Web Scraping pada Situs Wikipedia menggunakan Metode Ekspresi Regular

Mardi Siswo Utomo
Program Studi Teknik Informatika, Universitas Stikubank
email: mardiutomo@gmail.com

Abstrak

Teknik scraping dapat dilakukan teknik regular ekspresi, regular ekspresi ditentukan pola yang mengawali dan mengakhiri suatu konten utama pada halaman situs. Salah satu situs yang mempunyai berbagai macam informasi yang akan digunakan sebagai obyek scraping adalah wikipedia. Salah satu situs yang mempunyai berbagai macam informasi yang akan digunakan sebagai obyek scraping adalah wikipedia. Wikipedia sendiri adalah proyek ensiklopedia multibahasa dalam jaringan yang bebas dan terbuka.

Kemampuan dari aplikasi web scraping adalah Sistem berupa fungsi menggunakan aplikasi berbasis web digunakan untuk melakukan proses web scraping, Menggunakan CMS wordpress sebagai manajemen kontennya, Terimplementasi di dalam wordpress sebagai plugin.

Pada akhirnya untuk dapat di rangking dengan bagus oleh mesin pencari dibutuhkan konten unik, karena dengan teknik web scraping konten yang dihasilkan tidak unik maka teknik ini website tidak dapat menjadi urutan awal pada hasil mensin pencari

Kata Kunci : Web Scraping, Web Mining, Regular Ekspresi

PENDAHULUAN

Salah satu cara untuk memisahkan konten utama halaman situs dengan bagian-bagian yang tidak berhubungan dengan isi adalah dengan menggunakan teknik scraping (mardi 2012). Dengan teknik ini konten utama dari suatu halaman situs dapat diekstrak, dikoleksi dan selanjutnya dapat diproses oleh proses pengindekan .

Teknik scraping dapat dilakukan dengan berbagai cara diantaranya menggunakan analisa html DOM (document object model) dan dengan menggunakan teknik pemrograman regular ekspresi. Kedua teknik ini mempunyai keunggulan tersendiri dan menghasilkan hasil yang tidak jauh berbeda. Pada teknik DOM dibutuhkan Xquery untuk mengekstrak konten utama dari halaman situs sedangkan pada teknik regular ekspresi ditentukan pola yang mengawali dan mengakhiri suatu konten utama pada halaman situs.

Salah satu situs yang mempunyai berbagai macam informasi yang akan digunakan sebagai obyek scraping adalah wikipedia. Wikipedia sendiri adalah proyek ensiklopedia multibahasa dalam jaringan yang bebas dan terbuka, yang dijalankan oleh Wikimedia Foundation, sebuah organisasi nirlaba yang berbasis di Amerika Serikat. Nama Wikipedia berasal dari gabungan kata wiki dan encyclopedia. Wikipedia dirilis pada tahun 2001 oleh Jimmy Wales dan Larry Sanger dan kini merupakan karya referensi paling besar, cepat berkembang, dan populer di Internet. Proyek Wikipedia bertujuan untuk mengumpulkan seluruh ilmu pengetahuan manusia.

Keistimewaan Wikipedia adalah selain menyajikan informasi yang biasa ditemui di dalam sebuah ensiklopedia, Wikipedia juga memuat artikel-artikel yang biasanya ditemukan di dalam almanak, majalah spesialis, dan topik-topik berita yang masih hangat.dan banyak

orang yang memekai wikipedia ini untuk menyelesaikan tugas dan pekerjaan rumah.

METODE PENGEMBANGAN

Penelitian ini menggunakan model *prototyping* . di dalam model ini sistem dirancang dan dibangun secara bertahap dan untuk setiap tahap pengembangan dilakukan percobaan-percobaan untuk melihat apakah sistem sudah bekerja sesuai dengan yang diinginkan.

