Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 21, No. 1, November 2021, pp. 229∼236
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v21i1.1081

Ekstraksi Informasi Destinasi Wisata Populer Jawa Timur Menggunakan Depth-First Crawling

Information Extraction of the Popular Tourism Sites of East Java Depth-First Crawling

Sepyan Purnama Kristanto[1], Lutfi Hakim[2], Dianni Yusuf[3], Citra Ayu Indriyani[4]
[1,2,3,4]Politeknik Negeri Banyuwangi, Indonesia

Informasi Artikel

Genesis Artikel:

Diterima, 13 Maret 2021
Direvisi, 24 April 2021
Disetujui, 21 Juni 2021

Kata Kunci:

Web Mining
Web Crawling
Depth-First Crawling
Tripadvisor
Jawa Timur

ABSTRAK

Jawa Timur memiliki area yang luas menjadi salah satu Provinsi yang memiliki banyak tempat wisata yang wajib dikunjungi. Terdapat banyak tempat wisata alam dan buatan yang tersebar dibeberapa Kabupaten, membuat Jawa Timur berkembang dari aspek pariwisata. Dengan banyaknya tempat wisata tersebut mengakibatkannya Pemprov Jawa Timur kesulitan dalam melakukan kontrol dan *monitoring* perkembangan serta evaluasi kualitas tempat wisata tersebut. Hal ini menghambat proses optimasi yang berfokus pada tempat wisata favorit, hal ini yang melatarbelakangi perlu adanya proses klasterisasi berdasarkan data raster tempat wisata tersebut. Aplikasi *Trip Advisor* dipilih sebagai objek sumber data dikarenakan aplikasi ini memiliki banyak *fitur* unggulan antara lain, *review, rating, history* hingga rekomendasi yang dapat mendukung proses klasterisasi. Metode pengumpulan data menggunakan model *Web Mining* dengan salah satu metode terbaik dalam penggalian informasi pada *web* yaitu *Depth-First Crawling* dimana data akan diekstrak berdasarkan hubungan antara data tersebut sehingga mendapatkan data akurat. Proses *crawling* berfokus pada komponen Alamat, Kota, *Rating* serta *Review* dari masing-masing tempat wisata di Jawa Timur. Hasil penelitian menunjukkan prosentase *success rate* dalam extraksi informasi yang didapatkan dari ke 4 *crawling variable* dengan 211 jumlah data dengan masing-masing *success rate* 97% untuk alamat, 99% untuk Kota, 70% untuk *Rating* dan 60% untuk *Review*.

Keywords:

Web Mining
Web Crawling
Depth-First Crawling
Tripadvisor
East Java

ABSTRACT

East Java has a large area, becoming one of the provinces with many tourist attractions that must be visited. With so many tourist attractions, the East Java Provincial Government has difficulty controlling and monitoring developments and evaluating the quality of these tourist attractions. Furthermore This hampers the optimization process that focuses on favorite tourist attractions. The clustering process using raster data of these tourist attractions. The TripAdvisor application was chosen as the data source object because this application has many excellent features, including reviews, ratings, history, to recommendations that can support the clustering process. The data collection method uses the Web Mining model with one of the best methods of extracting information on the web, namely Depth-First Crawling, where the data will be extracted based on the relationship between the data so that it gets accurate data. The crawling process focuses on the Address, City, Rating, and Review components of each tourist spot in East Java. The results showed the percentage of success rate in extracting information obtained from the four crawling variables with 211 amounts of data with 97% success rate for addresses, 99% for cities, 70% for ratings, and 60% for reviews.

This is an open access article under the CC BY-SA license.

Penulis Korespondensi:

Sepyan Purnama Kristanto,
Jurusan Teknik Informatika,
Politeknik Negeri Banyuwangi,
Email: sepyan@poliwangi.ac.id

