INTEGRASI LAMAN WEB TENTANG PARIWISATA DAERAH ISTIMEWA YOGYAKARTA MEMANFAATKAN TEKNOLOGI WEB SCRAPING DAN TEXT MINING

Muhammad Rifqi Ma'arif

Program Studi Manajemen Informatika
STMIK Jenderal Achmad Yani Yogyakarta

rifqi@stmikayani.ac.id

Abstrak

Banyaknya sumber informassi yang ada di satu sisi akan memberikan manfaat, namun di sisi lain akan menimbulkan fenomena information overload. Information overload adalah banyaknya jumlah informasi yang diterima oleh manusia sehingga menimbulkan kesulitan dalam penerimaan dan pengolahannya. Fenomena information overload salah satunya terjadi pada informasi pariwisata yang ada di Daerah Istimewa Yogyakarta (DIY). Dengan banyaknya laman web yang menyajikan informasi mengenai pariwisata DIY, calon wisatawan harus menyediakan lebih banyak waktu untuk memilah dan mengakses sebanyak mungkin laman web guna mendapatkan informasi yang lengkap dan akurat.Penelitian ini bertujuan untuk membangun sebuah laman web yang mampu mengintegrasikan informasi dari laman-laman web yang lain yang memuat informasi mengenai pariwisata DIY. Integrasi informasi akan dibuat dengan memanfaatkan teknologi web scraping dan text mining. Dengan adanya laman web yang mengintegrasikan informasi dari laman-laman web yang lain, calon wisatawan tidak perlu lagi menghabiskan banyak waktu untuk mencari informasi pariwsata DIY yang lengkap dan akurat.

Kata Kunci: *web scraping*, *text mining*, pariwisata DIY.