TINJAUAN PUSTAKA

1. CMS Wordpress

WordPress adalah sebuah aplikasi sumber terbuka (*open source*) yang sangat populer digunakan sebagai mesin blog (*blog engine*). WordPress dibangun dengan bahasa pemrograman PHP dan basis data (*database*) MySQL. PHP dan MySQL, keduanya merupakan perangkat lunak sumber terbuka (*open source software*). Selain sebagai blog, WordPress juga mulai digunakan sebagai sebuah CMS (*Content Management System*) karena kemampuannya untuk dimodifikasi dan disesuaikan dengan kebutuhan penggunanya. WordPress adalah penerus resmi dari b2/cafelog yang dikembangkan oleh Michel Valdrighi. Nama WordPress diusulkan oleh Christine Selleck. (http://id.wikipedia.org)

2. Web Scraping

Scraping Web (juga disebut panen Web atau Web ekstraksi data) adalah sebuah perangkat lunak komputer teknik penggalian informasi dari situs web. Biasanya, program perangkat lunak tersebut mensimulasikan eksplorasi manusia dari Web oleh salah satu rendah menerapkan-Hypertext Transfer Protocol (HTTP), atau embedding Web browser tertentu penuh, seperti Internet Explorer (IE) dan Mozilla Web browser.

Web Scraping berkaitan erat dengan pengindeksan Web, yang indeks konten Web menggunakan bot dan merupakan teknik universal yang diadopsi oleh kebanyakan mesin pencari. Sebaliknya, menggores Web lebih memfokuskan pada transformasi konten Web yang tidak terstruktur, biasanya dalam format HTML, menjadi data terstruktur yang dapat disimpan dan dianalisa dalam database lokal pusat atau spreadsheet. Web Scraping juga terkait dengan otomasi Web, yang mensimulasikan browsing Web manusia menggunakan perangkat lunak komputer. Penggunaan Web Scraping termasuk perbandingan harga online, cuaca data monitoring, deteksi situs berubah, penelitian Web, Mashup konten Web dan Web integrasi data.

ANALISA SISTEM

1. Ruang lingkup produk

Sistem ini adalah Rekayasa Perangkat Lunak Komputer berbasis web yang bertujuan untuk melakukan pengambilan isi dari konten halaman web. Hal-hal yang diharapkan oleh pengguna agar dapat diwujudkan dalam sistem ini diantaranya adalah hal-hal sebagai berikut :

a. Sistem dapat secara otomatis mengekstrak konten utama dari suatu halaman web , dalam penelitian ini digunakan halaman dokuumen pada situs http://en.wikipedia.org.

b. Untuk mempersingkat penelitian sehingga fokus pada penelitian web scraping maka digunakan CMS wordpress untuk menangani proses manajemen website.

c. Sistem dapat secara otomatis menampilkan hasil scrap dalam bentuk halaman web site atau memasukan hasil ekstrak ke dalam artikel di wordpress, baik untuk judul, isi, rangkuman maupun kata kunci / tag.

d. Sistem ini dapat diintegrasikan dengan plugin wordpress yang lain sehingga mempermudah instalasi.

e. Sistem dapat di install di web server manapun yang mendukung instalasi CMS wordpress.

Dalam pengembangan aplikasi ini diharapkan dapat memberikan manfaat sbb :

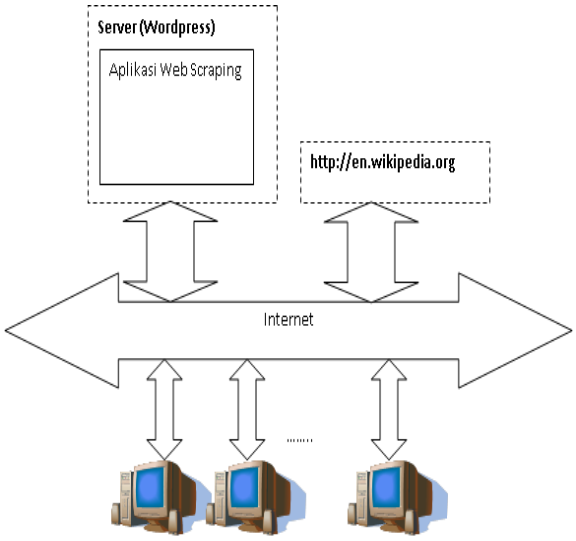
a. Sistem dapat mengekstrak konten utama pada halaman dokumen wikipedia.org

b. Sistem dapat membuat konten secara otomatis sesuai kebutuhan pada website dengan basis wordpress.

2. Perspektif produk

Aplikasi yang dibangun ditanam pada web server yang terkoneksi dengan jaringan internet. Aplikasi berjalan menggunakan service http dengan format transaksi data html, sehingga dapat dibuka menggunakan terminal yang terkoneksi ke jaringan komputer dan mampu / mempunyai browser WEB. Pengguna dapat melihat dokumen yang telah diekstrak dalam bentuk artikel dalam wordpress.