1. PENDAHULUAN

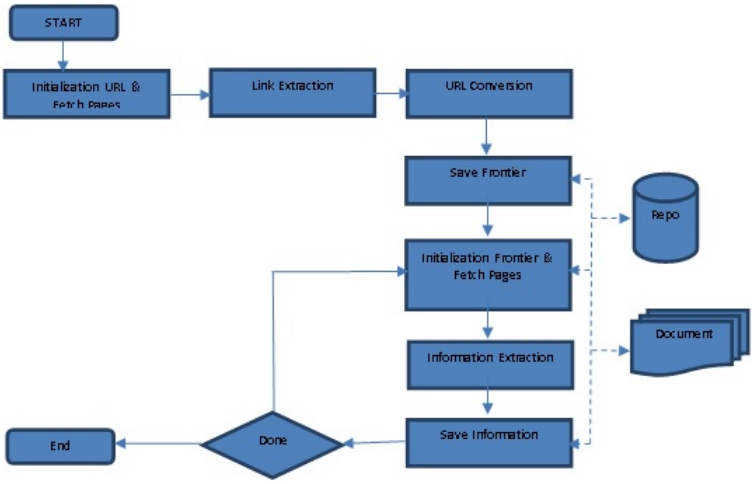
Destinasi wisata merupakan bagian yang tidak terpisahkan dari kehidupan manusia saat ini. Banyak orang berlomba-lomba dalam mencari dan mengunjungi tempat wisata, salah satunya yaitu tempat wisata yang ada di Jawa Timur [1]. Jawa Timur merupakan salah satu Provinsi yang memiliki banyak wilayah dan memiliki luas wilayah 47.800 km, dengan luas wilayah tersebut menjadikan Jawa Timur sebagai Provinsi terbesar diantara 6 Provinsi di Pulau Jawa. Provinsi ini memiliki perbatasan langsung dengan Samudra Hindia di Selatan serta Laut Jawa disebelah Utara yang membuat Jawa Timur memiliki beragam Fauna dan Flora. Banyaknya sumber daya yang dimiliki Jawa Timur manjadikan salah satu faktor yang membuat wisatawan *local* dan mancanegara datang berkunjung. Selaras dengan yang disampaikan oleh Dinas Pariwisata Jawa Timur yang mengatakan bahwa Jawa Timur memiliki potensi besar dalam sektor pariwisata sehingga perlu adanya *mapping* dan klasterisasi area wisata yang potensial [2].

Saat ini, Jawa Timur menjadi salah satu objek wisata yang sering dikunjungi banyak wisatan terutama dari Manca Negara, beberapa tempat yang paling sering dikunjungi yaitu : Gunung Bromo Pasuruan , Gunung Semeru di Malang atau beberapa tempat wisata lain seperti Ijen di Banyuwangi. Setiap lokasi tersebut menjadi tempat yang sering dikunjungi ketika para turis datang atau bahkan sebelum mereka kembali ke Negara Asal mereka. Banyaknya wisatawan yang berkunjung tidak lepas dari peran teknologi dalam menyebarkan informasi serta aktivitas setiap turis yang datang. Beragam portal serta aplikasi telah hadir dan membantu para *traveler* dalam menemukan tempat terbaik untuk dikunjungi termasuk di Jawa Timur. Salah satu portal informasi yang menyediakan informasi yang terbaik bernama *Tripadvisor* [3]. Portal ini menjadi salah satu portal yang dijadikan rujukan oleh banyak *traveler* sebelum mereka mengunjungi tempat tertentu untuk berkunjung atau berlibur. Banyak fitur menarik yang dimiliki *TripAdvisor* untuk mempermudah para *traveler* mengetahui kelebihan serta kekurangn tempat yang akan mereka kunjungi, beberapa fitur yang disukai diantaranya adalah *Review*, *Rating* serta Komentar dari temapat yang akan mereka kunjungi. Selain itu terdapat fitur lain yang terintegrasi dengan layanan jasa lainnya, seusai dengan program Pemerintah Dinas Provinsi yang sangat berfokus pada pengembangan tempat parisawata tentunya *TripAdvisor* ini merupakan media terbaik dalam melakukan analisa dan investigasi. Sebelum melakukan analisa serta pengembangan perlu dilakukan klasterirsasi tempat wisata berdasarkan beberapa parameter baik dari *review*, *rating* dan lainnya. [3] Klasterisasi ini bisa dilakukan dengan menggunakan data raster yang dimiliki Dinas Pariwisata Jawa timur, namun data raster ini tidak memiliki komponen atau fitur yang dapat digunakan oleh pihak terkait untuk media promosi serta kualitas dari tempat tersebut berdasarkan *review* pengunjung. Dengan banyaknya tempat wisata yang berada di Jawa Timur, proses klasterisasi menjadi lebih sulit dikarenakan banyak tempat wisata yang tidak terindentifikasi atau banyak juga data wisatawan yang tidak terdokumentasi dengan baik. Oleh karena itu, pada penelitian ini bertujuan untuk mengumpulkan informasi tempat wisata favorit di Provinsi Jawa Timur dengan menggunakan web sebagai sumber data utama. *World Wide Web* (WWW) sering kita sebut sebagai *web*, telah menjadi sumber informasi utama dan banyak dikenal orang. Ketika seseorang mencari inforamsi tujuan utama yang mereka lakukan adalah membuka sebuah *web* dan mengetikan informasi apa yang mereka butuhkan. *Web* sendiri merupakan jaringan milyaran dokumen yang saring terintegrasi dan dikelola oleh jutaan orang. *Web* sebagai *portal* informasi global memiliki kemudahan bagi setiap *user*, sehingga pencarian informasi menjadi sangat mudah dan murah. Penggalian informasi atau biasa disebut *web mining* merupakan salah satu solusi untuk mengumpulkan informasi tempat wisata populer yang memanfaatkan *web* sebagai sumber data utama. Penggunaan cara konvensional, seperti survei, wawancara atau dengan teknik statistika lainnya seringkali terhambat masalah dana serta waktu yang mengakibatkan proses tidak berjalan. Dengan berkembangnya teknologi pada industri 4.0 serta banyaknya informasi dan kumpulan data tersimpan, *web mining* diharapakn menjadi salah satu solusi terbaik dalam mengumpulkan informasi. *Web mining* merupakan isitilah yang digunakan dalam proses penerapan teknik *data mining* ke data *web* untuk mengekstraksi informasi yang sesuai dan relevan dari sumber data yang telah disediakan. [4]