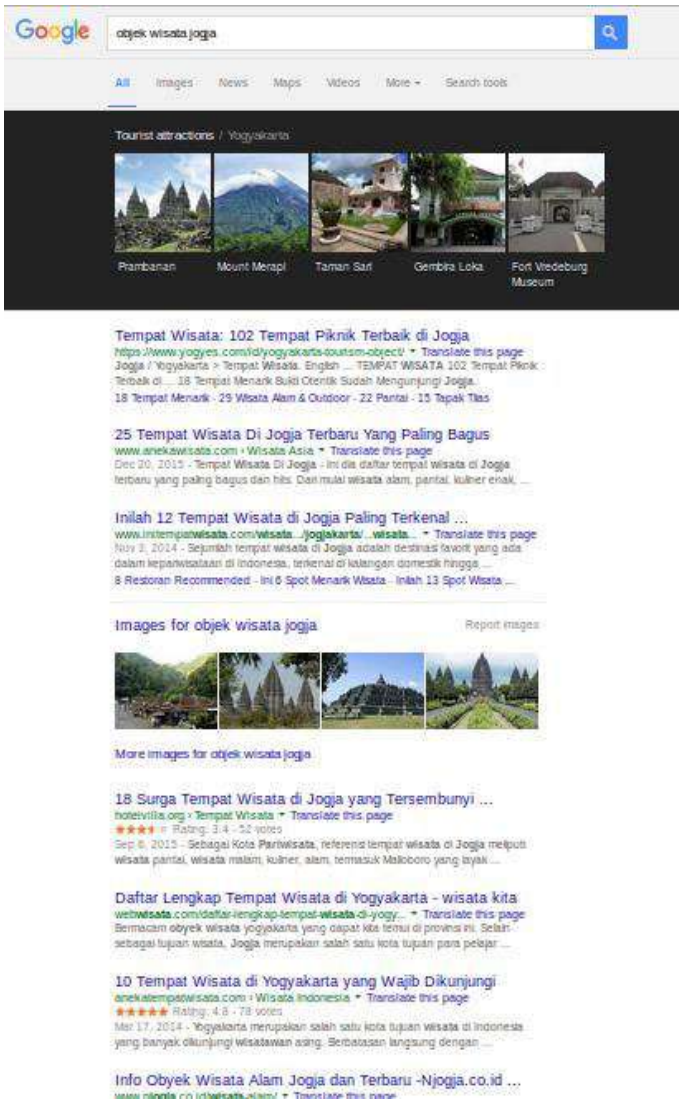
1. Pendahuluan

Daerah Istimewa Yogyakarta (DIY) sudah dikenal luas sebagai salah satu tujuan wisata bagi wisatawan domestik maupun mancanegara. Informasi mengenai objek-objek wisata yang ada di DIY beserta dengan hal-hal pendukungnya seperti sarana transportasi, fasilitas umum, dan lain sebagainya bisa dengan mudah ditemukan di internet. Dalam laman pencarian Google (google.co.id) misalnya, ketika seseorang menggunakan kata kunci "objek wisata jogja", maka dia akan menemukan puluhan laman yang membahas seluk beluk objek wisata yang ada di DIY (gambar 1).

Banyaknya sumber informasi yang ada di satu sisi akan memberikan manfaat kepada para calon wisatawan, karena antara satu laman dengan laman yang lain bisa saling melengkapi informasi yang disajikan. Namun di sisi lain informasi dengan jumlah yang sangat banyak dan beragam akan menyebabkan timbulnya *information overload.* Information overload adalah banyaknya informasi yang diterima oleh manusia sehingga sulit untuk mengolahnya. Karena adanya

information overload, manusia dituntut untuk dapat mengkombinasikan berbagai informasi yang didapatkan dari berbagai sumber sehingga menjadi satu kesatuan informasi yang utuh, akurat dan bermanfaat. Dengan banyaknya laman yang menyajikan informasi mengenai pariwisata di DIY, calon wisatawan harus menyediakan lebih banyak waktu untuk memilah dan mengakses sebanyak mungkin laman guna mendapatkan informasi yang lengkap dan akurat.

Berdasarkan permasalahan yang diuraikan di paragraf sebelumnya, maka dapat disimpulkan bahwa diperlukannya sebuah portal terintegrasi yang bisa secara otomatis mengambil informasi dari sebanyak mungkin halaman kemudian menyajikannya kepada para calon wisatawan dalam bentuk yang lebih ringkas namun lengkap dan akurat.



Gambar 1: Hasil pencarian mesin pencari Google untuk kata kunci “Objek wisata Jogja”.

Penelitian ini bertujuan untuk mengekspolarasi laman-laman yang menyajikan informasi mengenai pariwisata DIY kemudian mengumpulkan semua informasi yang ada di laman tersebut secara otomatis dengan menggunakan teknologi *web scraping*. Informasi yang sudah terkumpul kemudian akan diolah dengan menggunakan teknologi *text mining* sehingga menjadi informasi yang lebih ringkas namun lengkap dan akurat untuk disajikan kepada para calon wisatawan

Penggunaan teknik *web scraping* untuk mengambilan data dari laman web secara otomatis sudah dikenal luas. Beberapa penelitian di Indonesia yang terkait dengan implementasi web scraping diantaranya adalah peneltian dengan judul "Penerapan Teknik Web Scraping untuk Pencarian Artikel Ilmiah" (Josi dkk, 2015). Penelitian ini menggunakan teknologi web scraping untuk mencari judul-judul penelitian berdasarkan query pencarian yang diinputkan oleh pengguna. Adapun judul-judul penelitian tersebut didapatkan secara otomatis dengan teknik web scraping dari tiga laman web yaitu portal garuda, Indonesian Scientific Journal Database (ISJD), dan Google Scholar.

Penelitian kedua berjudul "Penerapan Teknik Web Scraping Untuk Pembuatan Web Service Harga Kebutuhan Pokok Nasional Pada Situs Kemendag Menggunakan Adapter Pattern Dengan Perangkat Bergerak Sebagai Klien" (Mangundiraja dkk, 2014). Penelitian ini bertujuan untuk mengimplementasikan teknik *web scraping* untuk mendapatkan data harga kebutuhan pokok nasional yang disediakan oleh situs Kemendag. Data harga yang didapatkan nantinya akan ditransformasikan kedalam format data JSON yang dapat diakses oleh berbagai perangkat khususnya perangkat bergerak melalui aplikasi klien melalui *web service* . *Web service* ini akan menyediakan data berupa harga, harga rata-rata, harga tertinggi, dan harga terendah yang dapat di-filter berdasarkan kebutuhan.

