Pemanfaatan Topic-Focused Crawler untuk Pembangunan Corpus Berita Bencana menggunakan Teknik Scrapy CSS Selector

Muhammad Syaifur Rohman*[1], Heru Agus Santoso[2], Galuh Wilujeng Saraswati[3], Nurul Anisa Sri Winarsih[4]
[1,2,3,4]Teknik Informatika, Universitas Dian Nuswantoro;
Jalan Imam Bonjol No. 207, Kota Semarang, Jawa Tengah 50131
e-mail: *[1]syaifur@dsn.dinus.ac.id, [2]heru.agus.santoso@dsn.dinus.ac.id,
[3]galuhwilujengs@dsn.dinus.ac.id, [4]nurulanisasw@dsn.dinus.ac.id

Abstrak

BPBD provinsi Jawa Tengah adalah suatu Lembaga Pemerintah non-departemen yang melaksanakan tugas penanggulangan bencana di Provinsi maupun Kabupaten/ Kota. Dalam melaksanakan tugasnya, BPBD berusaha memberikan beberapa layanan informasi kebencanaan melalui sebuah website, akan tetapi banyak masyarakat yang belum banyak berkunjung didalamnya. Oleh karena itu perlu di bangun sebuah aplikasi berbasis android. Untuk membuat aplikasi BPBD provinsi Jawa Tengah berbasis android ini memerlukan akses informasi dan data dari website BPBD pusat, sedangkan akses untuk menuju databasenya harus menghubungkan ke dalam server KOMNIFO. Dengan mempertimbangankan masalah keamanan, peneliti menggunakan crawling untuk membuat corpus berita kebencanaan. Teknik pengembangan software yang digunakan yaitu waterfall dengan proses: perencanaan, analisis, perancangan, implementasi dan pengujian. Peneliti membangun Topic-Focused Crawler untuk Pembangunan Corpus Berita Bencana Menggunakan Teknik Scrapy CSS Selector menggunakan bahasa pemrograman python dan database mysql.

Kata kunci—Topic-Focused Crawler, *Scrapy CSS, BPBD, php*

Abstract

Central Java provincial BPBD is a non-departmental Government Institution that carries out disaster management tasks in the Province and Regency / City. In carrying out its duties, BPBD tries to provide several disaster information services through a website, but many people have not visited much in it. Therefore it is necessary to build an Android-based application. To create an Android-based Central Java province BPBD application requires access to information and data from the central BPBD website, while access to its database must connect to the KOMNIFO server. By considering security issues, researchers use crawling to create a corpus of disaster news. The software development technique used is the waterfall with the process: planning, analysis, design, implementation and testing. Researchers build Topic-Focused Crawlers for Corpus News Development Disasters Using the Scrapy CSS Selector Technique using the python programming language and the mysql database.

Keywords— Topic-Focused Crawler, *Scrapy CSS, BPBD, php*

1. PENDAHULUAN

Seiring dengan meningkatnya pertumbuhan informasi di internet mendorong manusia untuk mengembangkan inovasi baru dalam mengolah sumber informasi tersebut sehingga dapat menghasilkan informasi baru [1]. Bagi Lembaga Pemerintahan seperti BPBD (Badan Penaggulangan Bencana Derah) Provinsi Jawa Tengah, website merupakan salah satu media yang digunakan untuk menampilkan informasi terkait dengan kebencanaan [2]. BPBD provinsi Jawa Tengah adalah suatu Lembaga Pemerintah non-departemen yang melaksanakan tugas penanggulangan bencana di Provinsi maupun Kabupaten/Kota. Informasi kebencanaan ini diperlukan oleh masyarakat agar selalu waspada dan persiapan proses mitigasi bencana.

Di Era industri 4.0 ini smartphone menjadi kebutuhan wajib seluruh penduduk didunia ditandai dengan banyaknya pengguna smartphone terutama berbasis Android [3]. Kondisi ini mendorong peniltian ini mengembangkan sebuah aplikasi berbasis android untuk meberikan informasi terkini mengenai berita kebencanaan yang bisa diikuti oleh seluruh masyarakat khususnya di provinsi Jawa Tengah apalagi aplikasi mobile hanya perlu diinstal tanpa banyak langkah diperlukan dibanding dengan platform lain. Sehingga masyarakat bisa melihat daftar berita bencana yang sedang terjadi di sekitar lingkunganya secara langsung.