Sumber daya yang di ekstrak merupakan kumpulan kombinasi data dokumen yang tersimpan dan tersedia dalam bentuk *web*. Pada *web mining* proses yang digunakan serupa dengan konsep *data mining* yang sering dilakukan pada *database* atau data *warehouse*. Sedangkan pada *web mining* proses pengumpulan data dilakukan pada suatu *website* yang telah ditentukan untuk mengunduh halaman *web* keseluruhan atau sesuai dengan konten yang dibutuhkan dengan menggunakan *crawling* [5]. *Crawling* sangat berperan penting pada *web mining* dengan melakukan proses otomasi pengumpulan informasi, proses ini melakukan fungsi khusus yang berjalan di *background*. Pada penelitian ini metode *web crawling* digunakan untuk pengumpulan informasi destinasi wisata pada aplikasi *TripAdvisor*. *TripAdvisor* dipilih karena *website* tersebut merupakan salah satu TOP #1 situs informasi wisata yang diakui di dunia dengan jumlah pengguna sebesar 750 juta [2]. Dengan memanfaatkan aplikasi *TripAdvisor* diharapkan proses ekstraksi informasi destinasi wisata lebih maksimal, dikarena kan aplikasi ini telah memiliki beragam fitur yang yang sesuai dengan kebutuhan pada proses klasterisasi.

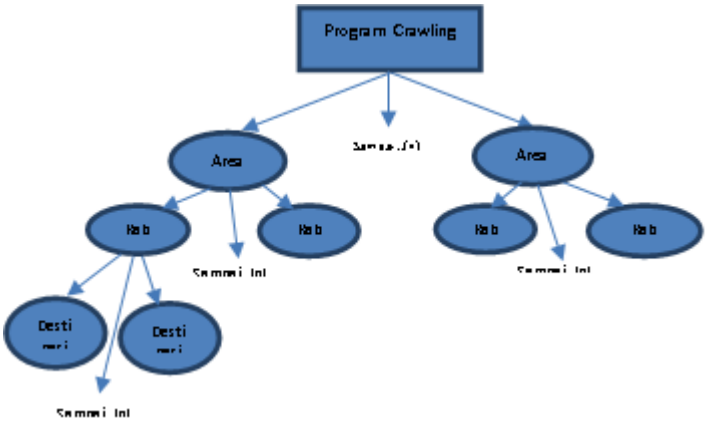
2. METODE PENELITIAN

Pada penelitian ini berfokus pada proses ekstrasi infomasi dari sebuah website atau disebut dengan *web mining* atau *web crawling*, pada prosesnya model *crawling* yang digunakan penelitian ini berfoskus pada teknik *Depth-First Crawling*. *Web Crawling* akan menggunakan BOT untuk mencari halaman *website* melalui URL (*Uniform Resource Locator*), serta mengembalikan data tersebut ke pengguna secara langsung. Dengan adanya *Crawler* pengguna tidak harus menelusuri halaman *web* satu per satu sehingga menghemat waktu serta dapat meningkatkan akurasi dari proses pengumpulan data pada gambar 1.



