar data yang telah diperoleh. Aplikasi *web scraping* (juga disebut *intelligent, automated, or autonomous agents*) hanya fokus pada cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi [5]. *Web scrapping* berhubungan dengan pengindeksan web yang merupakan suatu teknik *universal* yang dipakai hampir semua *search engine*. Perbedaannya *web scraping* lebih berfokus pada transformasi dari suatu web yang tidak terstruktur, umumnya dalam format HTML menjadi suatu format data terstruktur yang dapat disimpan dan dianalisa pada *database* atau lembar kerja.

B. HTML DOM

HTML DOM adalah standar untuk bagaimana untuk mendapatkan, mengubah, menambah, atau menghapus elemen HTML (*http://www.w3ii.com/id///js/js_HTMLdom.HTML :* diterjemahkan dengan *google translate*) [6]. HTML DOM mendefinisikan objek dan properti dari semua elemen HTML, dengan

Submitted 11-01-2017; Revised 11-01-2017; Accepted 11-01-2017

metode untuk mengaksesnya. Dengan DOM, *JavaScript* dapat mengakses semua elemen didalam dokumen HTML. DOM adalah *interface* yang bersifat netral terhadap platform dan bahasa yang membuat program dan script dapat mengakses secara dinamis dan mengupdate struktur, *style,* dan konten dokumen. HTML DOM menggunakan bahasa pemrograman untuk mengakses obyek-obyeknya, biasanya *JavaScript*. Semua elemen HTML diperlakukan sebagai obyek. Antarmuka pemrogramannya adalah metode dan properti dari setiap obyek.

C. *Cleaning* dan *Lowercase*

Cleaning adalah tahapan menyaring data, seperti menghilangkan kata yang terlalu panjang sesuai dengan batas yang ditentukan. *Lowercase* adalah proses penyeragaman semua huruf menjadi huruf kecil semua [1].

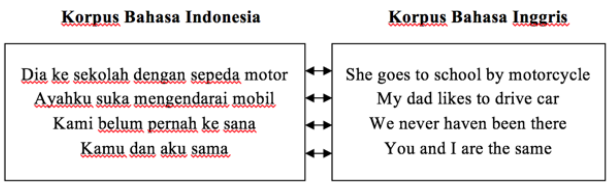
D. *Tokenisasi*

Tokenisasi adalah tahap pemrosesan dimana teks *input* dibagi menjadi unit-unit kecil yang disebut token, yang dapat berupa suatu kata, suatu angka, atau suatu tanda baca (*http://nlp.stanford.edu/ir-book/html/ htmledition/tokenization-1.html* : diterjemahkan menggunakan *google translate*) [7]. Proses ini cukup rumit untuk sebuat program komputer karena beberapa karakter dapat dijadikan sebagai pembatas (delimiter) dari token-token itu sendiri. Pembatas dari token tersebut antara lain spasi, tab dan baris baru, sedangkan karakter ( ) < > ! ? " . , terkadang dianggap sebagai pembatas dan juga bukan pembatas tergantung pada kondisi pemakainya.

E. *Korpus*

Korpus merupakan kumpulan teks berupa kata atau kalimat dalam ukuran besar dan terstruktur. Korpus dapat berisi *text* dalam satu bahasa *(*korpus *monolingual)* atau berbagai macam bahasa *(*korpus *multilingual)* dan dapat disimpan dalam bentuk *file text*. Salah satu kegunaan korpus yaitu sebagai data *training* untuk mendukung *probabilistic translation model* yang dibutuhkan oleh *Cross Language Information Retrieval* (CLIR) dan *Machine Translation* (MT) [8].

