UPDATE PILOT PROJECT

Web Scraping Terhadap Keyword Bioteknologi

Ikang Fadhli
Nutrifood Indonesia

12 November 2021

Contents

List of Figures

List of Tables

1 PENDAHULUAN

1.1 Latar Belakang

Komite Gama memiliki kebutuhan untuk melakukan inventarisasi terhadap:

1. Kegiatan pendidikan (formal atau non-formal). Termasuk program pendidikan yang telah dilakukan atau akan direncanakan.
2. Kegiatan penelitian (riset dan inovasi). Termasuk program pendidikan yang telah dilakukan atau akan direncanakan.

pada sektor **produksi pangan**, **rantai pasokan pangan**, dan **konsumsi pangan**.

1.2 Masalah

Informasi terkait kegiatan-kegiatan di atas tersebar di berbagai situs seperti pada portal:

1. Neliti[1].
2. BRIN[2].
3. SINTA[3].
4. dan berbagai situs lainnya.

Bagaimana cara kita bisa mendokumentasikan data kegiatan-kegiatan yang ada pada situs-situs tersebut?

1.3 Tujuan

Melakukan *big data mining* untuk mendokumentasikan kegiatan-kegiatan di situs-situs tersebut sekaligus membuat *data base* yang *reliable* terhadap topik terkait.

[1] https://www.neliti.com/
[2] https://www.brin.go.id/
[3] https://sinta.ristekbrin.go.id/

2 METODE

2.1 *Web Scraping*

Metode yang akan digunakan dalam *big data mining* kali ini adalah *web scraping*, yakni mengambil data yang terlihat secara visual pada suatu situs dengan cara melakukan *parsing html file*.

Untuk melakukan itu, saya membuat *custom algorithm* persitus yang dituju menggunakan bahasa pemrograman **R**.

2.2 *Data Carpentry*

Data yang diambil dari *web scraping* akan dibersihkan dan dibuat konsisten baik struktur dan formatnya. Data akan diekspor dalam bentuk tabel Microsoft Excel.

3 *PILOT PROJECT*

Saya akan mencoba mengambil beberapa data yang relevan terkait suatu *keyword* pada *pilot project* ini.

3.1 *Keyword* yang Digunakan

Keyword yang digunakan pada *pilot project* ini adalah **bioteknologi**.

3.2 Hasil *Big Data Mining*

Berikut adalah hasil penelusuran *keyword* **bioteknologi**.

3.2.1 Kegiatan Penelitian

Hasil penelusuran pada kegiatan penelitian di bidang **bioteknologi** saya dapatkan dengan cara mencari *keyword* tersebut di situs **neliti**[4] dan **SINTA**[5]. Semua jurnal yang muncul akan saya ambil informasinya sebagai bukti pelaksanaan riset dan inovasi terkait *keyword* tersebut.

3.2.1.1 Situs `neliti` Pada situs ini, saya spesifik akan mencari **jurnal** yang terkait dengan *keyword*. Sebagai contoh:


Journal article Bestari

Bioteknologi dan Dilema Rasionalitas Teknologi

Nurudin Nurudin

Diterbitkan 2001

Unduh teks lengkap | Buat referensi | Share | View original

Abstrak

The development of biotechnology has had been between two rationalities both technological pole and ethical ones. Meanwhile technological rationality becomes an orientation of its development and human life both individual and social have be dehumanized. It is necessary that its development directs to the ethical rationality

Figure 1: Contoh Halaman Depan dari Situs Jurnal Hasil Penelusuran di neliti

Informasi yang akan diambil dari halaman situs tersebut antara lain:

[4]https://www.neliti.com/
[5]https://sinta.ristekbrin.go.id/

1. Judul,
2. Penulis,
3. Tahun penerbitan,
4. Abstrak,
5. *Link* situs.

Pada saat saya melakukan *web scraping* pada `11 November 2021 17:20 WIB`, tercatat ada 195 buah jurnal di situs `neliti`.

Data hasil *web scraping* terlampir.