Dengan mempertimbangankan masalah keamanan, peneliti menggunakan crawling untuk membuat corpus berita kebencanaan sehingga akses langsung ke server utama bisa dihindari. Oleh karena itu, Aplikasi BPBD provinsi Jawa Tengah ini melakukan *Topic Focused Crawler* menggunakan web scrapping berbasis CSS selector.

Metode *crawling* ini digunakan untuk mengumpulkan data dan informasi yang di lakukan secara otomatis melalui penyaringn ke semua halaman internet untuk membuat index dari infomasi yang dicarinya. Metode ini digunakan untuk mengumpulkan informasi sebanyak mungkin dari halaman web, mengubungankan struktur web dan mengunduh halaman secara otomatis [4]. Pada saat ini, terdapat metode crawling yang berfokus pada sebuah topik (*topic-focused crawler*) diciptakan untuk crawling bersasarkan tema yang diinginkan oleh pengguna [2][5]. Crawler ini berfokus pada topik tidak peduli dalam mengejar cakupan halaman yang tinggi tetapi mengakses halaman yang berhubungan dengan subjek secara selektif dan karenanya *topic-focused crawler* memiliki kelebihan seperti penggunaan sumber daya yang rendah, mudah melakukan pembaruan basis data indeks, memiliki kemampuan untuk memilih informasi web melalui blok topik dan mengumpulkan bagian-bagian berbeda dari hasil yang dikumpulkan untuk meningkatkan pemanfaatan seluruh halaman web.

Terdapat beberapa perangkat lunak yang digunakan untuk crawling website anatar lain: DFtoExcelOnLine, Zamzar, CometDocs, PDFTables, Tabula, import.io, kimonolab, myTrama, Mozenda, QuBole, ScraperWiki, Apache Nutch, dan Scrapy [4][6]. Di antara perangkat lunak itu, peneliti memilih Scrapy sebagai alat utama fokus penelitian dalam mengekstraksi data. Ada beberapa alasan untuk memilih framework ini, pertama karena *scrapy* besifat *open source* sehingga dapat dengan mudah digunakan oleh siapapun[1]. Kedua scrapy bersifat *fleksibel* sehingga hanya diperlukan sedikit adaptasi jika ingin menjelalajahi beberapa situs yang berbeda[7]. Ketiga, Platform ini memiliki mekasisme mengekstrak data dan informasi dari dokumen berbentuk HTML dengan memilih bagian tertentu dari bagian dokumen HTML [8]. Selain itu untuk karena platform ini menggunakan bahasa pemrograman python[9], sehingga programmer pemula bisa menyesuaikan pemrograman kedalam bentuk yang lebih sederhana. Fitur ini tentu akan membantu pengembang untuk mengembangkan dan memelihara program crawlingnya menjadi lebih tepat sararan.

Beberapa penelitian sebelumnya terkait dengan *web scraping* dapat digunakan didalam beberapa objek penelitian anatar lain: Pembuatan *framework* sistem penemuan informasi kembali tentang petroleum menggnakan web scrapy menggunakan media *phyton*. dan pembuatan rancangan topic-focused crawler dengan megimplenetasikan scrapy[2]. Selain itu algoritnma ini memiliki kemampuan untuk peningkatan search engine dengan mengkombinasikan metode ini dengan proses klasifikasi *Naïve Bayes*[6], AWeb Scraping Methodology for Bypassing Twitter API Restrictions [8], Penerapan *web scraping* pada website *company Profile* [10]. Pemanfaatan *crawling* berita dalam bahasa indonesia untuk pembutaan *corpus* berita menggunakan *scrapy* dan *Xpath* [11].

Berdasarkan latar belakang diatas, penulis merancang dan mengimplementasikan web scraping pada aplikasi BPBD Provinsi Jawa Tengah. Bahasa pemrograman phyton dan database MYSQL dipilih penulis untuk media penyimpanan data dan *framework scrapy* digunakan untuk update berita kebencanaan sehingga aplikasi yang dibuat peneliti mempunyai update berita yang sama dengan website utama BPBD provinsi Jawa Tengah.