Gambar 2.1 adalah contoh beberapa korpus paralel bahasa indonesia dan bahasa inggris :

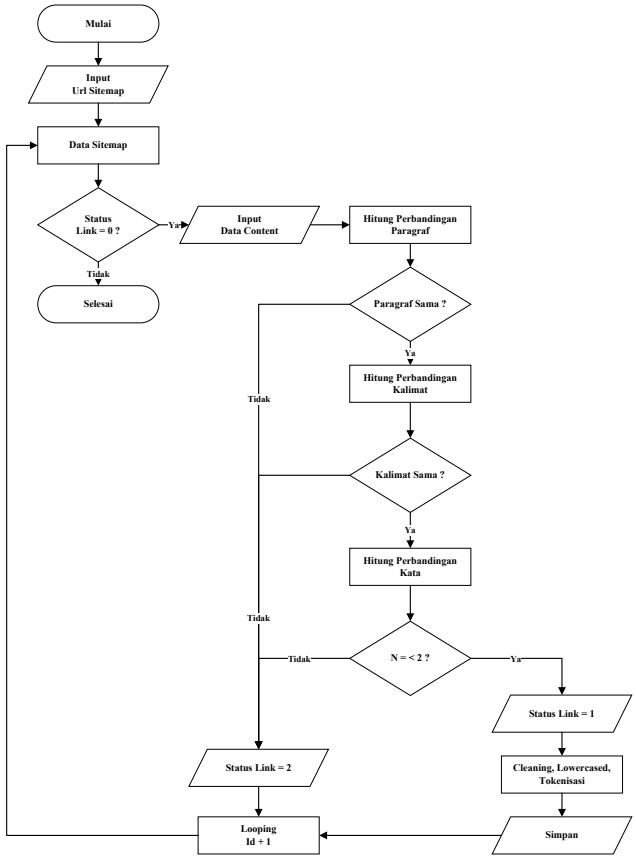


Gambar 2.1 Korpus paralel bahasa Indonesia dan bahasa Inggris

III. PERANCANGAN SISTEM

A. *Peracnganan Flowchart Sistem*

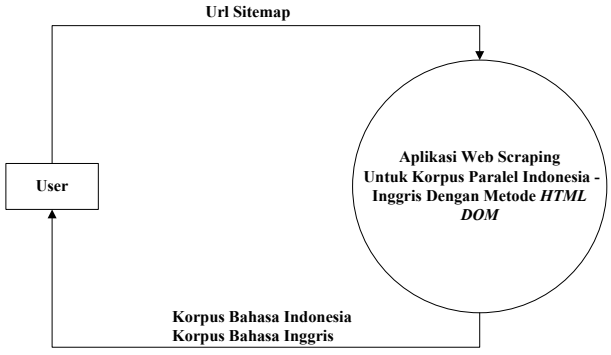
Flowcart sistem dalam aplikasi ini adalah sistem berjalan secara otomatis mulai dari pengambilan *sitemap* pada *website* sasaran sampai secara terus menerus sampai selesai mendapatkan semua data berupa korpus paralel Bahasa Indonesia - Inggris. Gambar 3.1 ini adalah cara kerja sistem dari aplikasi *web scraping* dengan metode HTML DOM dalam berjalan :



Gambar 3.1 *Flowchart* Sistem

B. *Perancangan Diagram Konteks*

Diagram konteks adalah diagram yang memberikan gambaran umum terhadap kegiatan yang berlangsung dalam sistem. Gambar 3.2 menunjukan diagram konteks dari sistem.



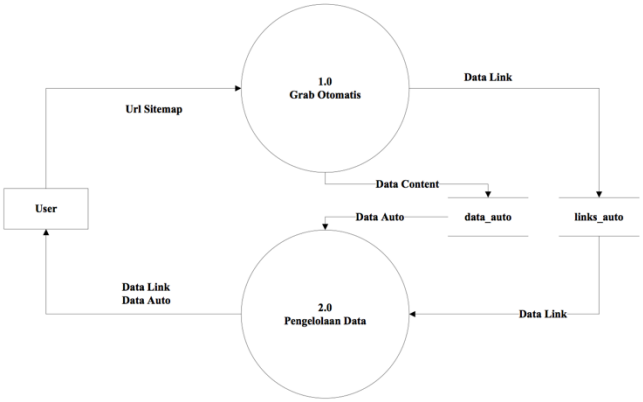
Gambar 3.2 Diagram Konteks Sistem

C. *Perancangan Diagram Overview Sistem*

Diagram *overview* adalah diagram yang menjelaskan urutan-

Korespondensi : Vivensius Mitra

urutan proses dari diagram konteks. Seperti pada Gambar 3.3, sistem ini dibagi menjadi empat dua.