3.2.1.2 Situs `SINTA` Pada situs SINTA, saya akan spesifik mencari jurnal yang terkait *keyword*. Berikut adalah hasil pencariannya:

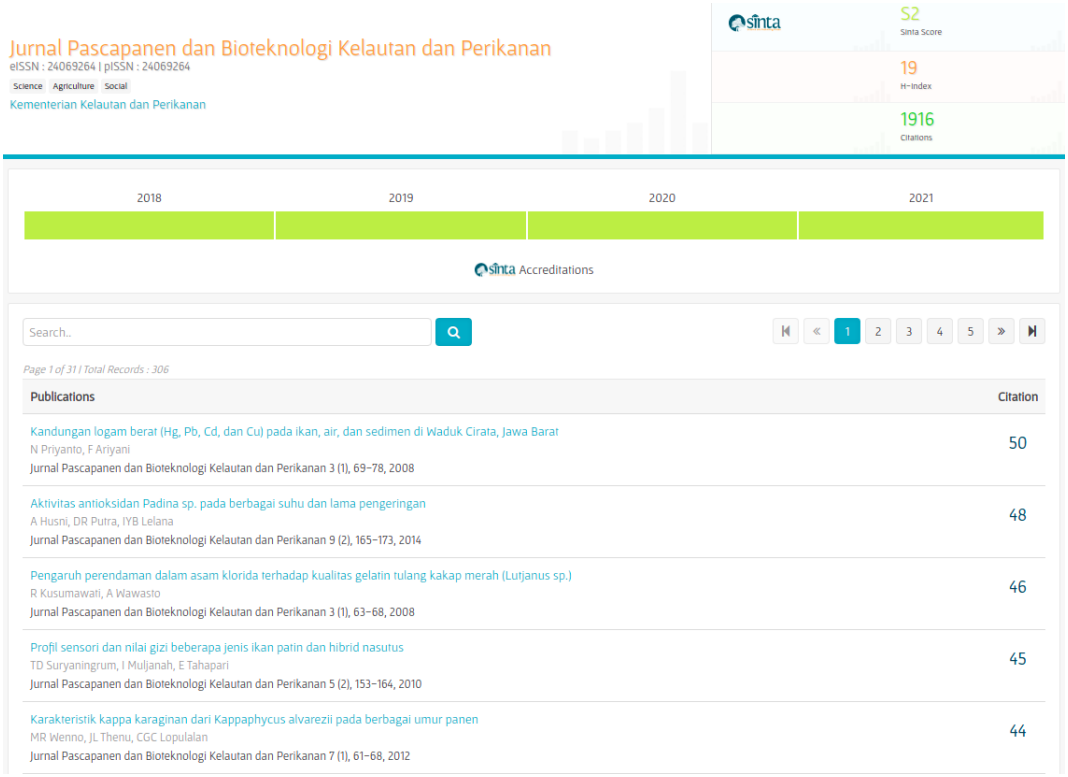

Figure 2: Contoh Halaman Depan dari Situs SINTA Hasil Penelusuran

Informasi yang akan diambil dari halaman situs tersebut antara lain:

1. Judul,
2. Penulis,
3. Nama jurnal,
4. Berapa banyak tersitasi,
5. *Link* situs.

Pada saat saya melakukan *web scraping* pada `11 November 2021 21:57 WIB`, tercatat ada 443 buah jurnal di situs `SINTA`.

Data hasil *web scraping* terlampir.

3.2.2 Kegiatan Lainnya

Salah satu kesulitan yang dihadapi adalah pada saat mencari informasi seputar kegiatan pendidikan formal atau non formal (pelatihan). Sebagai iterasi pertama, saya akan coba menelusuri *keyword* di situs `BRIN` dan mengambil semua artikel atau berita terkait.