Pada gambar 1 diperlihatkan Komputer Server berfungsi sebagai web server yang teristall wordpress. Web server akan mengambil halaman web dari wikipedia.org kemudian mengkestrak konten utama dari halaman tersebut dan menyimpannya kedalam bentuk artikel di wordpress.



Gambar 1. Perspektif produk

Fungsi-fungsi Produk

Produk Aplikasi dibangun dengan antarmuka web, sehingga semua fungsi dapat langsung diakses dari halaman aktif manapun. Walaupun demikian tidak ada menu atau fungsi yang secara eksplisit merujuk ke sistem web scraping

3. Aturan Bisnis Aplikasi

Aturan bisnis digunakan sebagai acuan kemampuan dari aplikasi yang akan dibuat. aturan bisnis untuk web scraping adalah sbb:

a. Sistem berupa fungsi menggunakan aplikasi berbasis web digunakan untuk melakukan proses web scraping.

b. Menggunakan CMS wordpress sebagai manajemen kontennya

c. Terimplementasi di dalam wordpress sebagai plugin.

PERANCANGAN SISTEM

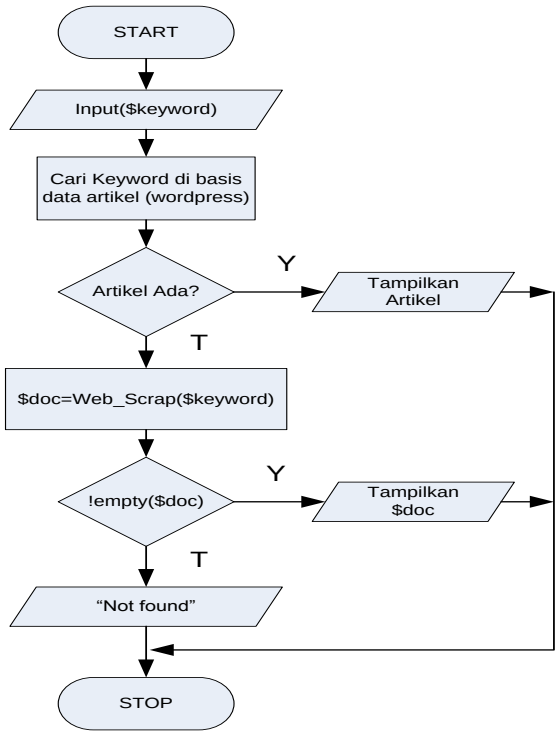
1. Diagram Alir Pencarian Artikel pada Wordpress

Aliran Proses pencarian artikel pada website berbasis wordpress dengan web scraping terinstall diperlihatkan pada gambar 2. Variabel keyword dimasukan sebagai dasar pencarian artikel, jika artikel di temukan di dalam basis data maka artikel akan di tampilkan.

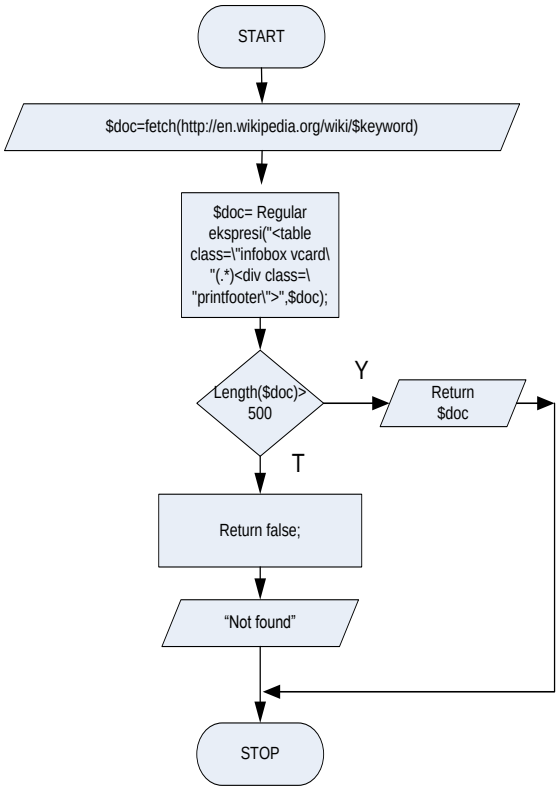
Jika artikel tidak ditemukan maka dilakukan proses web scraping untuk kata kunci yang dimasukan. Jika web scraping berhasil menemukan artikel maka akan ditampilkan ke halaman web apabila tidak maka akan ditampilkan error Not Found.