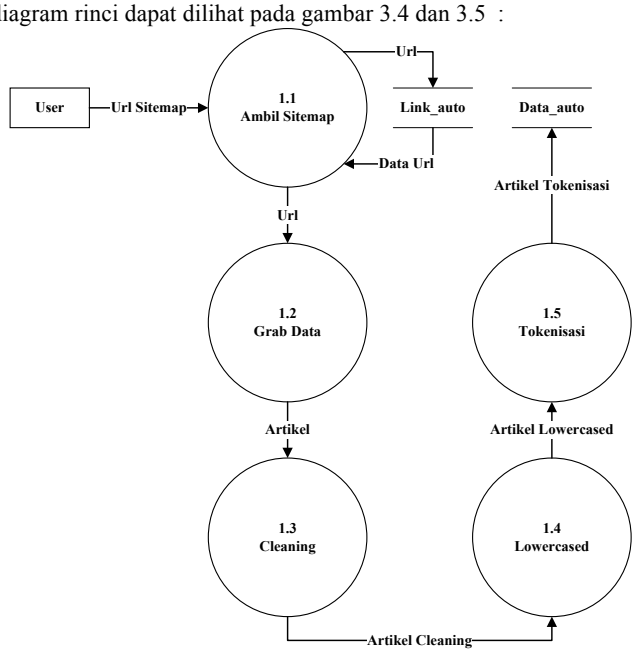


Gambar 3.3 Diagram *Overview* Sistem

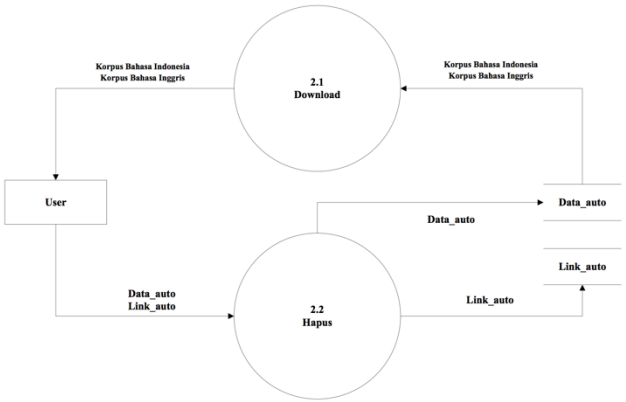
D. Perancangan Diagram Rinci Sistem

Diagram Rinci Level 2

Diagram rinci menguraikan lebih lanjut mengenai proses dari diagram *overview* yang memperlihatkan arus data masuk dan arus data keluar. Berdasarkan diagram *overview* di atas, maka model diagram rinci dapat dilihat pada gambar 3.4 dan 3.5 :



Gambar 3.4 Diagram Rinci *Level* 2 Proses 1.0 Grab Otomatis



Gambar 3.5 Diagram Rinci *Level* 2 Proses 2.0 Pengelolaan Data

E. Perancangan Basis Data

Peracangan Tabel Data

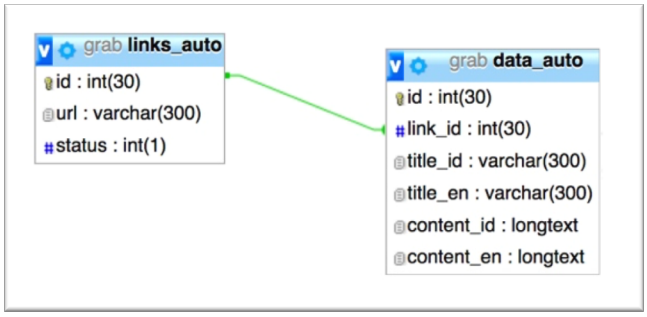
Tabel link_auto merupakan tabel yang digunakan untuk menyimpan data url yang diperoleh dari sitemap situs berita dua bahasa. Spesifikasi tabel link_auto dapat dilihat pada tabel 3.1.

