Implementasi Teknik *Web Scraping* pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia

Yoga Sahria
Magister Teknik Informatika/Fakultas Teknologi Industri, Universitas Islam Indonesia
Email: 17917225@students.uii.ac.id

Keywords:
Web Scraping;
SINTA; Topik.

Abstrak

Semakin berkembang pesat teknologi, menyebabkan penelitian kesehatan indonesia semakin maju. Perkembangan teknologi ini mendorong manusia untuk meneliti kususnya di bidang kesehatan di Indonesia. Dengan banyaknya Penelitian kesehatan indonesia diperlukan analisis informasi yang digunakan untuk memetakan penelitian indonesia yang lebih cepat dan efisien untuk mendapatkan topik penelitian kesehatan. Penelitian ini dibuat untuk mengetahui topik penelitian kesehatan di Indonesia yang terdapat pada Jurnal SINTA. Metode yang digunakan yaitu dengan mengimplementasikan web scraping pada Jurnal SINTA. Teknik scraping mengambil judul jurnal kesehatan, judul penelitian, author, afiliasi dan kemudian dianalisis hasil pengumpulan data tersebut. Pada proses scraping dilakukan dengan identifikasi kelas tag HTML.Tag HTML yang digunakan yaitu tag yang mengapit judul jurnal kesehatan, judul penelitian, author dan afiliasi untuk kemudian dibuatkan tamplate scraping. Data yang diperoleh kemudian dianalisis sehingga dapat diketahui tren topik penelitian kesehatan di Indonesia di Jurnal SINTA. Penelitian ini dibuat menggunakan bahasa Python dengan modul–modul yang mendukung untuk diterapkan. Penelitian ini dapat memproses web scraping dari Jurnal SINTA dan kemudian data disimpan ke dalam bentuk format CSV. Dokumen format CSV yang diperoleh kemudian diolah menggunakan python untuk diperoleh suatu model. Hasil keluaran berupa tren topik penelitian kesehatan di Indonesia, banyaknya penelitian kesehatan, afiliasi penelitian kesehatan di Indonesia.

1. PENDAHULUAN

Semakin pesatanya perkembangan teknologi kesehatan dan peneliti kesehatan, semakin banyak penelitian kususnya di bidang kesehatan di Indonesia, meningkatnya teknologi kesehatan semakin kebutuhan informasi yang mendorong untuk mengolah informasi yang dilakukan dengan mudah, cepat dan efisien. Dengan adanya internet hasil penelitian dengan mudah disimpan kedalam portal jurnal secara *onlin*e. Semakin banyaknya penelitian semakin banyak pula penyimpanan media digital yang mendorong terjadinya banyak jumlah dokumen elektronik yang tersimpan dalam repository jurnal atau publikasi ilmiah. Namun, pada umumnya hal ini tidak disertai dengan pertumbuhan jumlah analisis informasi atau pengetahuan yang dapat disarikan dari dokumen karya ilmiah tersebut. Salah satu fasilitas pendukung untuk melihat penelitian kesehatan di Indonesia adalah jurnal SINTA (*Science and Technology Index*).

Jurnal Sinta (*Science and Technology Index*) yaitu jurnal yang dibuat untuk menyimpan hasil penelitian-penelitan di Indonesia[1], salah satunya adalah penelitian kesehatan di Indonesia. Jurnal SINTA Sistem informasi penelitian berbasis web menawarkan akses cepat, mudah dan komprehensif untuk mengukur kinerja para peneliti, lembaga dan jurnal di Indonesia. Sinta memberikan tolok ukur dan analisis, identifikasi kekuatan penelitian masing-masing lembaga untuk mengembangkan kemitraan kolaboratif, untuk menganalisis tren penelitian dan direktori ahli.