Gambar 1. Alur Proses Crawling

Terdapat beberapa penelitian terkait yang membahas tentang *web mining* dengan menggunakan *web crawling* sebagai teknik pengumpulan data, beberapa diantara adalah *mining user tweet* [6], lowongan pekerjaan, artikel berita atau *list movie* pilihan [7]. Setiap data yang di *Crawling* memiliki bentuk dan karakteristik yang berbeda mulai dari yang terstruktur serta yang memiliki bentuk tidak terstruktur. Beberapa contoh bentuk data diantaranya mulai artikel yang ada di *website*, *tweet* dan *posting* [8] atau komentar pengguna pada aplikasi [9], hingga data cuaca serta kejadian-kejadian tertentu yang ada disekitar kita. Sumber objek yang di gunakan sebagai sumber informasi utama dapat terdiri dari *multiwebsite* atau juga dapat menggunakan *singlewebsite* sebagai fokus utama dari data yang ingin di *mining*. [10] [14] *Depth-First Crawling* digunakan pada penelitian ini untuk menghasilkan informasi yang lebih akurat, model ini akan menelusuri proses *crawling* dari *node* pertama kemudian dilanjutkan pada *node* selanjutnya pada proses di level berikutnya. [14] [15] Model penelusuran informasi bergerak dari arah kiri hingga masuk ke *node* paling dalam dari sebuah *website*, jika proses telah menemukan batas informasi maka *node* akan bergerak mundur atau *backtrack* untuk melanjutkan ke *node* lainnya seperti pada gambar 2.



Gambar 2. Proses *Crawling* Menggunakan *Depth-First Crawling*

Langkap penyelesaian pada Depth-First Crawling (DFS):

1. *Node* ujung akar kedalam area
2. Ambil *node* dari awal area lalu cek apakah simpul merupakan solusi
3. Jika *node* merupakan solusi pencarian selesai dan hasil dikembalikan
4. Jika *node* bukan solusi, masukkan *node* tetangga atau *child node* berikutnya untuk mencari solusi berikutnya
5. Jika Langkah tersebut atau area tersebut kosong dan setiap *node* sudah dicek, pencarian selesai dan hasil dikembalikan dengan notifkasi solusi tidak ditemukan.

Pada Proses *crawling* yang akan digunakan dalam mengumpulkan informasi terkait lokasi, *review*, serta *rating* dari sebuah tempat wisata yang ada di Provinsi Jawa Timur dengan menggunakan *website* dari *TripAdvisor*. *TripAdvisor* dipilih karena memiliki reputasi yang sangat baik dikalangan *traveler* diantara beberapa *website* tempat wisata lainnnya seperti wisataku.com, jatim*tourism*, atau beautifulindonesia.com.[12] selain memiliki kualitas yang baik, *tripadvisor* memiliki sistem managemen data yang baik sehingga setiap data yang akan *dimining* dapat dengan mudah di proses untuk selanjutnya. Mayoritas proses *web crawling* menggunakan bahasa pemprograman Python dengan *tool* pendukung seperti selenium. Pada penelitian ini *tool* yang akan digunakan adalah Selenium dengan bahasa pemprograman Python. Pada proses *web crawling* ini dilakukan dengan beberapa tahapan yaitu:

1. Initialization URL and Fetch Page
   Langkah awal dalam melakukan *crawling* adalah melakukan inisiasi alamat *website* yang akan di *mining*, Proses ini disebut *Initialization* URL dimana Main URL yang akan digunakan adalah *webpage* utama dari *TripAdvisor*.

```
"https://www.tripadvisor.co.id/Search?q=jawa%20timur&searchSessionId=BB69
D9804C330F2203AC50F4E5A0E0691588750688832ssid&sid=09BC0B9607A309B30EEF31F
FB3F549C81588750727956&blockRedirect=true&ssrc=A&geo=2301797&rf=6&o=150";
```

Gambar 3. URL yang digunakan

2. Link Extraction
   Tahap berikutnya dalam melakukan *crawling* setelah melakukan inisiasi adalah melakukan ekstrasi URL yang telah diinputkan. Pada proses ini terdapat *Frontier* sebagai alamat atau URL yang belum dikunjungi, proses extrak menggunakan model Regular *Expression* yang mengindentifikasi pola pada Main Url dalam data *text*. Pada waktu yang sama selain melakukan identifikasi, proses ekstrasi berjalan simultan dengan melakukan *parsing* berdasarkan tag HTML.
3. URL Conversion
   Pada tahap ini URL hasil proses ekstraksi yang masih memiliki bentuk yang absurd serta tidak baku (Raw URL) tidak dapat langsung disimpan sebagai data *Frontier*. Raw URL harus diubah terlebih dahulu menjadi bentuk baku (Standart URL). Konversi dilakukan dengan menambahkan Main URL dari *website* utama ke Raw URL.
   Contoh:
   Raw URL
   $Wisata\_Rating - sda1231231 - d2523421 - Rating - S - Gbromo.html$
   Standart URL
   $https://tripadvisor.co.id/Wisata\_Rating - sda1231231 - d2523421 - Rating - S - Gbromo.html$
4. *Saving Frontier*
   Standart URL selanjutnya di simpan ke Repo untuk digunakan pada tahap berikutnya.
5. Initialization Frontier and Fetch the Page
   Pada tahap ini data URL selanjutnya diambil dari *Repository* untuk di *explore*. Pada penelitian ini hasil *explorer* merupakan detail informasi dari tempat wisata.
6. Information Extraction
   Selanjutnya URL hasil dari inisialisasi *Frontier* di extrak dengan menggunakan model Parsing untuk menghasilkan informasi berdasarkan pada struktur halaman HTML. Pada penelitian inii beberapa informasi yang diekstrak yaitu Alamat, Kota, *Rating* dan *Review*. Tabel 1 memperlihatkan jenis informasi yang di ekstrak dari halaman HTML. Beberapa informasi berada pada satu halaman HTML seperti informasi *rating* serta *review* dari tempat wisata tersebut.