Tabel 3.1 Spesifikasi tabel link_auto

Nama Field	Tipe	Keterangan	Keterangan
id	int(30)	Primary Key	Menyimpan id url
url	varchar(255)	Not Null	Menyimpan url
status	int(1)	Not Null	Menyimpan status url

Peracangan Realasi Antar Tabel

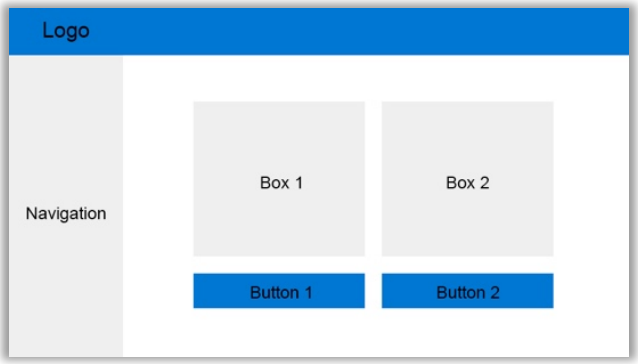
Perancangan diagram relasi antar tabel menggambarkan adanya relasi antar tabel yang terdapat dalam Aplikasi *Web Scraping* Untuk Korpus Paralel Indonesia - Inggris dengan Metode HTML DOM. Relasi antar tabel ini berfungsi untuk meminimalisir resiko data *redundancy* dan pemborosan *memory*. Relasi antar tabel ditunjukkan pada Gambar 3.6.



Gambar 3.6 Relasi Antar Tabel

F. Perancangan Antar Muka

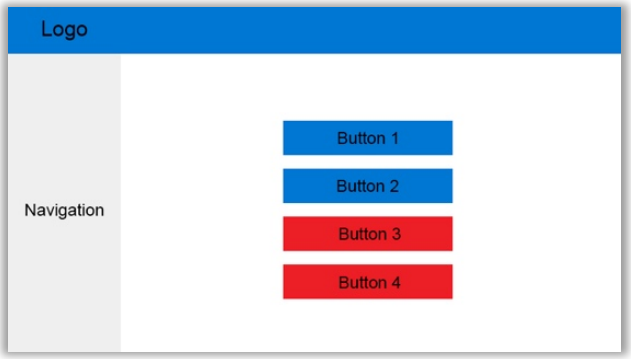
Halaman *grab* otomatis adalah halam utama dalam aplikasi ini, pada halaman ini *user* dapat memulai melakukan proses mengambil data korpus paralel pada *website* sasaran. Rancangan halaman *grab* otomatis dapat dilihat pada gambar 3.7.



Gambar 3.7 Perancangan antarmuka halaman dasbor

Halaman *grab* otomatis adalah halam utama dalam aplikasi ini, pada halaman ini *user* dapat memulai melakukan proses mengambil data korpus paralel pada *website* sasaran. Rancangan halaman *grab* otomatis dapat dilihat pada gambar 3.10.

Korespondensi : Vivensius Mitra



Gambar 3.8 Perancangan antarmuka halaman *grab* otomatis

IV. HASIL DAN ANALISIS PENGUJIAN

A. Hasil Perancanga Antarmuka

Halaman Dasbor

Halaman dasbor adalah halaman pertama yang terbuka ketika *user* mengakses aplikasi *web scraping* untuk korpus paralel Indonesia - Inggris dengan metode HTML DOM. Dalam halaman dasbor terdapat 2 (dua) buah box yang masing-masing menampilkan gambar bendera indonesia dan inggris untuk menandakan posisi bahasa dan di bawahnya terdapat jumlah kalimat dari masing-masih bahasa yang tersimpan di dalam database, serta 2 (dua) buah button yang berfungsi untuk mengunduh data korpus paralel dari database. Tampilan dari halaman dasbors dapat dilihat pada gambar 4.1.