2. Diagram Alir Fungsi Web Scraping

Aliran proses web scraping ditampilkan pada gambar 3.. Fungsi akan memulai dengan mengambil halaman web pada url http://en.wikipedia.org/wiki/$keyword.
Kemudian dilakukan proses scraping pada halaman tersebut dan pabila dokumen setelah di ekstrak menghasilkan konten dengan panjang lebih dari 500 karakter maka sistem berhasil menemukan dokumen yang berhubungan dengan keyword dan mengembalikannya dalam bentuk hasil fungsi.



Gambar 2. Diagram Alir Pencarian Artikel pada Wordpress



Gambar 3. Diagram aliran proses fungsi web scraping

IMPLEMENTASI

1. Implementasi Fungsi Web Scraping Fungsi Web Scraping

Pada gambar 3 diperlihatkan alur proses dari fungsi web scraping, dimana fungsi terlebih dahulu melakukan pengambilan halaman web di website en.wikipedia.org. Proses pengambilan halaman atau biasa disebut dengan fetching dapat dilakukan dengan perintah PHP file_get_content, tetapi pada kasus en.wikipedia.org metode file_get_content tidak dapat dilakukan dikarenakan adanya leech detection di webserver wikipedia.

Metode lainnya untuk mengambil konten web adalah CURL, metode ini dapat mengirim informasi lengkap dan detil layaknya sebuah web browser sehingga web server menganggap permintaan dilakukan oleh seorang pengguna dengan menggunakan web browser. Pada penelitian ini tidak secara langsung menggunakan perintah CURL untuk mengambil halaman web, tetapi menggunakan pustaka snoopy. Pustaka snoopy berisi fungsi-fungsi PHP untuk melakukan fetch di suatu halaman web dengan perintah dasar CURL.

Pada program 1 diperlihatkan potongan kode php yang melakukan fetch, sebelum melakukan fetch terlebih dahulu dilakukan deklarasi obyek $snoopy dari kelas Snoopy, kemudian user agent string di set opera sehingga web server akan mengenali fungsi sebagai browser opera. Sebelum keyword di gunakan terlebih dahulu di filter menjadi huruf kecil semua dan mengganti karakter + menjadi karakter spasi.

Program 1. Implementasi Operasi Fetch Halaman WEB en.wikipedia.org

```
$snoopy = new Snoopy;
$snoopy->agent='opera';
$keyword=strtolower($keyword);
$keyw=str_replace("  "," ",$keyword);
$keyw=str_replace(" ","+",$keyword);
$uristr="http://en.wikipedia.org/w/index.php?titl
e=Special%3ASearch&search=$keyword";
```

```
$snoopy->fetch($uristr);

$striklan= $snoopy->results;
```

Pada program 2 . Konten utama dibatasi dari header oleh teks ”<!-- bodytext -->” dan dibatasi oleh footer oleh ”<!-- /bodytext -->”, sehingga pada program 3 digunakan perintah regular ekspresi dengan mendeteksi pola <!-- bodytext -->(.*)<!-- /bodytext -->.

Program 2 Pola Konten utama halaman dokumen wikipedia

```
...bagian header

<!-- bodytext -->

.....

Konten Utama

.....

<!-- /bodytext -->

....bagian footer
```

Pada program 3 diperlihatkan operasi web scraping dengan menggunakan metode regular ekspresi. Konten utama halaman web hasil en.wikipedia.com diapit oleh bagian header dan footer seperti terlihat

Program 3. Perintah ekspresi regular untuk memisahkan konten utama wikipedia

```
$str= $snoopy->results;

ereg(”<!-- bodytext -->(.*)<!-- /bodytext -->", $str ,$match);

$str=$match[0];
```

Implementasi Wordpress Plugin

Setelah fungsi selesai ditulis maka untuk mempermudah penggunaan dan integrasi dengan wordpress maka struktur program fungsi web scraping di rubah menjadi struktur plugin pada wordpress.

Struktur program plugin pada wordpress mengharuskan ditambahkannya header remark yang berfungsi untuk memuat informasi seputar plugin tersebut, seperti terlihat pada potongan program 4.