Tabel 1. Informasi dan HTML

No	Informasi	HTML
1	Alamat	address
2	Kota	city
3	Rating	rating
4	Review	rating

   Data informasi dibatasi hanya 4 dikarenakan informasi tersebut merupakan informasi utama yang dibutuhkan oleh Dinas Pariwisata untuk melakukan klasterisasi tempat wisata terfavorit. Ke empat data tersebut yang yang mimiliki data html yang salam selanjutnya dilakukan proses ektraksi dengan menggunakan model Regular *Expression*.
7. Saving the Information
   Hasil informasi yang telah diolah selanjutnya disimpan pada *repository*

```
from selenium.webdriver.common.keys import Keys
from urllib.parse import urlparse
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import requests
```

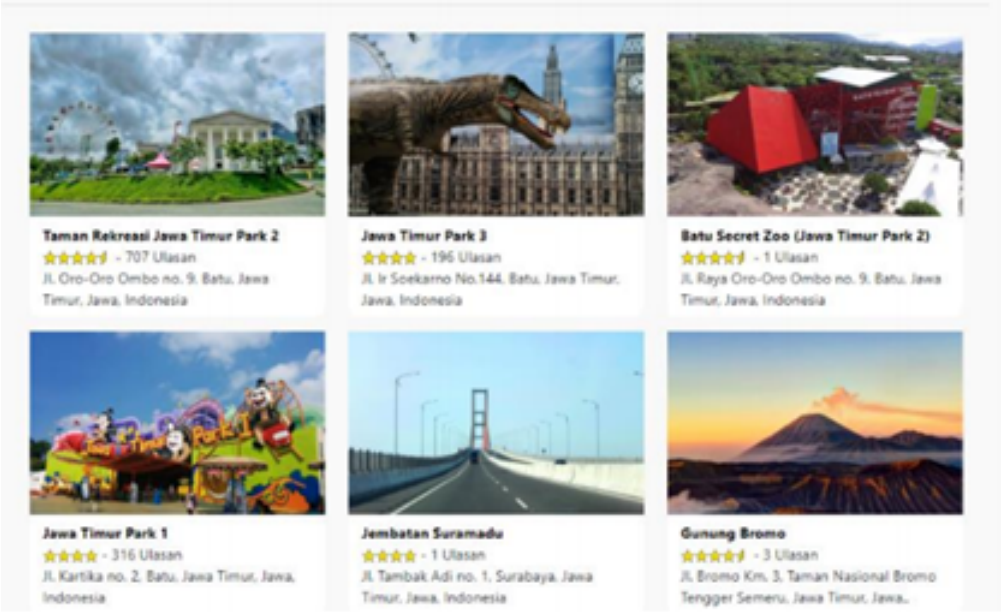
Gambar 4. Proses import selenium

Proses *Crawling* menggunakan selenium sebagai sistem otomasi agar *web browser* dapat berjalan otomatis, sehingga proses ektraksi dapat dilakukan secara menyeluruh. Selenium digunakan karena memiliki performa yang baik dalam melakukan proses *Crawling* data pada suatu *website*. [13] Sebelum menggunakan selenium sebagai *tool* untuk melakukan proses *crawling* serta otomasi, kita harus melakukan *import library* tersebut pada halaman atas atau *header* dari program python tersebut seperti pada gambar 4.

3. HASIL DAN ANALISIS

Proses ekstraksi dilakukan dengan cara inisiasi URL *TripAdvisor*, inisiasi dilakukan dengan menggunakan link utama *Tripadvisor* sebagai berikut:

https://www.tripadvisor.co.id/Search?q=jawa%20timur&searchSessionId=19E1ED5BB1D0F80E4484525E4BE094641619-238631639ssid&searchNearby=false&sid=18BF25139AEB46CBB91431EB17B7CCD01619238680414&blockRedirect=true" link utama selanjutnya digunakan untuk proses Link Extraction dengan menggunakan Regular *Expresion* sehingga mendapatkan 211 informasi seputar destinasi wisata di area Jawa Timur yang dapat kita lihat dari gambar 5. Setiap destinasi wisata memiliki komponen penunjang atau informasi tambahan antara lain *rating*, ulasan, lokasi hingga beberapa foto dari tempat wisata tersebut pada gambar nomor 6.