Nama file : ***index.php***

Gambar 4.1 Antarmuka Halaman Dasbor

Penjelasan :

1. Kolom dengan bedera Indonesia menampilkan jumlah kalimat bahasa Indonesia yang dihasilkan oleh aplikasi.
2. *Button Download TXT* yang terdapat dibawah kolom dengan bendera Indonesia berfungsi untuk mengunduh kalimat bahsa Indonesia ke dalam komputer.
3. Kolom dengan bedera Inggris menampilkan jumlah kalimat bahasa Inggris yang dihasilkan oleh aplikasi.
4. *Button Download TXT* yang terdapat dibawah kolom dengan bendera Inggris berfungsi untuk mengunduh kalimat bahsa Inggris ke dalam komputer.

Halaman Grab Otomatis

Halaman grab otomatis adalalah halaman dimana semua proses yang berjalan pada sistem. Tampilan dari halaman grab otomatis terdapat pada gambar 4.2.

Gambar 4.2 Antarmuka Halaman Grab Otomatis

Penjelasan :

1. *Button* Mulai *Grab* (Otomatis) : berfungsi untuk menjalankan perintah web scraping yang diawali pengambilan data sitemap, pengambilan data kalimat pada setiap berita dan mejalankan proses tokenisasi, *cleaning* serta *lowercased* kemudian menyimpan hasilnya ke dalam database secara bersamaan pada website Berita 2 (dua) Bahasa secara otomatis.
2. Button *Grab* Data : berfungsi untuk melanjutkan grab data jika proses tersebut terhenti karena adanya gangguan koneksi pada internet.
3. Button Hapus Semua *Link* : berfungsi untuk menghapus semua link yang terdapat di dalam database.
4. Button Hapus Semua Data : berfungsi untuk menghapus semua data kalimat yang terdapat di dalam database.

Bentuk *output* akhir dari aplikasi *web scraping* untuk korpus paralel Indonesia - Inggris dengan metode HTML DOM yang dihasilkan seperti pada gambar 4.3.


full-id.txt

1 mentan tantang fakultas pertanian ptn di indonesia timur kembangkan lahan pertanian
2 samarinda , kalimantan timur (b2b) - menteri pertanian ri andi amran sulaiman menantang sekaligus memberi peluang kepada fakultas - fakultas pertanian pada perguruan tinggi negeri (ptn) kawasan timur indonesia (kti) , untuk mengembangkan lahan pertanian seluas 1.000 hektar , pemerintah akan memberi modal awal berupa benih unggul , pupuk , dan alat mesin pertanian (alsintan) kemudian membelinya setelah panen pertama
3 lahan pertanian 1.000 hektar disediakan oleh universitas untuk digarap oleh mahasiswa dengan masyarakat sekitar kampus , mereka wajib dibimbing oleh dosen fakultas pertanian
4 untuk tahap awal , pemerintah akan membantu menyediakan benih unggul , pupuk dan alsintan dan akan kami beli hasil panen pertama , kata mentan amran sulaiman pada rapat koordinasi konsorsium ptn - kti di samarinda , provinsi kalimantan timur pada selasa malam (11 / 10)
5 menurutnya , langkah tersebut untuk menghimpun potensi yang dimiliki oleh setiap ptn untuk mendukung langkah pemerintah khususnya kementerian pertanian (kementan) meningkatkan kontribusinya mendukung pencapaian swasembada pangan di wilayah kt
6 kalau setuju saya berikan pupuk gratis berikut benih dan alsintan , cukup seribu hektar dulu setiap universitas dengan melibatkan semua potensi di fakultas pertanian bersama masyarakat sekitar , bisa menanam padi , jagung , kedelai dan hortikultura maka mahasiswa pun semakin mencintai bidang ilmu yang telah mereka pilih dan tekuni di kampus , kata mentan
7 tantangan tersebut disambut baik oleh ketua kptn - kti prof dr h masjaya dan gubernur kalimantan timur (kaltim) awang faroek ishak , para rektor , wakil rektor , dan dekan fakultas pertanian dan peternakan di wilayah kt
8 dukungan ptn masjaya menyatakan dukungannya pada program kementan membangun pertanian di kalimantan untuk menyukseskan swasembada pangan di kti khususnya kalimantan , mengingat luas daratannya masih luas dan kalangan ptn semakin optimis setelah pemerintah ri membuktikan komitmennya di bidang pangan dengan menekan impor jagung hampir 60 % pada 2015
9 kami siap mendukung pak menteri untuk mensukseskan dan siap membantu program - program yang dicanangkan terutama yang terkait ketahanan pangan di kalimantan , kata masjaya
10 mentan menambahkan , kalangan akademisi dari ptn baik dari tenaga pendidik maupun mahasiswa berperan penting mendukung swasembada pangan , dengan menjadi tenaga pendamping dan penyuluh pertanian untuk para petani di kalimantan
11 selain itu , kalangan akademisi juga dapat berperan dengan menciptakan inovasi - inovasi di bidang teknologi pertanian dan pemerintah sangat siap mensukseskan program swasembada pangan di kalimantan
12 mourinho main tebak foto , heboh ketika kenali paul pogba ketika bocah
13 jose mourinho terlihat seperti orang yang paling santai di dunia ketika dia kembali ke ruang kerjanya , kakinya disilangkan dan tangannya berada di lutut
14 manajer manchester united diberi waktu untuk mengenali siapa para pemain yang dilatihnya saat ini
15 dari foto - foto mereka ketika masih bocah , sebelum mengungkapkan kesan pertama dari para pemainnya
16 mantan bos chelsea , terkenal tenang menangani situasi tekanan di pertandingan besar , tampak kehilangan percaya diri karena ia dihadapkan pada foto dari beberapa orang khususnya yang diambil langsung dari album keluarga