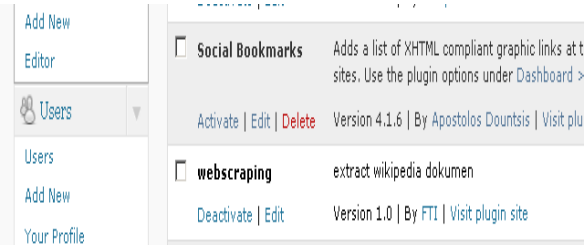
Program 4. Header plugin wordpress web scraping

```
/*

Plugin Name: Wikipedia web scraping

Plugin URI: http://unisbank.ac.id

Description: extract wikipedia dokumen

Version: 1.0

Author: FTI

Author URI: http://unisbank.ac.id

*/
```

Instalasi Plugin Web Scraping

Plugin pada wordpress diinstall melalui menu administrator di url http://[namadomain]/wp-admin/ setelah terlebih dahulu memasukan username dan password untuk administrator. Plugin dipasang pada menu plugins->add new , plugin dapat di upload ataupun langsung di download dari repository wordpress.

Pada penelitian ini plugin tidak terdapat di repository wordpress, sehingga digunakan menu upload untuk menambahkan plugin. Plugin akan aktif setelah diaktifkan dengan meng klik url activate dibawah nama plugin. Pada gambar 4 diperlihatkan tampilan layar daftar plugin yang terpasang di wordpress.



Gambar 4. Tampilan layar daftar plugin wordpress

Aktifasi Fungsi Web Scraping

Setelah plugin terpasang tidak serta merta fungsi webscraping telah digunakan pada wordpress. Untuk dapat berfungsi seperti proses yang digambarkan pada gambar 2 maka dibutuhkan modifikasi pada bagian template / theme yang digunakan. Pada gambar 2

diperlihatkan bahwa proses yang di sisipi fungsi web scraping adalah proses pencarian.

Pada wordpress proses pencarian melibatkan file search.php pada theme yang aktif untuk menghasilkan halaman konten. File search.php pada bagian tampilan ”Not Found” di ganti dengan pemanggilan fungsi web scraping sehingga sebelum menampilkan not found terlebih dahulu dilakukan pencarian pada situs en.wikipedia.org, seperti terlihat pada program 5

Program 5. Perubahan pada search.php

```
<?php if ( have_posts() ) { ?>

<h1 class="page-title"><?php printf( __( 'Search Results for: %s', 'twentyten' ), '<span>' . get_search_query() . '</span>' ); ?></h1>

<?php } else {

global $s;

include "wp-includes/class-snoopy.php";

if((function_exists('webscrap')) and (!empty($s))) $doc=webscrap($s);

if(strlen($doc)>500) echo "<div id='post-0' class='post no-results not-found'>
                                        <h2 class='entry-title'>$s</h2>
                                        <div class='entry-content'>

        <p>$doc</p>

        </div><!-- .entry-content -->
                                    </div><!-- #post-0 -->";


else {  ?>

 <div id="post-0" class="post no-results not-found">

<h2 class="entry-title"><?php _e( 'Nothing Found', 'twentyten' ); ?></h2>
                                        <div class="entry-content">

        <p><?php _e( 'Sorry, but nothing matched your search criteria. Please try again with some different keywords.', 'twentyten' ); ?></p>


        <?php get_search_form(); ?>


        </div><!-- .entry-content -->
                                    </div><!-- #post-0 -->


<?php  } } ?>
```

HASIL DAN PEMBAHASAN

Google.com trend

Pengujian pertama dilakukan dengan memasukan kata kunci-kata kunci yang terdapat dalam trend pencarian google.com. Seperti pada tabel 1 adalah daftar kata-kata kunci pencarian yang populer pada tanggal 8 Februari 2011.

Tabel 1. Hasil pengujian dengan kata kunci dari trend google.com

No	Kata Kunci	Hasil
1	aol huffington post	Search Result
2	sting wwe	Search Result
3	sean payton	Dokumen
4	gizmodo	Dokumen
5	mike munchak	Dokumen
6	chicago code	Dokumen
7	reyes	Search Result
8	groupon	Dokumen
9	aguilera	Search Result
10	national anthem	Dokumen

Kata-kata kunci dari google trend tersebut sebagai masukan pada sistem web scraping. Kemudian pada kolom hasil di perlihatkan apakah web scraping berhasil memberikan keluaran berupa halaman web. Pada tabel juga diperlihatkan apakah hasil dari web scraping

menghasilkan dokumen atau halaman hasil cari. Dari tabel 4.1 diperlihatkan 6 dari 10 kata kunci populer pada tanggal 8 februari 2011 dapat diproduksi halaman situsnya dengan menggunakan teknik web scraping.