Penelilitian lain terkait dengan implementasi teknik web scraping adalah implementasi *web scraping* untuk ekstraksi halaman web hadist yang diterjemahkan kedalam bahasa Indonesia (Zaira, 2011), dan penerapan *web scraping* untuk membuat katalog buku secara otomatis (Soputri, 2015). Dari beberapa penelitian yang penulis uraikan diatas, kesemuanya hanya sebatas memanfaatkan *web scraping* untuk mengambil dan menggabungkan data dari laman-laman web yang ada kemudian menyajikannya secara mentah kepada pengguna. Dalam penelitian ini, teknik *web scraping* akan digabungkan dengan

teknik *text mining*, sehingga nantinya informasi yang dikumpulkan akan diolah terlebih dahulu untuk disajikan kedalam bentuk atau format yang lebih baik.

2. Metode Penelitian

Penelitian ini dilaksanakan dalam tiga tahapan. Tahapan yang pertama adalah menentukan laman yang akan dijadikan sebagai sumber informasi. Dalam tahapan ini, dilakukan penelusuran pada mesin pencari untuk menemukan laman-laman yang secara khusus menyajikan informasi pariwisata di Daerah Istimewa Yogyakarta. Adapun laman-laman yang menyajikan informasi pariwisata secara umum, ataupun laman web umum yang mengulas informasi pariwisata tidak menjadi objek dalam penelitian ini. Dari hasil penelusuran dengan menggunakan mesin pencari Google, diambil 5 situs yang secara spesifik menyajikan informasi pariwisata di Daerah Istimewa Yogyakarta baik secara keseluruhan maupun secara khusus di bagian tertentu dari situs tersebut. Daftar situs yang digunakan dalam penelitian ini diperlihatkan oleh tabel 1.

Tabel 1. Daftar situs pariwisata DIY

No	URL	URL Halaman Khusus*
1	https://gudeg.net	https://gudeg.net/obyek-wisata.html
2	http://wisatajawa.co.id/	http://wisatajawa.co.id/tag/wisata-jogjakarta/
3	http://pariwisata.jogjakota.go.id	http://pariwisata.jogjakota.go.id/travel
4	http://initempatwisata.com	http://www.initempatwisata.com/wisata-indonesia/jogjakarta/
5	http://visitingjogja.com	-
6	http://yogyes.com	https://www.yogyes.com/id/yogyakarta-tourism-object/

*) URL Halaman khusus adalah alamat spesifik dari suatu situs yang berisi informasi wisata DIY, apabila situs tersebut tidak membahasa wisata DIY saja.

Tahapan kedua adalah melakukan ekstraksi informasi dari laman sumber menggunakan teknik *web scraping*. *Web scraping* (Zheng, 2007) 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. Secara umum, ada 4 tahapan dalam penggunaan Web scraping untuk menggambil data secara otomatis dari sebuah laman web sebagai berikut (Turland, 2010):

1) Mempelajari dokumen HTML dari website yang akan diambil informasinya untuk tag HTML yang mengapit informasi yang akan diambil.
2) Menelusuri mekanisme navigasi pada website yang akan diambil informasinya untuk ditirukan pada aplikasi web scraper yang akan dibuat.
3) Berdasarkan informasi yang didapat pada langkah 1 dan 2 di atas, aplikasi web scraper dibuat untuk mengotomatisasi pengambilan informasi dari website yang ditentukan.
4) Informasi yang didapat dari langkah 3 disimpan dalam format data tertentu.

Tahapan yang ketiga adalah melakukan pengelompokan informasi yang didapatkan sebelumnya dari laman sumber menggunakan *text mining*. *Text mining* (penambangan teks) adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru dalam bentuk sebuah informasi, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Feldman & Sanger, 2007). *Text mining* merupakan teknik yang digunakan untuk menangani permasalahan *information extraction* atupun *information retrieval*.