3.2.2.1 Situs `BRIN` Berikut adalah hasil penelusuran *keyword* di situs `BRIN`:

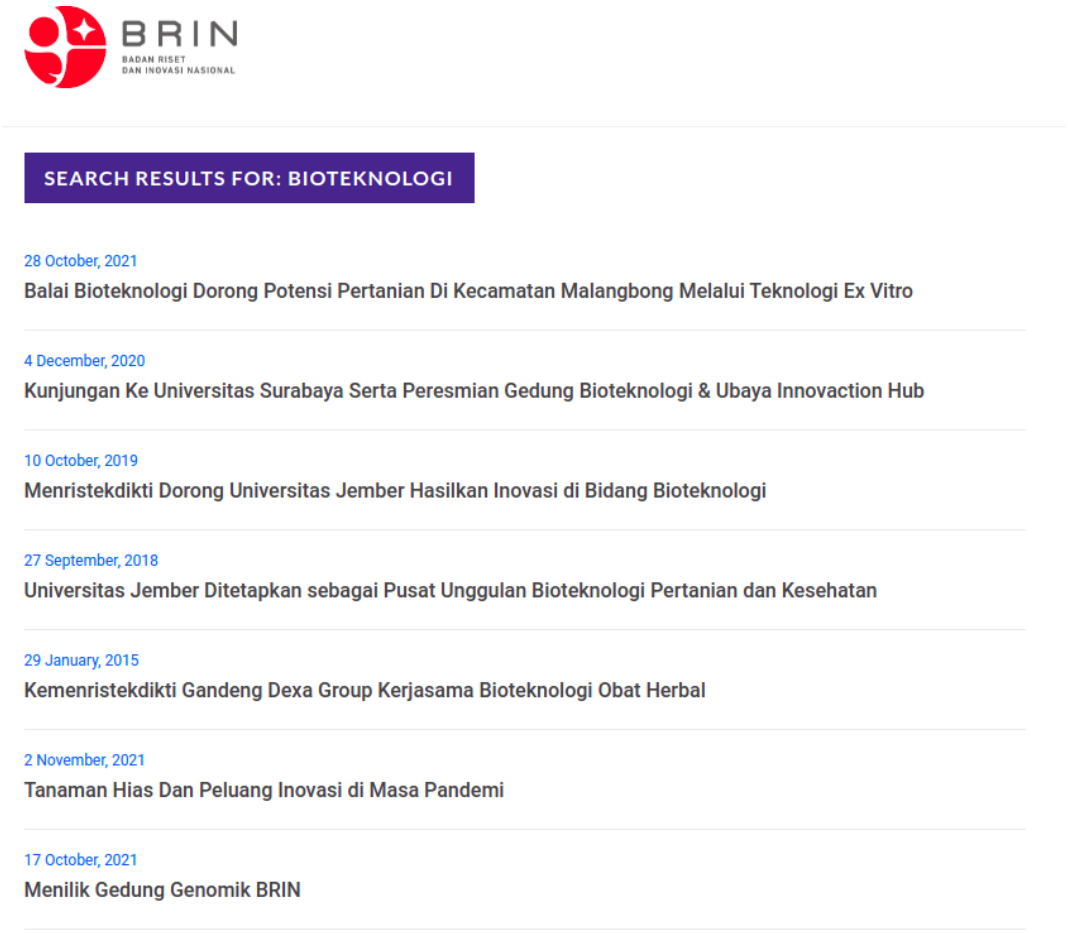

Figure 3: Contoh Halaman Depan dari Situs BRIN Hasil Penelusuran

Informasi yang akan saya ambil adalah:

1. Judul artikel,
2. Tanggal artikel,

3. Isi artikel.

Kelak kita akan coba pilah, apakah ada unsur pendidikan formal, **pelatihan** atau ***training*** dari artikel tersebut.

Pada saat saya melakukan *web scraping* pada `11 November 2021 22:21 WIB`, tercatat ada 88 buah artikel di situs `BRIN`.

Dari data tersebut, akan saya beri tanda mana saja artikel yang memiliki kata `pelatihan`, `pendidikan`, dan `training`.

Data hasil *web scraping* terlampir.