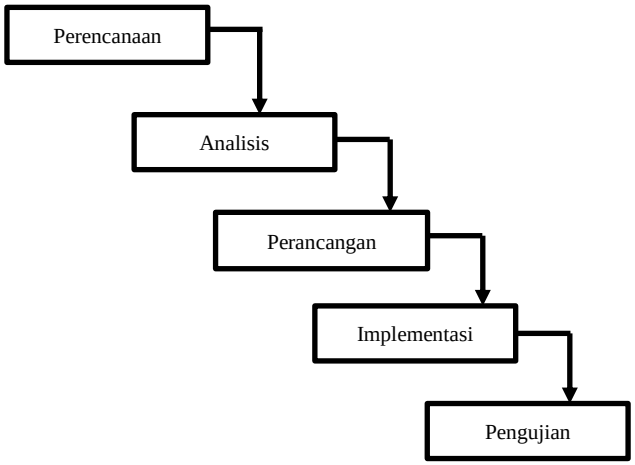
2. METODE PENELITIAN

2.1 Objek Penelitian

Objek penelitian ini adalah sebuah website BPBD Provinsi Jawa Tengah yang bernama (http://bpbdjateng.com). Website ini merupakan sarana dari BPBD dalam memberikan pelayanan informasi kebencanaan ke seluruh masyarakat Jawa Tengah. Ada beberapa konten dalam website yang bisa digunakan untuk focused Crawling Scrapy yaitu: berita kebencanaan, PPID, dan artikel kegiatan. Penelitian ini hanya mengekstrak artikel/berita yang ada dalam kategori tersebut.

2.2 Metode Pengembangan

Teknik pengembangan software waterfall digunakan untuk membangun Topic-Focused Crawler Untuk Pembangunan Corpus Berita Bencana Menggunakan Teknik Scrapy CSS Selector.



Gambar 1. Metode Pembangunan

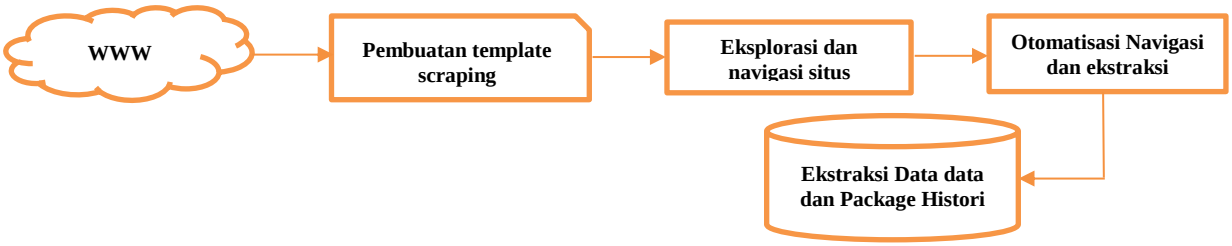
Gambar 1 menunjukan tahap -tahapan dalam membangun aplikasi yang akan digunakan pada penelitian ini. Metede tersebut meliputi beberapa tahapan diantaranya: (1) Perencanaan, (2) Analisis, (3) Perancangan, (4) Implementasi, dan (5) Pengujian

2.1 Tahap perencanaan

Pada Tahap ini, peneliti menentukan objek penelitian yang akan diimplementasikan teknik web Crawling Scrapy. Menentukan jumlah kategori berita yang akan yang akan diekstraksi menjadi basis Data aplikasi BPBD Provinsi Jawa Tengah.

2.2 Analisis Kebutuhan

Dalam tahapan ini peneliti melakukan Analisa kebutuan yang di perlukan untuk membangun aplikasi ini. Kegiatan analisa ini mencakup 3 aktivitas [10] yang terdiri dari analisia pembuatan web scraping menggunakan scrapy CSS selector sesuai dengan tema case study, Analisa spesifikasi aplikasi yang akan dibangun dan analisa aritekstur topic focused crawler menggunakan scrapy CSS selector. Scrapy melakukan crawling websire BPBD Jawa Tengah dengan menggunakan fitur selector CSS. Fitur ini dipilih karena objek penelitian ini (website BPBD Jawa Tengah), merupakan sebuah website yang hampir semua node yang tersedia di halaman tersebut menggunakan style CSS yang sama, sehingga artikel/informasi yang dihasilkan dari proses penyeleksian relative lebih spasifik. Gambar 2 menjelaskan aritekstur topic focused crawler menggunakan scrapy CSS selector.