Pada dasarnya proses kerja dari *text mining* banyak mengadopsi penelitian *data mining*, namun yang menjadi perbedaan adalah pola yang digunakan oleh *text mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur, sedangkan dalam data mining pola yang diambil adalah dari data yang terstruktur (Han & Kamber, 2006). Tahapan-tahapan dalam *text mining* secara umum adalah *text preprocessing* dan *feature selection* (Berry & Kogan , 2010

2.1 Text Preprocessing

Dalam melakukan *text mining,* teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses *text preprocessing. Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Dalam penelitian ini ada tiga tahapan *text preprocessing* yang dilakukan sebagai berikut:

a) Tokenisasi, yaitu tahap pemotongan string input berdasarkan kata yang menyusunnya
b) *Stopword removal*, yaitu tahapa membuang kata-kata yang tidak berpengaruh terhadap proses klasifikasi seperti kata depan, kata sambung, dan lain sebagainya.
c) *Case folding*, yaitu tahapan untuk menyeragamkan bentuk huruf menjadi huruf besar atau huruf kecil.

2.2 *Feature Selection*.

Tahapan *feature selection* merupakan tahapan penting dalam *text mining.* Salah satu fungsi penting yang disediakan oleh proses ini adalah untuk dapat memilih term atau kata apa saja yang dapat dijadikan sebagai wakil penting untuk kumpulan dokumen yang akan kita analisis.Dalam penelitian ini, metode feature selection yang digunakan adalah metode *n-gram dan term frequency.*

2.2.1 N-gram

Fitur *n-gram* digunakan dalam proses pembuatan model dengan membagi suatu kalimat menjadi beberapa bagian kata. Dalam *n-gram*, 'n' menunjukkan jumlah kata yang akan dikelompokkan menjadi satu bagian. Misalnya, apabila n=2 atau biasa disebut dengan bigram, maka sebuah kalimat akan dibagi kedalam masing-masing dua kata pada setiap bagian. Contoh pada kalimat *"Integrasi informasi pariwisata Jogja"*, maka dengan fitur bigram akan dipecah menjadi "*integrasi informasi*", "*informasi pariwisata*", dan " *pariwisata jogja*".

2.2.2 Term Frequency

Term frequency merupakan salah satu metode yang digunakan untuk melakukan perhitungan pembobotan *term.* Fitur *term frequency* dilakukan dengan menghitung frekuensi kemunculan term tertentu pada suatu dokumen.

2.3 Pengelompokan Informasi

Setelah berhasil melakukan text preprocessing dan feature selection, langkah selanjutnya adalah melakukan pengelompokan informasi. Dalam proses pengelompokan ini, terlebih dahulu ditentukan 4 kategori tempat pariwisata dan kata-kata kuncinya. Kata-kata kunci tersebut yang nantinya menjadi dasar untuk menentukan sebuah informasi/dokumen dimasukkan kedalam kategori tertentu. Proses kategorisasi dilakukan dengan mencocokkan *term frequency* dari masing-

masing dokumen dengan daftar kata-kata kunci yang sudah ditentukan. Tabel 2 berikut memperlihatkan kata kata-kata kunci untuk masing-masing kategori.

Tabel 2. Daftar kata kunci untuk setiap kategori

No	Kategori	Kata Kuci
1	Pantai	Pantai, ombak, laut, biru, pasir, kapal, nelayan
2	Alam*	Gunung, bukit, pojon, pepohonan, hijau, sungai, danau, air terjun, sawah, udara sejuk, pemandangan
3	Belanja	Belanja, barang-barang, harga, kerajinan, oleh-oleh
4	Kuliner	Makanan, enak, lezat, gudeg, manis, pedas, asam, asin, maknyus, menggugah selera, minuman, segar
5	Budaya	Budaya, seni, tari, kebudayaan, museum, sejarah, tradisi, tradisional, kearifan lokal
6	Taman	Taman, tempat bermain, rekreasi keluarga

*) Non pantai

3. Hasil dan Pembahasan

Proses awal dalam penelitian ini adalah pengambilan informasi dari laman web yang sudah ditentukan menggunakan teknik *web scraping*. Dalam menjalankan program web scraping, terlebih dahulu harus diketahui struktur HTML dari laman web untuk menentukan dalam tag HTML yang mana informasi inti direpresentasikan.

Tabel 3. Daftar tag HTML yang digunakan untuk memformat tampilan informasi inti pariwisata DIY di setiap laman.