Web Scraping adalah metode pengumpulan data melalui internet, meskipun *web scraping* bukan sesuatu hal yang baru, belakangan ini metode web *scraping* sangat populer digunakan untuk pemenuhan data mining[2]. Sebelumnya metode ini dikenal kedalam beberapa istilah, diantaranya *screen scraping*, *data mining*, *web harvesting* ataupun metode lain yang sejenis[3]. Menurut teori, web *scraping* adalah cara untuk mengumpulkan data menggunakan metode yang berbeda dengan penggunaan API (*Aplication Programming Interface*)[4]. Cara seperti ini biasanya dimulai dengan penulisan kode program. Yang dimana digunakan sebagai automatisasi *query* untuk melakukan request data terhadap *server*. Data hasil *request* tersebut dapat dilakukan ekstrasi untuk menghasilkan informasi yang akan dicari[5]. Berikut langkah web scraping dapat dilihat pada gambar 1.

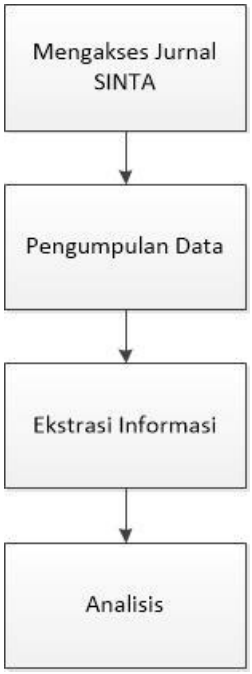
Gambar 1.Web *Scraping*

Manfaat *web scraping* adalah untuk mengambil informasi agar informasi yang diambil lebih berfokus sehingga memudahkan dalam melakukan pencarian sesuatu. *Web Scraping* sebuah teknik lunak komputer teknik penggalian informasi dari situs web *online*[6], adapun cara mengembangkan teknik *web scraping* yaitu dengan cara pertama Pembuat program mempelajari dokumen HTML dari website yang akan diambil informasinya untuk di tag HTML[7].

Berdasarkan masalah yang telah diuraiakan peneliti membangun analisis pemetaan yang mampu membantu mengumpulkan data dari jurnal SINTA. Penelitian ini didasarkan pada teknik *Web Scraping* untuk mengekstrak data penelitian kesehatan di laman situs jurnal secara otomatis disimpan untuk kemudian dianalisis.

2. METODE

Penelitian pada penelitian ini menerapkan teknik web scraping dilakukan dengan observasi pada Jurnal SINTA yang afiliasi dari kementrian kesehatan indonesia yang terindeks oleh jurnal SINTA adapun gambaran metode penelitian sebagai berikut:



Gambar 2. Metode Penelitian

Tahapan yang pertama adalah mengakses jurnal SINTA yang akan dijadikan sebagai sumber informasi. Pada tahap ini dilakukan pencarian jurnal kesehatan indonesia yang afiliasinya dari kementrian kesehatan untuk menemukan jurnal khusus menyajikan publikasi penelitian kesehatan di Indonesia. Dari hasil pencarian pada jurnal SINTA, diambil 4 Jurnal dari Kementrian Kesehatan Indonesia yang secara spesifik menyajikan publikasi yang terindeks SINTA pada S2 pada url http://sinta2.ristekdikti.go.id/journals/detail ?. Daftar jurnal yang digunakan dalam penelitian ini diperlihatkan oleh tabel 1.

Tabel 1. Nama Jurnal dan Id

No	Nama Jurnal
1	Ekologi Kesehatan (EK) id =21
2	Media Penelitian dan Pengembangan Kesehatan (MPPK) id=2791
3	Buletin Penelitian Kesehatan (BPK) id=979
4	Penelitian Sistem Kesehatan (PSK) id=2518

Tahapan kedua yaitu melakukan pengumpulan data dengan teknik *web scraping*. *Web Scraping* adalah proses pengambilan sebuah informasi dokumen yang bersifat semi struktur dari internet yang berupa halaman-halaman web yang berbentuk HTML atau XHTML. Pengumpulan data ini diambil dari halaman Jurnal SINTA dengan berdasarkan id URL jurnal kesehatan yang berafiliasi dari Kementrian Kesehatan Indonesia.