Gambar 5. List tempat wisata

Pada gambar 5 terlihat data *list* tempat wisata yang ada di Jawa Timur dengan menampilkan beberapa komponen penunjang antara lain *rating*, ulasan, serta lokasi dmana tempat wisata itu berada. Setiap tempat wisata yang ditampilkan berbentuk *cardview* dengan luas area maksimal di Provinsi Jawa Timur. Jika ingin mendapatkan data detail terkait tempat wisata tersebut, *user* dapat menekan nama atau *header* dari tempat wisata atau juga dapat merubah tampilan dari *cardview* menjadi *listview* seperti pada gambar 6.

Gambar 6. Detail Destinasi Wisata

Setelah data URL sudah *fix* proses berikutnya adalah menggunakan URL tersebut untuk melakukan proses *crawling*, masukkan URL tersebut kedalam program *crawling*. Pada program *crawling* selain memasukkan URL yang akan digunakan, parameter informasi yang dibutuhkan juga dimasukkan pada aplikasi.

```
review1 = driver.find_elements_by_class_name("result-title") review2 =
driver.find_elements_by_class_name("address-text") address-text =
driver.find_elements_by_class_name("city-text") city-text =
driver.find_elements_by_class_name("inner") inner =
```

Gambar 7. Parameter Input

Hasil dari proses *crawling* akan dimunculkan dalam data exel dengan formal (CSV), list tempat wisata ditampilkan berdasarkan area yang dipilih serta telah disesuaikan dengan cakupan luas area yang digunakan. Dengan metode Depth-First destinasi wisasta dimunculkan berserta rekomendasi kedekatan atau kemiripan tempat wisata yang dicari. Node yang mewakili dari key pencarian digunakan sebagai KeyWord untuk mencari similarity serta korelasi antara tempat satu dengan tempat lainnya. Dalam hal ini data awal hanya menampilkan 10 informasi atau destinasi selanjutnya sistem akan melakukan *crawling* otomatis berdasarkan halaman atau destinasi yang muncul pertama kali dengan menggunakan keterkaitan terhadap lokasi atau area dimana tempat wisata itu berada. Setelah dilakukan proses *crawling* hasil utama dari proses tersebut adalah bentuk data abstrack seperti pada gambar 8, yang berisikan URL serta ada beberapa yang menampilkan halaman HTML yang tercampur dengan beberapa *text*. Dari data tersebut selanjutnya dilakukan normalisasi serta pembersihan dari data awal agar terbentuk data real atau proses ini sering di sebut *stripping*.

Gambar 8. Contoh hasil crawling

Dari hasil yang telah didapatkan pada proses *crawling* masih berbentuk data mentah yang berisikan data halaman serta komponen atau tag dari halaman yang tidak lengkap dan acak. Tahapan ini sudah menjadi akhir proses *crawling*, namun untuk membaca data hasil *crawling* masih sulit dilakukan karena data yang diterima masih berupa data *raw* yang bercampur antara main *text* dan *punctuation* agar data dapat dibaca dengan baik selanjutnya dilakukan pemprosesan lebih lanjut dengan menggunakan *stemming*, *stopword* serta *post tagging*. Pada proses ini data *Raw* yang masih berbentuk HTML dilakukan proses *preparation* dengan menggunakan *tool* dari NLTK.

Informasi yang telah di dapatkan masih memiliki beberapa kekurangan terkait informasi, terdapat beberapa tempat wisata yang tidak memiliki variabel informasi yang lengkap sesuai yang dibutuhkan. Hasil proses ekstraksi ditampilkan pada tabel 2 yang menampilkan data hasil ekstraksi beserta variabel informasi atau kelengkapan yang diperoleh. Prosentase kelengkapan juga ditambilkan untuk mengetahui sejauh mana setiap data memiliki kelengkapan komponen variabel yang dibutuhkan dalam proses klasterisasi. Setiap data yang tidak memiliki kelengkapan komponen atau variabel disebabkan data atau tempat wisata tersebut merupakan tempat wisata yang memiliki *rating* rendah atau dimungkinkan tempat wisata tersebut jarang dikunjungi oleh wisatawan sehingga sangat jarang wisatawan yang memberikan *review* atau *rating*.