No	URL	Tag HTML
1	https://gudeg.net/obyek-wisata.html	`<div class="widget-body"> <div class="widget-body widget-dir-home"><h4>`
2	http://wisatajawa.co.id/tag/wisata-jogjakarta/	`<div class="blog__item__info padding-top"><h2 class="blog__item__title" itemprop=`
3	http://pariwisata.jogjakota.go.id/travel	`<div class="eight columns"><h5>`
4	http://www.initempatwisata.com/wisata-indonesia/jogjakarta/	`<div class="main-box-inside "><div class="vce-loop-wrap">`
5	http://visitingjogja.com	`<div class="prl-homepage-widget"><h5 class="prl-block-title alizarin">`
6	https://www.yogyes.com/id/yogyakarta-tourism-object/	`<div class="wrapper"> <div id="main-content"> <div id="article">`

Data mengenai tag HTML yang berisi informasi inti serta jumlah informasi yang berhasil diambil dari suatu laman web diperlihatkan pada table 3. Dari 52 artikel yang berhasil didapatkan, tabel 4 memperlihatkan hasil pengelompokan yang dilakukan dengan mencocokan *term frequency* dan kata-kata kunci yang sudah ditentukan.

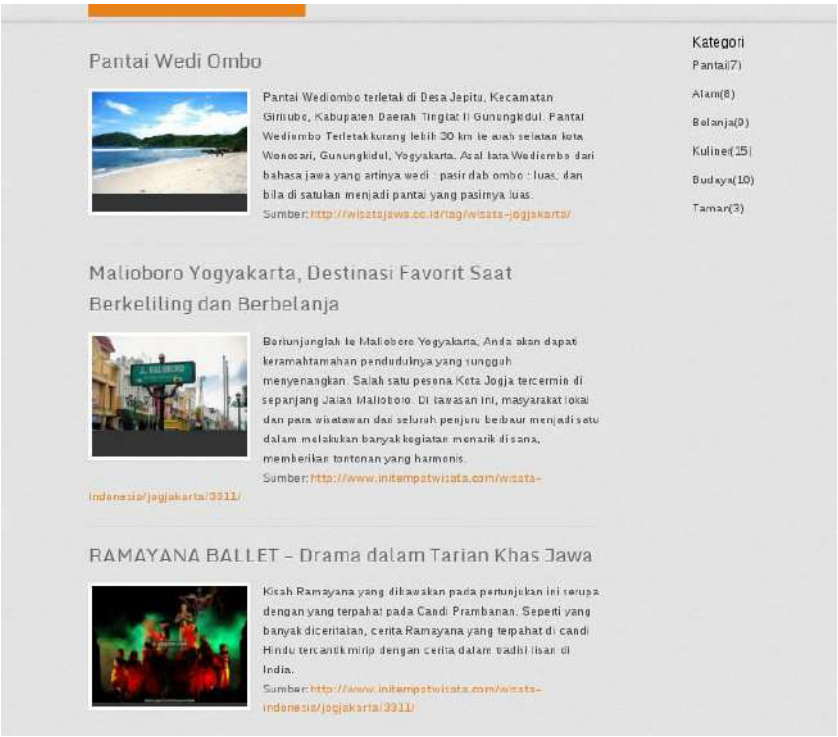
Tabel 4. Hasil pengelompokan laman web dengan text mining

No	Kategori	Jumlah Laman	Contoh Laman	
			URL	Judul
1	Pantai	7	http://wisatajawa.co.id/pantai-wediombo/#more-1175	Pantai Wediombo
2	Alam*	8	http://www.initempatwisata.com/wisata-indonesia/jogjakarta/pantai-jogan-gunungkidul-keunikan-air-terjun-yang-mempesona/3395/	Pantai Jogan Gunungkidul, Keunikan Air Terjun Yang Mempesona
3	Belanja	9	https://www.yogyes.com/id/yogyakarta-tourism-object/other/malioboro/	Malioboro Menyusuri Jalan Karangan Bunga Dan Surga Cinderamata Di Jantung Kota Jogja
4	Kuliner	15	https://www.yogyes.com/id/yogyakarta-travel-guide/8-must-try-culinary-in-kaliurang/	8 Kuliner Wajib Di Kaliurang No. 7 Bisa Bikin Kamu Ketagihan
5	Budaya	10	https://gudeg.net/direktori/63/museum-sonobudoyo-yogyakarta.html	Museum Sonobudoyo Yogyakarta
6	Taman	3	https://www.yogyes.com/id/yogyakarta-tourism-object/other/gardu-action/	Gardu Action Asyiknya Selfie Dan Pasang Aksi Memanfaatkan Barang Tak Terpakai