Pada tahap ketiga yaitu ekstrasi informasi yang dihasilkan dari pengumpulan data dengan *scraping* kemudian data yang tidak tersetruktur dijadikan tersetruktur sehingga dapat mempermudah untuk pengolahan data.

Pada tahap keempat yaitu Analisis data. Hasil Analisis penelitian ini diharapkan dapat membantu Peneliti Kesehatan, Pelayan Kesehatan dan Pemerintah dalam mengambil keputusan. untuk mendukung hasil analis penelitian penelitian ini menggunakan visualisasi *grafik, chart, wordcloud* yang bertujuan untuk memvisualisasi data untuk mengkomunikasikan informasi secara jelas dan efisien untuk kemudian dianalisis kepada pengguna lewat grafik informasi yang dipilih.

3. HASIL DAN PEMBAHASAN

Hasil penelitian ini adalah mengimplementasikan teknik *scraping* pada jurnal SINTA dengan menggunakan bahasa pemograman *Python* dan hasil *scraping* ditampung di exel dengan format CSV kemudian data dianalisis untuk mengetahui topik penelitian kesehatan Indonesia.

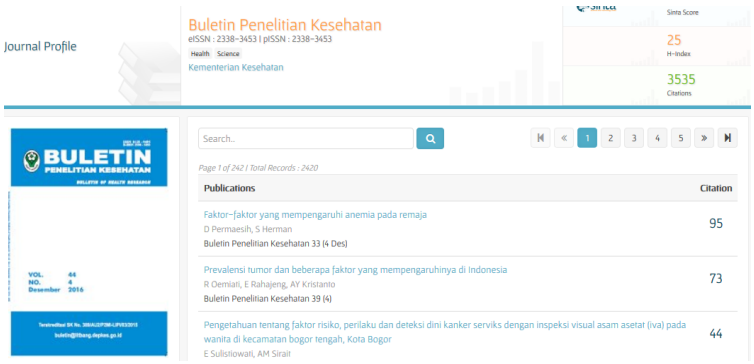
3.1. Akses Jurnal SINTA

Berikut halaman web jurnal SINTA yang di *scraping* berdasarakan id 4 jurnal yang dikumpulkan data nya yaitu terdiri dari jurnal ekologi kesehatan, media penelitian dan pengembangan kesehatan, buletin penelitian kesehatan, penelitian sistem kesehatan. keempat jurnal tersebut berafiliasi dari kementrian kesehatan Indonesia. Berikut halaman web jurnal SINTA yang di scraping dapat dilihat pada gambar 3,4,5 dan 6.

Gambar 3. Jurnal Penelitian Sistem Kesehatan

Gambar 4. Jurnal Ekologi Kesehatan

Gambar 5. Jurnal Media Penelitian Kesehatan



Gambar 6. Penelitian Sistem Kesehatan

3.2. *Scraping* Data

Dalam menjalankan teknik *web scraping*, terlebih dahulu harus diketahui struktur HTML dari laman web untuk menentukan dalam tag HTML yang mana informasi inti direpresentasikan. Berikut structur HTML yang digunakan pada Jurnal SINTA dapat dilihat pada gambar 7.

```
<table class="uk-table">
  <caption>Page 1 of 38 | Total Records : 378</caption>
  <thead> ... </thead>
  <tbody>
    <tr>
      <td>
        <!--<img class="journal-cover uk-align-left" src="img/cover-journal.png"/>-->
        <dl class="uk-description-list-line">
          <dt class="uk-text-primary">
            <a class="paper-link" href="https://scholar.google.co.idjavascript:void(0)"
            target="_blank"> event
              Faktor-faktor yang mempengaruhi kejadian TB paru dan upaya penanggulangannya
            </a>
          </dt>
          <dd>HSP Manalu</dd>
          <dd class="indexed-by">
```

Gambar 7. HTML Jurnal SINTA

Setelah mengetahui tag laman web yang akan diakusisi. Dalam penelitian ini menggunakan teknik *Scraping* yang ditulis dalam bahasa pemrograman Python untuk mempermudah proses *scraping*. Berikut daftar url yang menjadi target *scraping* data di jurnal Sinta pada tabel 2.