Tabel 2. Hasil Informasi

No	Destinasi	Alamat	Kota	Rating	Review
1	Gunung Bromo	Probolinggo	Probolionggo	5	This picturesque volcano in Bromo Tengger Semeru National Park has a stairway to the crater
2	Gunung Semeru	Malang	Malang	4, 6	If you're a serious hiker, summiting this volcano will be worth the effort and sweat. It's a tough climb near the summit due to the lose gravel, but the view from atop is simply breathtaking - it'll be worth it! the whole trail from Ranu Pani village - Ranu Kumbolo lake - Kalimati - Semeru is beautiful. If you have time, opt to stay a night at Ranu Kumbolo lake where you'll wake up to the beautiful lake view - as if you've been transported into a fairytale. 10/10 would recommend
3	Gunung Ijen	Banyuwangi	Banyuwangi	5	Really great experience we got when visit this magnificent place. Very beautiful place with beautiful people but PLEASE, keep this place clean as possible as so many tissuethrown away. Such a shame behavior from what we called a nature lover.
4	Jatim Park	Malang	Malang	4, 7	The facilities were clean and adequate. Most of the facilities were well maintained. There weretoilet, food court, prayer room, souvenir shop, and a spacious parking lot. The place has appealing design. There were many interesting photo spots. There are three exciting sections at Jatim Park 2, the zoo, animal museum, and small amusement park with some rides.
5	B29 Mountain	Lumajang	Lumajang	4, 7	My first time doing mt climbing was at the top of B29. I was so happy i can saw a beautiful side of mountains in Indonesia. Really can't believe that we need to use motorbike at first, it's little bit scary. But after we arrived at the top of B29 i was thankful and feel blessed for the beautiful scenery.

Tabel 2 merupakan contoh hasil *crawling* yang didapatkan dan telah dilakukan proses *striping* sehingga informasi telah di rubah dan lebih mudah di baca. Data serta informasi yang telah diambil selanjutnya dimasukkan ke *database* MYSQL agar tidak perlu dilakukan *crawling* berulang kali ketika dibutuhkan untuk pengolahan selanjutnya. Mayoritas informasi yang dibagi beberapa jenis berdasarkan kelengkapan berkas atau komponen. Dari 211 data inforamsi tempat wisata sebesar 30% merupakan tempat wisata yang tidak memiliki kelengkapan komponen serta 70% merupakan tempat wisata yang memiliki kelengkapan komponen variabel atau berkas yang sesuai. Data yang telah didapatkan selanjutnya dapat dianalisas kembali dengan model analisa spasial dari lokasi persebaran tempat wisata untuk mengetahui hubungan lokasi tempat wisata dengan minat para wisatawan untuk datang mengujungi tempat tersebut atau dapat juga dilakukan analisa *sentiment* terhadap tempat wisata tersebut dengan menggunakan data *review* wisatawan. Hasil dari analisa tersebut dapat digunakan sebagai pedoman dalam menentukan kebijakan serta pengembangan tempat wisata tersebut, atau bisa juga digunakan sebagai penentu model *marketing* untuk mengenalkan tempat wisata ke wisatawan Lokal atau Mancanegara.

Tabel 3. Hasil Ekstrasi Informasi dari *website*

No	Informasi	Total Success	Total Unsuccess	Procentage Rate	Keterangan
1	Alamat	201	10	97%	Valid
2	Kota	210	1	99%	Valid
3	Rating	131	80	70%	Valid
4	Review	111	100	60%	Valid

Pada tabel 3 dapat kita lihat hasil extraksi yang telah dilakukan, dari keseluruhan data beberapa data tidak memiliki kelengkapan variabel, *review* pengunjung tidak selalu dimiliki oleh tempat wisata. Hal ini dimungkinkan bukan karena wisatawan tidak ingin me*review* melainkan beberapa tempat wisata jarang di kunjungi oleh wisatawan. Sangat dimungkinkan karena beberapa tempat wisata kurang dikenal oleh wisatawan lokal bahkan mancanegara, kurang dikenalnya tempat wisata tersebut mengakitkan wisatawan tidak memiliki cukup informasi terkait destinasi tersebut.