Gambar 2. Ilustrasi cara kerja Web Scraping

Berdasarkan Gambar 2, proses pembutan penelitian ini di mulai dari membuat template scraping yang digunakan untuk mempelajari cara mengambil informasi dari dokumen HTML website dan menandai HTML tersebut. Tahap kedua adalah dengan mengeksplorasi navigasi situs yang dikenai scrapping. Tahapan terakhir adalah proses ekstraksi informasi yang dilakukan secara otomatis, tujuanya untuk mengambil informasi yang dibutuhkan secara otomatis dan disimpan ke dalam basis data.

2.3 Perancangan Basis Data

Pada tahap ini, penulis berfokus pada perancangan basis data untuk menampung ekstraksi berita atau artkel yang akan dilakukan scraping. Dalam pembuatan aplikasi BPBD Provinsi Jawa Tengah memerlukan tabel berita yang berasal dari website (http://bpbdjateng.com). Tabel ini terdiri dari 12 kolom. Field pertama berisi id categoty_id, title, slug, content, image, status, date, featured, created_at, update at dan delete_at.

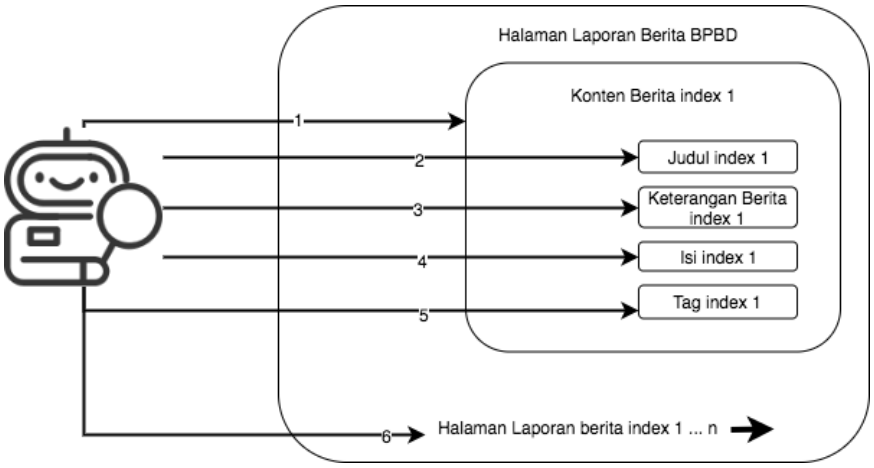
2.4 Implementasi

Implmentasi teknik scraping kedalam bahasa pemrograman PHP merupakan langkah keempat dalam tahapan ini. Berdasarkan penelitian [10][11] menyatakan bahwa bahasa pemrograman PHP dipilih karena bersifat dinamis karena sebagai webase sselain itu bersifat open source dan bisa di jalankan oleh semua sistem operasi. Sehingga dengan adanya kemudahan yang di miliki bahasa PHP ini akan membantu dalam pembuatan corpus berita bencana.

Pada tahap implementasi yang menggunakan bahasa pemrograman PHP, metode scrapying ini akan secara otomatis mengambil konten website. Untuk mengecek proses crawling, robot scrapper harus dibuat untuk mengambil konten-konten didalam website secara otomatis tanpa harus menunjuk satu persatu konten yang harus di ambil.

Pada tahap pertama (1) sesuai dengan gambar 3, robot harus mengetahui bagaimana ciri-ciri konten artikel berita yang bisa dicrawling. Karena bisa jadi dalam satu halaman web ada banyak sekali konten selain konten artikel berita. Dengan menggunakan CSS Selector, ciri konten artikel berita adalah yang mengandung CSS "div.post". Pada tahap kedua setelah berhasil mendefinisikan dimana CSS yang mengandung konten artikel berita, maka hal yang bisa dilakukan adalah mengambil seluruh konten didalamnya yang digolongkan menjadi beberapa bagian seperti judul, keterangan artikel, isi, dan tag. Pada Judul diambil dari CSS "h2.post-title" pada keterangan (2), keterangan dari CSS "p.postmetadataw" pada keterangan (3), pada isi berita diambil dari CSS "div.entrytext" pada keterangan (3) lalu yang terakhir tag diambil dari CSS "div.up-bottom-border" pada sub "p.postmetadata" ditunjukan dari keterangan (5). Pada tahap ketiga, robot akan mencari artikel lain dalam halaman tersebut yang mengandung CSS "div.post" sampai dengan tidak tersisa lagi yang mengandung "div.post". Selanjutnya setelah habis pada tahap ke empat robot akan menemui kondisi jika "div.post" habis maka akan ke halaman selanjutnya melalui CSS "div.alignright".