Tabel 2. Target Web Scraping

No	URL
1	http://sinta2.ristekdikti.go.id/journals/detail?id=21
2	http://sinta2.ristekdikti.go.id/journals/detail?id=2791
3	http://sinta2.ristekdikti.go.id/journals/detail?id=979
4	http://sinta2.ristekdikti.go.id/journals/detail?id=2518

Dari target *web scraping* jurnal kesehatan dalam bidang kesehatan di Indonesia kususnya dari kementrian kesehatan berikut Cuplikasi *coding scraping* data menggunakan bahasa pemograman *python* dapat dilihat pada gambar 8.

```
judul = list()
author = list()
afiliasi = list()

for i in page:
    print(i, end=" ")
    page = requests.get('http://sinta.ristekdikti.go.id/journals/detail?page='+str(i)+'&id=2518')
    soup = BeautifulSoup(page.text, 'html.parser')


    table = soup.find('tbody')
    tr_list = table.find_all("tr")


    for tr in tr_list:

        judul.append(tr.find("dt").get_text().replace("\n",""))
        author.append(tr.find("dd").get_text())
```

Gambar 8. *Coding Scraping*

Cuplikasi hasi *scraping* data menggunakan bahasa pemograman *python* dapat dilihat pada gambar 9.

No	judul	author	Nama	Afiliasi
1	Faktor-faktor yang mempengaruhi kejadian TB pa...	HSP Manalu	Jurnal Ekologi Kesehatan	Kementrian Kesehatan
2	Analisis kualitatif bakteri koliform pada depo...	DWIS Bali	Jurnal Ekologi Kesehatan	Kementrian Kesehatan
3	Prevalensi Cacing Pada Murid Sekolah Dasar Waj...	M Mardiana, D Djarismawati	Jurnal Ekologi Kesehatan	Kementrian Kesehatan
4	Pengamatan tempat perindukan Aedes aegypti pad...	H Hasyimi, M Soekirno	Jurnal Ekologi Kesehatan	Kementrian Kesehatan
5	Analisis kualitatif bakteri koliform pada depo...	NLPM Widiyanti, NP Ristiati	Jurnal Ekologi Kesehatan	Kementrian Kesehatan

Gambar 9. Hasil *Scraping*

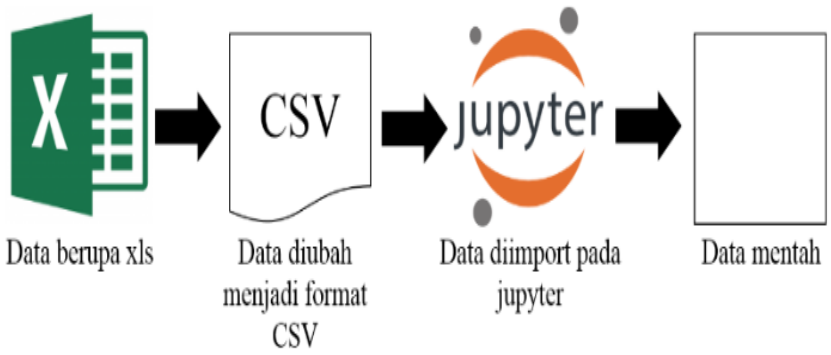
Dari *scraping* data tersebut diperoleh 5211 judul penelitian kesehatan di Indonesia yang di publikasikan di jurnal Sinta yang berafiliasi dari Kementrian Kesehatan yang dilakukan oleh para peneliti kesehatan. Berikut hasil rincian pengumpulan data dengan teknik *web scraping* dapat dilihat pada tabel 3.