4. KESIMPULAN

Hasil penelitian yang telah dilakukan menghasilkan beragam hasil data yang dikelompokan berdasarkan Alamat, Kota, *Rating* serta *Review* berbagai macam tempat wisata yang berada di Jawa Timur. Kumpulan data tersebut sesuai dengan kebutuhan akan informasi yang diingikan oleh Dinas Provinsi Jawa Timur untuk proses klasterisasi, penelitian itu juga dapat diambil kesimpulan terkait *web crawling* dapat di fungsikan sebagai metode untuk melakukan pengumpulan dan ektraksi informasi dari sebuah web. Sumber data yang dapat digunakan sebagai objek *web crawling* dapat menggunakan berbagai jenis *website* baik *ecommerce*, blog atau *website* berita tergantung kebutuhan yang ingin kita capai. Hasil proses *web crawling* yang telah dilakukan pada *website TripAdvisor* menunjukkan beberapa rincian informasi yang tidak sempur, dari proses tersebut menghasilkkan *success rate* sebesar 81,5%. Data yang didapatkan masih memiliki banyak kekurangan terutama beberapa destinasi tidak memiliki kelengkapan dokumen atau variabel yang dibutuhkan dikarenakan banyak wisatawan tidak mengisi atau memang wisatawan kurang mengenal tempat wisata tersebut sehingga probalitisa kemunculan tempat wisata tersebut di halaman pencarian *TripAdvisor* sangat kecil yang mengakitbakan wistawan enggan datang ke tempat wisata itu. Oleh sebab itu penelitian dapat dikembangkan dengan menambah sumber data dalam hal ini adalah *website*, pengembangan dapat juga di arahkan dengan integrasi dengan *Social* Media Platform atau bisa juga diintegrasikan dengan beberapa platform lain yang terkait seperti biro perjalanan. Sehingga semakin banyak sumber data maka informasi terkait tempat wisata favorit serta *success rate* akan semakin baik sehingga data yang didapatkan jauh lebih variatif.

REFERENSI

[1] Muktiyah Kumala, A. S. (2017). Analisis Potensi Sektor Pariwisata Sebagai Sektor Unggulan di Wilayah Jawa Timur. Ilmu Ekonomi, 474-481.

[2] Novan, Kurnia.2015. "Pariwisata Jawa Timur" [Online]. Tersedia: http://disbudpar.jatimprov.go.id/pariwisata-jawatimur.html. [Diakses: 15 Maret 2021]

[3] R. Hanifah and I. S. Nurhasanah, Implementasi Web Crawling untuk Mengumpulkan Web Crawling Implementation for Collecting, JTIIK: Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 5, no. 5, pp. 531536, 2018

[4] E. Susanti and K. Mustofa, Ekstraksi Informasi Halaman Web Menggunakan Pendekatan Bootstrapping pada Ontology-Based Information Extraction, IJCCS:Indonesian Journal of Compuing and Cybernetic System, vol. 9, no. 2, pp. 111-121, 2015

[5] R. Qian, K. Zhang, and G. Zhao, "A topic-specific Web crawler based on content and structure mining," Proc. 2013 3rd Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2013, pp. 458461, 2014

[6] A. B. Archana and J. Kumar, "Location based semantic information retrieval from web documents using web crawler," Proc. 2015 Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2015, pp. 370375, 2016

[7] L. B. Ilmawan, Membangun Web Crawler Berbasis Web Service untuk Data Crawling Pada Website Google Play Store, ILKOM Jurnal Ilmiah., vol. 10, no. 2, pp. 215224, 2018

[8] Z. Shi, M. Shi, and W. Lin, "The Implementation of Crawling News Page Based on Incremental Web Crawler," Proc. - 4th Int. Conf. Appl. Comput. Inf. Technol. 3rd Int. Conf. Comput. Sci. Appl. Informatics, 1st Int. Conf. Big Data, Cloud Comput. Data Sci. Eng. ACIT-CSII-BCD 2016, pp. 348351, 2017

[9] Y. Wang, Z. Hong, and M. Shi, "Research on LDA Model Algorithm of News-oriented Web Crawler," Proc. - 17th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2018, pp. 748753, 2018

[10] N. C. C. A. Phitaloka, Web Content Mining di Sektor Perbankan Pada Lq45 untuk Pendukung Keputusan Investasi Saham, Telematika: Jurnal Informatikan dan Teknologi Informasi, vol. 16, no. 1, p. 18, 2019

[11] S. P. Kristanto, J. A. Prasetyo, and E. Pramana, "Naive Bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH," Proc. - 2019 2nd Int. Conf. Comput. Informatics Eng. Artif. Intell. Roles Ind. Revolut. 4.0, IC2IE 2019, pp. 2428, 2019

[12] S. Budi, Text Mining untuk Analisis Sentimen Review Film Menggunakan Algoritma K-Means, TechnoCom : Jurnal Teknologi Informasi, vol. 16, no. 1, pp. 18, 2017

[13] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation," in Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, Jan. 2016, pp. 13481353

[14] Kustanto, Cynthia, Mutia S, Ratna, Viqarunnisa, Pocut, "Penerapan Algoritma Breadth-first Search dan Depth-first Search Pada FTP Search Engine for ITB Network", Teknik Informatika, Institut Teknologi Bandung, Bandung

[15] Google. Depth First Search. [Online]. Tersedia: https://saungkode.wordpress.com/2014/04/16/penelusuranpohon-biner-berdasarkan-kedalaman-dengan-algoritma-dfs-stack-dan-secara-melebar-level-orderdengan-algoritma-bfs-queue-dan-implementasinya-dalam-bahasa-c/. [Diakses: 23 Maret 2021]