Dan yang terakhir tahap kelima, robot akan menuju kehalaman halaman selanjutnya, menemui "div.post" dan mengambil seluruh isi dari artikel berita bencana BPBD Jawa Tengah dengan keterangan judul, keterangan, isi dan tag. Lalu jika sudah tidak ada CSS "div.alignright" atau dalam kata lain sudah menemui halaman terakhir maka robot akan otomatis berhenti mengambil data.



Gambar 3. Cara Kerja Robot Scaper

Selanjutnya seluruh hasil data scrapping yang dilakukan oleh robot akan disimpan kedalam sebuah array dan bisa siap untuk dikirmkan ke database yang sudah ditentukan untuk di akses melalui aplikasi android.

2.5 Pengujian (Testing)

Pada tahap terkahir ini, web scraping yang telah diimplemetasikan pada halaman update news pada situs http://bpbdjateng.com dilakukan pengujian. Pengujian yang dilakukan dalam bentuk blackbox pengecekan input output dalam setiap modul testing untuk mencari kesalahan coding dan kesalahan logika dengan pendekatan [10]. Tabel 1 menampilkan pengujian aplikasi

webscraping dengan pendekatan top-down aplikasi dibagi menjadi tujuh modul utama. Apabila modul yang telah diuji berjalan dengan baik dan sesuai dengan fungsinya maka menghasilkan kesimpulan valid.

Tabel 1. Pengujian Aplikasi

No	Skenario pengujian	Output pengujian	Hasil
1	Melakukan update berita kebencanaan	Berhasil melakukan pengelolaan update artikel secara otomatis	Berhasil
2	Melakukan setting lokasi bencana sesuai dengan pilihan	Berhasil melakukan setting lokasi	Berhasil
3	Melakukan pengelolaan file manager	Berhasil melakukan pengelolaan file manager	Berhasil
4	Melakukan pengelolaan Backup Manager	Berhasil melakukan pengelolaan backup manager	Behasil
5	Melakukan pengelolaan web setting	Berhasil melakukan Web Setting	Behasil
6	Melakukan pengiriman berita notifikasi secara realtime	Berhasil mengirimkan berita notifikasi secraa realtime	Berhasil
7	Melakukan pengelolaan User & Role manager	Berhasil melakukan pengelolaan user & role manager	Behasil