Tabel 3. Hasil *Scraping*

Nama Jurnal	Jumlah Judul
Ekologi Kesehatan	378
Media Penelitian dan Pengembangan Kesehatan	972
Buletin Penelitian Kesehatan	2420
Buletin Penelitian Sistem Kesehatan	1441

3.3. Analisis

Pada tahap analisis ini yaitu dengan memvisualisasikan hasil data scraping yang di kumpulkan melalui jurnal SINTA adapun langkah-langkahnya dapat dilihat pada gambar 10.



Gambar 10. Load Data

Gambar tersebut merupakan tahap analisis data untuk dibaca ke dalam *tools python jupyter*. Data sebelumnya yaitu format xls kemudian data diubah menjadi csv. Adapun data yang *load* adalah data keseluruhan data hasil *scraping* pada jurnal SINTA.

Setelah *load data* selanjutnya memvisualisasikan topik penelitian kesehatan Indonesia tersebut kedalam bentuk *wordcloud*. Visualisai penelitian ini untuk menentukan topik memilih *wordcloud* karena visualiasai *wordcloud* kata-kata memiliki bobot frekuensi kemunculan yang tinggi dalam suatu topik yang dapat ditampilkan menonjol dari sisi ukuran sehingga lebih mudah diketahui dan intuitif. Berikut hasil visualisasi tren topik dapat dilihat pada gambar berikut.

Gambar 11. Topik 1

Gambar 12. Topik 2

Gambar 13. Topik 3

Gambar 14. Topik 4

Dari gambar tersebut terlihat terdapat beberapa topik yang menonjol memiliki kata kunci mirip/sama (redudanci) yang muncul. Sehingga untuk akurasi dan efisiensi kategori maka topik yang redudan akan dikelompokan menjadi satu klaster dapat dilihat pada tabel 4 dan 5.

Tabel 4. Tren Topik Penelitian

Topik	Kata Kunci
1	a. Air Minum
	b. Malaria
	c. aedes aegypti
2	a. Malaria
	b. Demam Berdarah
3	a. Malaria
	b. *aedes aegypti*
	c. Rumah Sakit
4	a. Pelayanan Kesehatan

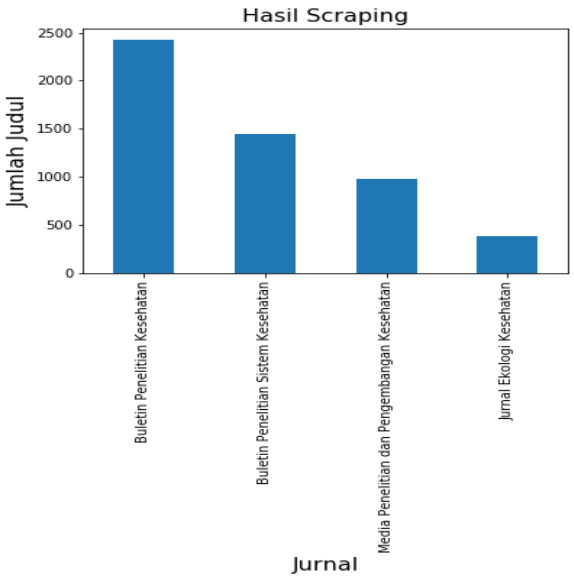
Adapun klaster Berdasarkan kemiripan makna dari kata-kata kunci yang muncul di setiap topik, maka topik tersebut dikelompokkan ke dalam klaster dapat dilihat pada tabel 5.

Tabel 5. Klusterisasi Topik

Klaster	Topik	Kata Kunci
Klaster 1	1, 2, 3	Malaria
Klaster 2	1,3	*aedes aegypti*
Klaster 3	Yang tidak ada kesamaan kata kunci	Air Minum, Demam Berdarah, Rumah Sakit, Pelayanan Kesehatan

Berdasarkan klusterisasi tren topik penlitian kesehatan di Indonesia yaitu Malaria, *aedes aegypti*, Air Minum, Demam Berdarah, Rumah Sakit, Pelayanan Kesehatan.