Gambar 4.3 Hasil *Output* Akhir dalam bentuk *file* dengan format *.txt*

B. Analisis Pengujian

Hasil Pengujian Blackbox

Pengujian sistem merupakan bagian yang penting dalam siklus pembangunan perangkat lunak. Pengujian dilakukan untuk menjamin kualitas dan untuk mengetahui kelemahan dari Perangkat lunak. Tujuan dari pengujian ini adalah untuk menjamin bahwa Perangkat lunak yang memiliki kualitas yang baik yaitu mampu untuk mempersentasikan kajian pokok dari spesifikasi, analisis, perancangan, dan implementasi dari perangkat lunak itu sendiri [9]. Hasil pengujian terdapat dalam tabel 4.1.

Tabel 4.1 Pengujian Sistem Blackbox

Korespondensi : Vivensius Mitra

No.	Item Uji	Detail Pengujian	Jenis Pengujian	Hasil
1.	Grab Otomatis	Pengambilan data sitemap	*Black box*	Berhasil
2.	Grab Otomatis	Pengambilan Data Kalimat	*Black box*	Berhasil
3.	Button Hapus	Penghapusan Link	*Black box*	Berhasil
4.	Button Hapus	Penghapusan Data	*Black box*	Berhasil
5.	Button Download	Download data Id	*Black box*	Berhasil
6.	Button Download	Download data en	*Black box*	Berhasil

Setelah melakukan pengujian blackbox maka dan semua item uji yang dinyatan berhasil maka selanjutnya dilakukan pengujian hasil dari output aplikasi berupa korpus paralel bahasa Indonesia dan Inggris.

Pengujian kalimat hasil output berguna untuk mengetahui apakah sistem yang dibangun dalam penelitian ini sudah mencapai tujuan atau tidak. Pengujian diambil dari hasil output yang dihasilkan oleh aplikasi web scraping berupa pasangan kalimat bahasa indonesia dan bahasa inggris yang menghasilkan korpus paralel. Proses pengambilan data menggunakan nilai perbandingan kata rentang 0 - 2 pada website berita dua bahasa dengan menggunakan aplikasi web scraping memakan waktu 12 jam dan menghasilkan 38.712 pasang korpus paralel bahasa indonesia - inggris yang dapat digunakan untuk kebutuhan mesin penerjemah.