*) Non pantai

Bagian akhir dari penelitian ini adalah membangun sebuah laman web baru yang menyajikan informasi-informasi yang suah diperoleh dengan *web*

scraping dan dikategorisasikan dengan menggunakan *text mining*. Gambar 2 memperlihatkan tampilan depan laman web yang dibangun.



Gambar 2 Tampilan laman web yang dibangun

4. Penutup

Banyaknya laman yang menyajikan informasi mengenai pariwisata DIY dengan tingkat akurasi dan kelengkapan yang beragam menyebabkan para calon wisatawan harus meluangkan lebih banyak waktu dan tenaga untuk mengakses sebanyak mungkin laman kemudian mengkombinasikan informasi-informasi yang ada. Berdasarkan permasalahan tersebut, pada penelitian kali ini telah dikembangkan sebuah portal terintegrasi yang bisa secara otomatis mengambil informasi dari sebanyak mungkin halaman kemudian menyajikannya kepada para calon wisatawan dalam bentuk yang lebih ringkas namun lengkap dan akurat.

Hal-hal yang akan dilakukan dalam penelitian ini meliputi tiga hal. Hal pertama yang dilakukan adalah melakukan proses pengambilan informasi laman web yang ada dengan teknologi *web scraping.* Kemudian, selanjutnya informasi pariwisata yang berhasil dikumpulkan dikelompokkan kedalam beberapa kategori secara otomatis menggunakan text mining. Terakhir, dalam penelitian ini dikembangkan sebuah laman web baru untuk menyajikan informasi yang sudah dikelompokkan kedalam beberapa kategori tersebut.

Daftar Pustaka

Berry, M.W. and Kogan, J., 2010. Text Mining. *Applications and Theory. West Sussex, PO19 8SQ, UK: John Wiley & Sons*.

Feldman, R. and Sanger, J., 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*, University of Illinois at Urbana-Champaign.

Himawan, Hidayatulah, 2009, E-tourism : Antara Konsep dan Implementasi Dalam Mendukung Industri Pariwisata Indonesia, Yogyakarta, Seminar Nasional Informatika, UPN "Veteran"

Josi, A. and Abdillah, L.A., 2014. Penerapan teknik web scraping pada mesin pencari artikel ilmiah. *arXiv preprint arXiv:1410.5777*.

Mangundiraja, Nehru; Pinandito, Syahputra Aryo; Kharisma, Agi Putra, 2014. Penerapan Teknik Web Scraping Untuk Pembuatan Web Service Harga Kebutuhan Pokok Nasional Pada Situs Kemendag Menggunakan Adapter Pattern Dengan Perangkat Bergerak Sebagai Klien, Repositori Jurnal Mahasiswa PTIIK Universitas Brawijaya, Volume 3, No. 11

Soputri, S. ,2015, Implementasi Web Scraping Untuk Pencarian Data Buku Perpustakaan Studi Kasus: Perpustakaan Ukdw & Jogjalib.net. (Undergraduate thesis, Duta Wacana Christian University, 2015). Retrieved from http://sinta.ukdw.ac.id

Tanaamah, Andeka R. & Manuputty, Augie D., 2008, Kepariwisataan berbasiskan E-Tourism di Indonesia, Salatiga, Fakultas Teknologi Informasi, UKSW

Turland, M., 2010. *Guide to Web Scraping with PHP*. Marco Tabini & Associates, Inc..

Zaira, Zamrudi. 2011, Implementasi Ekstraksi Web Untuk Hadits Yang Diterjemahkan Dalam Bahasa Indonesia, Depok : Universitas Indonesia.

Zheng, S., Song, R., Wen, J. R., & Wu, D. , 2007, Joint optimization of wrapper generation and template detection. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 894-902). ACM.