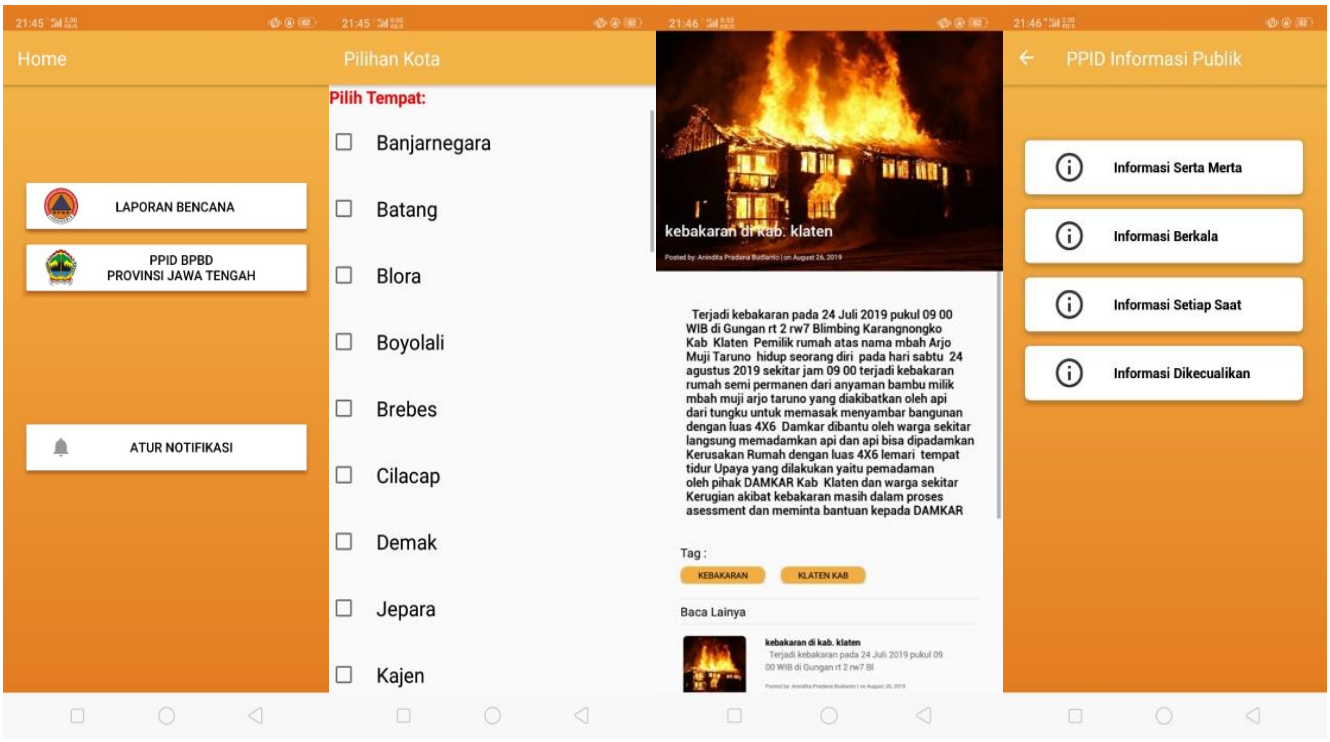
Tahapan pengujian diterapkan pada semua modul menghasilkan valid dan dapat berjalan sesuai fungsinya, Maka pembuatan corpus berita BPBD memalui teknik crawling dikatan berjalan dengan sempurna. Aplikasi ini mempunyai output berupa dapat mengirimkan notifikasi yang terkirim secara otomatis ke smartphone sesuai dengan lokasi/ wilayah yang telah dipilih.

3. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah sebuah Corpus Berita Bencana yang akan digunakan dalam pembuatan aplikasi BPBD provinsi Jawa Tengah yang dapat di unduh pada google play dengan nama BPBD Jawa Tengah. Aplikasi ini mampu memberikan update berita terkini mengenai bencana yang terjadi dilokasi sekitar Anda, sesuai pemilihan lokasi yang telah pengguna tentukan pada saat login aplikasi. Hasil scrapy tersebut akan di tampilkan sebagai news feed pada notifikasi aplikasi dengan melakukan penyimpanan *historynews* yang diambil sehingga jika tedapat sebuah berita yang pernah ditampilkan akan tersimoan dalam *history*.

3.1 Home Page

Pada saat aplikasi BPBD Jawa Tengah dijalankan maka pertama kali pengguna akan masuk pada Home Page. Saat aplikasi ini mulai dijalankan, proses web scraping juga mulai dioperasikan secara otomatis. Scraping yang dijalankan otomatis ini akan dijalankan dengan cron-job sehingga data akan diupdate dengan waktu tertentu. Lalu data yang belum ada akan masuk sebagai record yang baru. Seluruh record data hasil dari scrapping akan bisa diakses oleh android. Berikut ini tampilan Home Page.



Gambar 4. Tampilan Aplikasi BPBD Provnsi Jawa Tengah

Tampilan Home aplikasi BPBD Jawa Tengah berbasis android ini memiliki 3 fungsi menu utama anatra lain: Laporan bencana, PPID BPBD Provinsi Jawa tengah, dan Atur Notifikasi. Menu Laporan bencana menyediakan informasi kepada pengguna mengenai berita kebencanaan yang terjadi saat itu juga dan bersifat realtime. Laporan bencana ini akan terhubung dengan menu atur notifikasi karna pengiriman data notifikasi akan disesuaikan dengan subscribe pilihan lokasi yang telah di inputkan oleh pengguna. Untuk tampilan Menu PPID BPBD Jawa Tengah sudah dikelompokan berdasarkan 4 kategori yaitu: informasi Serta Merta, informasi berkala, informasi setiap saat dan informasi dikecualikan.