Adapun hasil visualisasi bentuk grafik jurnal yang di *scraping* berdasarkan jumlah judul masing-masing ke 4 jurnal afiliasi kementrian kesehatan dapat dilihat pada gambar 15 sebagai berikut:



Gambar 15. Visualisasi Jumlah Judul

Adapun hasil visualisasi bentuk *Wordcloud* untuk author yang aktif dalam penelitian kesehatan di Indonesia dapat dilihat pada gambar 16 sebagai berikut:

Gambar 16. Visualisasi Author

4. KESIMPULAN

Banyaknya laman yang menampilkan informasi mengenai jurnal kesehatan di Indonesia namun pada penelitian ini peneliti fokus pada Jurnal SINTA yang afiliasinya dari kementrian kesehatan Indonesia. Teknik *web scraping* akan memberikan dampak yang sangat bagus apabila diterapkan dalam sistem manajemen pengetahuan berbasis komputer. Teknik *web scraping* mengotomasi proses akusisi informasi khususnya informasi-informasi yang bersumber dari jurnal di internet.

Hal-hal yang dilakukan dalam penelitian ini meliputi empat hal. Hal pertama yang dilakukan adalah mengakses jurnal sinta untuk melakukan proses pengambilan informasi judul, author dan afiliasi. Hal yang kedua yaitu dilakukan pengumpulan data dengan menggunakan teknik web scraping untuk mengekstrak data dari jurnal SINTA. Hal yang ketiga ekstrasi informasi dari hasil pengumpulan data tersebut dengan cara menstrukturkan data yang sebelumnya tidak tersetruktur. selanjutnya menganalisis tren topik penelitian kesehatan di Indonesia.

Penelitian ini berhasil menyimpan otomatis data hasil *scraping* di Jurnal SINTA pada database. Dengan adanya penelitian ini, peneliti kesehatan dapat dengan mudah untuk mengetahui tren topik penelitian kesehatan di Indonesia. Selain itu dapat mengumpulkan informasi mengenai artikel/jurnal ilmiah

UCAPAN TERIMAKASIH

Terima kasih dan apresiasi setinggi-tingginya penulis sampaikan kepada pihak-pihak yang membantu terlaksananya penelitian ini. Yang pertama kepada Dhomas Hatta Fudholi, Ph.D Dosen Magister Teknik Informatika Fakultas Teknologi Industri Universitas Islam Indonesia Yogyakarta yang berkenan membimbing penelitian ini kususnya untuk proses pengambilan/*scraping* data, serta yang kedua kepada Bu Dwi Pamulatsih sebagai pakar tenaga kesehatan. Salain itu juga memberikan ucapan terima kasih kepada Dosen STIKES Al Islam Yogyakarta untuk memberikan review dan masukan atas penelitian yang dilakukan.

REFERENSI

[1] R. Roberts, N. Callow, L. Hardy, D. Markland, and J. Bringer, “Http://Sinta1.Ristekdikti.Go.Id ©2017,” pp. 200–221, 2017.

[2] L. Wall *et al.*, “About the Tutorial Copyright & Disclaimer,” p. 2, 2015.

[3] A. Josi, L. A. Abdillah, and Suryayusra, “Penerapan teknik web scraping pada mesin pencari artikel ilmiah,” 2014.

[4] K. Jarmul and R. Lawson, *Python Web Scrapping: Fetching Data From the Web*. 2017.

[5] A. Priyanto and M. R. Ma’arif, “Implementasi Web Scrapping dan

Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 25–33, 2018.

[6] M. S. Utomo, "Web Scraping pada Situs Wikipedia menggunakan Metode Ekspresi Regular," *J. Teknol. Inf. Din.*, vol. 18, no. 2, pp. 153–160, 2013.

[7] M. R. Ma'arif, "Integrasi Laman Web Tentang Pariwisata Daerah Istimewa Yogyakarta Memanfaatkan Teknologi Web Scraping Dan Text Mining," *Teknomatika*, vol. 9, no. 1, pp. 71–80, 2016.