Penentuan ukuran sampel yang akan dalam penelitian ini dilakukan dengan menggunakan rumus Slovin sebagai berikut (Sugiyono : 2006) :

$$n = \frac{N}{1 + Ne^2}$$

dimana :

n : Jumlah sampel

N : Jumlah populasi

e : batas toleransi kesalahan, pada kasus ini nilai e = 10%(0,1)

Dari rumus di atas, maka besar jumlah sampel (n) korpus paralel yang digunakan adalah sebagai berikut :

$$n = \frac{38712}{1 + 38712\,(0{,}1)^2} = 99{,}74 \text{ (dibulatkan keatas menjadi 100)}$$

Tingkat Persentase Data

Tingkat persentase data yang dihasilkan oleh aplikasi *web scraping* dengan metode HTML DOM yang mengambil sumber data dari situs berita dua bahasa. Keseluruhan link url yang diambil didalam situs berita dua bahasa adalah sebanyak 8.951 buah dan data yang berhasil diambil sebagai korpus paralel sebanyak 4.175 buah. Untuk lebih jelas dapat dilihat dalam perhitungan berikut.

$$n = \frac{4.175}{8.951} \text{ x } 100 = 46{,}6\ \%$$

Berdasarkan perhitungan tersebut maka total artikel yang berhasil diambil oleh aplikasi web scraping dengan metode HTML DOM sebesar 46,6% dan sebesar 53.4 % gagal diproses. Aplikasi ini hanya mengambil sesuai dengan algoritma yang dirancang hanya untuk kalimat yang paralel dan penulisan kalimat yang sejajar antara bahasa indonesia dan bahasa inggris. Penyaringan kalimat yang sangat rinci menyebabkan data yang diambil tidak banyak dan salah satu faktor yang mempengaruhi data yang ditidak diambil karena penulisan yang dilakukan oleh pihak pengelola situs berita dua bahasa tidak paralel antara bahasa indonesia sebagai sumber dan bahasa inggris sebagai terjemahannya.

Hasil Pengujian Kecepatan

Pada pengujian kecepatan dalam pengumpulan paralel korpus ini tools yang digunakan sebagai pengukur kecepatannya adalah sebuah stopwatch pada sebuah perangkat Smartphone Apple. Internet Service Provider (ISP) yang digunakan adalah IndiHome dengan kecepatan 10 MBps dan 20 MBps dan browser yang digunakan adalah Mozila Firefox. Perangkat keras yang digunakan adalah MacBook dengan Sistem Operasi IOS, Prosesor Intel Core m3, RAM 8 GB, SSD 256 dan PC DELL dengan Sistem Opersi Windows 7, Prosesor Core i3, RAM 2GB, HDD 128. Pengujian ini dilakukan untuk melihat kecepatan aplikasi dalam pengumpulan korpus paralel pada dapa situs Berita dua Bahasa dengan perangkat keras yang berbeda dan kecepatan internet yang berbeda. Hasil pengujian ini dijabarkan pada tabel 4.2.

Tabel 4.2 Hasil Pengujian Kecepatan Pengumpulan Korpus Paralel

No.	Jumlah Halaman	Kecepatan Internet	Kecepatan Aplikasi		Jamlah Kalimat Paralel
			MacBook	PC DELL	
1.	4.175	10 MBps	19h32m16s	37h14m32s	38.712
2.	4.175	20 MBps	12h49m37s	22h54m9s	38.712

Dari hasil pengujian kecepatan aplikasi dalam mengumpulkan data korpus paralel pada tabel 4.2, spesifikasi perangat keras dan kecepatan internet yang digunakan sangat mempengaruhi waktu yang dibutuhkan oleh aplikasi dalam mengumpulkan korpus pararalel dari situs Berita dua Bahasa. Dapat dilihat pada tabel perangkat keras MacBook lebih cepat dibandingkan PC DELL.