3.2 Hasil Update Berita dengan Teknik *Scraping*

Seluruh list berita yang ada pada website pusat (http://bpbdjateng.com) akan secara otomatis dapat terlihat juga pada aplikasi android BPBD Provinsi Jawa Tengah. Admin website (http://bpbdjateng.com) tidak perlu melakukan update data secara manual ke dalam aplikasi ini karena semua update berita yang diada pada werbite akan secara otomatis terhubung dengan aplukasi androidnya.

3.3 Lagalitas WebScraping

Data scraping mengambil data dari screen outputs atau mengekstrak data dari kode HTML. Aplikasi hanya mengambil yang disediakan / dihasilkan oleh interface website yang di-scrape. Scrapes yang dilakukan pada peneltian ini bersifat “mutual benefit” yang dapat membantu proses promosi satu sama lain dalam regional yang berbeda.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil implementasi webscraping pada website BPBD Provinsi Jawa tengah dapat ditarik kesimpulan sebagai berikut:

1) Website BPBD Provinsi Jawa Tengah ini yang dihasilkan dengan menerapkan teknik web scraping menggnakan teknik CSS selector berhasil mengekstrak informasi yang di simpan dalam bentuk news list.
2) Website berhasil menyimpan otomatis data hasil scraping dan history pada database.
3) Website melakukan web scraping secara otomatis sesuai penjadwalan yang dihasilkan crontab atau task scheduler selama task scheduler tersebut aktif.
4) Dengan adanya task ini pada website BPBD Provinsi Jawa Tengah ini, memudahkan pembuatan corpus berita pada aplikasi Anroid BPBD Jawa Tengah dan distribusi informasi secara otomatis dan berkesinambungan pada website yang melakukan scraping.
5) Web scraping merupakan tindakan legal karena pada penelitian ini data diambil dari pihak yang sudah memberikan izin demi tujuan yang lain yakni meningkatkan kinerja BPBD dalam memberikan persebaran informasi kebencanaan dan menudahkan masyarakat dalam memperoleh berita kebencaan. Web scraping di sini tidak melakukan pencurian data, tidak manipulasi informasi, dan sebagainya. Bahkan web scraping dapat memberikan mutual simbiosme dengan meningkatnya trafik atas sumber asli link di-scraping.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada BPBD Jawa Tengah yang telah memberi dukungan terhadap penelitian ini berupa kesempatan dalam bentuk kerja sama dengan UDINUS Program Studi Teknik Informatika.

DAFTAR PUSTAKA

[1] T. Arierf and R. Hemawan, "Perbandingan Metode Web Scraping Menggunakan Css Selector Dan Xpath Selector Taufiq," *J. Ilm. ILMU Komput. Univ. Udayana Vol.*, vol. X, no. 2, pp. 32–38, 2017.

[2] D. Xie and W. Xia, "Design and Implementation of The Topic-focused Crawler Based on Scrapy," vol. 851, pp. 487–490, 2014.

[3] P. Raulamo-jurvanen, K. Kakkonen, and M. Mäntylä, "Using Surveys and Web-Scraping to Select Tools for Software Testing Consultancy."

[4] A. S. Nisafani, R. A. Hendrawan, and A. Wibisono, "Eliciting Data From Website Using Scrapy : An Example," no. or Ixml, pp. 7–12, 2017.

[5] C. Kim, S. Park, and Y. Kim, "Design and implementation of crawling algorithm to collect deep web information for web archiving," 2018.

[6] A. D. Supriatna, "Web Scraping and Naïve Bayes Classification for Job Search Engine Web Scraping and Naïve Bayes Classification for Job Search Engine," 2018.

[7] Y. Fan, "Design and Implementation of Distributed Crawler System Based on Scrapy Design and Implementation of Distributed Crawler System Based on Scrapy," 2018.

[8] I. P. Nacional and E. Culhuacan, "Research Article A Web Scraping Methodology for Bypassing Twitter API," pp. 1–7.

[9] Y. Ren, "A Framework of Petroleum Information Retrieval System Based On Web Scraping With Python," *2018 15th Int. Conf. Serv. Syst. Serv. Manag.*, pp. 1–6, 2018.

[10] S. Tinggi, I. Komputer, and D. Bangsa, "Penerapan Web Scraping Pada Websitecompany Profile," pp. 37–43.

[11] T. Rizaldi and H. Ariefputranto, "Pemanfaatan News Crawling Untuk Pembangunan Corpus Berita Menggunakan Scrapy dan Xpath," pp. 291–295, 2017.