Google.co.uk trend

Pengujian kedua dilakukan dengan memasukan kata kunci-kata kunci yang terdapat dalam trend pencarian google.co.uk. Seperti pada tabel 2 adalah daftar kata-kata kunci pencarian yang populer pada tanggal 8 Februari 2011.

Tabel 2. Hasil pengujian dengan kata kunci dari trend google.com

No	Kata Kunci	Hasil
1	paul getty	Dokumen
2	sting wwe	Search Result
3	gma news	Dokumen
4	sean payton	Dokumen
5	huffington post aol	Search Result
6	darth vader commercial	Search Result
7	mike munchak	Dokumen
8	reyes	Search Result
9	chicago code	Dokumen
10	groupon	Dokumen

Kata-kata kunci dari google trend tersebut sebagai masukan pada sistem web scraping. Kemudian pada kolom hasil di perlihatkan apakah web scraping berhasil memberikan keluaran berupa halaman web. Pada tabel juga diperlihatkan apakah hasil dari web scraping menghasilkan dokumen atau halaman hasil cari. Dari tabel 4.2 diperlihatkan 6 dari 10 kata kunci populer pada tanggal 8 februari 2011 dapat diproduksi halaman situsnya dengan menggunakan teknik web scraping.

KESIMPULAN

Teknik web scraping dengan metode regular ekspresi dapat di implementasikan on the fly sehingga hanya dilakukan saat seorang pengguna membutuhkan dan dokumen yang dimaksud tidak terdapat didalam website Dari tren pencarian google.com dan google.co.uk pada tanggal 8 februari 2011, dihasilkan 60% halaman dokumen yang berhubungan dengan kata-kata kunci pada trend pencarian google.

Untuk dapat di rangking dengan bagus oleh mesin pencari dibutuhkan konten unik, karena dengan teknik web scraping konten yang dihasilkan tidak unik maka teknik ini website tidak dapat menjadi urutan awal pada hasil mensin pencari.

Untuk menghasilkan artikel yang unik perlu digabungkan dengan teknik text sumarization ataupun random sentences producer.

Masalah hukum dan HAKI menjadi masalah utama pada situs yang menggunakan teknik web scraping ini. Tetapi dapat di tambahkan pada footer sumber asli dari dokumen tersebut.

DAFTAR PUSTAKA

Google Trends. (n.d.). *Google*. http://google.com/trends. Diakses tanggal 8 Februari 2013

Google Trends - Hot Searches. (n.d.). *Google*. http://google.co.uk/trends. Diakses tanggal 8 Februari 2013

Hadiono K, 2010, *Aplikasi Web Scrapping Untuk Koleksi Konten Utama Halaman Situs*, Unisbank.

JISC Briefing Paper, (2006), *Text mining*, JISC, Inggris

Mardi Siswo Utomo, (2012). *Implementasi PHP Sebagai Penghasil Konten Otomatis Pada Halaman Situs*,Dinamik,Unisbank

Murhadin, Endy, (2003). *PHP Programming Fundamental dan MySQL Fundamental*, http://ikc.cbn.net.id/umum/endy-php.php

Nugroho, Bunafit, (2004). *PHP & MySQL Dengan Editor Dreamweaver MX*, Andi, Yogyakarta

Pressman R, (1997). *Software Engineering*, Mc Graw Hill, USA

Prothelon's, (2005). *Web Desain, PHP Programming, Language Learning*, http://prothelon.com/mambo/tutorial

Wikipedia, (n.d.). Wikipedia - Wikipedia, the free encyclopedia. *Wikipedia, the free encyclopedia*. http://en.wikipedia.org/wiki/wikipedia. Diakses tanggal 2 januari 2013

Wikipedia. (n.d.). Wikipedia, the free encyclopedia. *Wikipedia, the free encyclopedia*. http://en.wikipedia.org/. Diakses tanggal 2 januari 2013

WordPress › Blog Tool, Publishing Platform, and CMS. (n.d.). *WordPress › Blog Tool, Publishing Platform, and CMS*. http://wordpress.org. Diakses tanggal 2 januari 2013

World Wide Web Consortium (W3C). (n.d.). *World Wide Web Consortium (W3C)*. http://www.w3.org. Diakses tanggal 8 maret 2013