V. KESIMPULAN DAN SARAN

A. *Kesimpulan*

Berdasarkan hasil dari pengujian dan analisis terhadap sistem Aplikasi Web Scraping Untuk Korpus Paralel Indonesia - Inggris Dengan Metode HTML DOM, maka ditarik kesimpulan sebagai berikut :

1. Sistem mampu menghasilkan dokumen korpus paralel melalui proses scraping dengan metode HTML DOM dari website Berita dua Bahasa dengan alamat URL (http://www.berita2bahasa.com/) dan mampu menghasilkan dokumen yang berisi kumpulan berita dan artikel Bahasa Indonesia sebagai sumber dan Bahasa Inggris sebagai terjemahan.
2. Sistem dapat menjalankan proses *cleaning*, *lowercased*, tokenisasi dan menyimpan data korpus paralel Bahasa Indonesia - Inggris secara otomatis tanpa melalui tahapan manual.
3. Aplikasi *Web Scraping* Untuk Korpus Paralel Indonesia - Inggris Dengan Metode HTML DOM, telah menghasilkan korpus paralel Bahasa Indonesia - Inggri sebanyak 38.712 pasang korpus paralel.
4. Pengumpulan korpus paralel dengan Aplikasi Web Scraping dengan Metode HTML DOM sangat berpengaruh pada spesifikasi perangkat keras dan kecepatan internet yang digunakan, perbandingan ini telah dibuktikan sesuai pada tabel 4.2, meskipun waktu yang dibutuhkan cukup lama namun sesuai pada jumlah korpus paralel yang diperoleh, dan sistem ini jauh lebih praktis dibandingkan dengan proses manual dalam pengumpulan korpus paralel.

B. *Saran*

Berdasarkan dari hasil analisis, implementasi dan pengujian sistem aplikasi web scraping untuk korpus paralel Indonesia - Inggris dengan metode HTML DOM, maka peneliti menyarankan beberapa hal sebagai berikut :

1. Sistem yang telah dibangun ini bisa dikembangkan ke arah pembangunan algoritma baru yang dapat mengenal dan

Korespondensi : Vivensius Mitra

memperbaiki penulisan dan tanda baca secara otomatis pada masing-masing bahasa yaitu bahasa Indonesia-Inggris. Banyaknya presentase data yang gagal diproses membuka peluang bagi para peneliti untuk mengembangkan algoritma yang lebih baik untuk meningkatkan persentase data yang didapatkan dan artikel yang dapat diambil tanpa mengurangi kualitas korpus.

2. Kepada pihak pengelola situs Berita dua Bahasa agar lebih memperhatikan penulisan artikel, karena masih terdapat banyak artikel yang tidak sesuai antara artikel bahasa Indonesia dan artikel bahasa Inggris.

DAFTAR PUSTAKA

[1] Septiandri, Edy. 2015. Rancang Bangun Aplikasi Information Retrieval Untuk Mengkoleksi Data Paralel Korpus Teks Bahasa Inggris – Bahasa Indonesia.Skripsi. Pontianak : Fakultas Teknik, Universitas Tanjungpura.

[2] Sujaini, Herry. 2012. An Approach to Improving Corpus Quality for Indonesian - English Statistical Machine Translation. International Journal of Engineering Research & Technology (IJERT) ISSN : 2278-0181 Vol. 4 Issue 02.

[3] Larasati, Septina Dian. 2012. IDENTIC Corpus: Morphologically EnrichedIndonesian-English Parallel Corpus. /[12 Oktober 2016] Unduh:http://ufal.mff.cuni.cz/~larasati/identic/.Muchtar, Januar. 2009.

[4] Turland, M. 2010. php architect's Guide to Web Scraping with PHP. Introduction-Web Scraping, str, 2.

[5] Josi, Ahmat., Abdillah, L.A., & Suryayusra. 2014. Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah. Jurnal Sistem Invormasi (JSI), 5(2), 159- 164.

[6] HTML DOM.. http://www.w3ii.com/id///js/js_htmldom.html

[7] Tokenization. http://nlp.stanford.edu/IR-book/html/htmleditiontokenization-1.html.

[8] Siagian, Adelina Irmadewita. 2012. Implementasi Corpus Generator Dengan Parallel Text. Skripsi. Medan : Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.

[9] Ariani, Sukamto Rosa. 2009. Black-Box Testing, Testing dan Implementasi Sistem.

Korespondensi : Vivensius Mitra