PYTHON WEB SCRAPING

PYTHON WEB SCRAPING

Rolly M. Awangga
Informatics Research Center

Kreatif Industri Nusantara

Penulis:
Rolly Maulana Awangga

ISBN : 978-602-53897-0-2

Editor:
M. Yusril Helmi Setyawan

Penyunting:
Syafrial Fachrie Pane
Khaera Tunnisa
Diana Asri Wijayanti

Desain sampul dan Tata letak:
Deza Martha Akbar

Penerbit:
Kreatif Industri Nusantara

Redaksi:
Jl. Ligar Nyawang No. 2
Bandung 40191
Tel. 022 2045-8529
Email : awangga@kreatif.co.id

Distributor:
Informatics Research Center
Jl. Sariasih No. 54
Bandung 40151
Email : irc@poltekpos.ac.id

Cetakan Pertama, 2019

'Jika Kamu tidak dapat
menahan lelahnya
belajar, Maka kamu harus
sanggup menahan
perihnya Kebodohan.'
Imam Syafi'i

CONTRIBUTORS

ROLLY MAULANA AWANGGA, Informatics Research Center., Politeknik Pos Indonesia, Bandung, Indonesia

CONTENTS IN BRIEF

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

Listings

FOREWORD

Sepatah kata dari Kaprodi, Kabag Kemahasiswaan dan Mahasiswa

KATA PENGANTAR

Buku ini diciptakan bagi yang awam dengan git sekalipun.

R. M. AWANGGA

Bandung, Jawa Barat
Februari, 2019

ACKNOWLEDGMENTS

Terima kasih atas semua masukan dari para mahasiswa agar bisa membuat buku ini lebih baik dan lebih mudah dimengerti.

Terima kasih ini juga ditujukan khusus untuk team IRC yang telah fokus untuk belajar dan memahami bagaimana buku ini mendampingi proses Intership.

R. M. A.

ACRONYMS

ACGIH	American Conference of Governmental Industrial Hygienists
AEC	Atomic Energy Commission
OSHA	Occupational Health and Safety Commission
SAMA	Scientific Apparatus Makers Association

GLOSSARY

git	Merupakan manajemen sumber kode yang dibuat oleh linus torvald.
bash	Merupakan bahasa sistem operasi berbasiskan *NIX.
linux	Sistem operasi berbasis sumber kode terbuka yang dibuat oleh Linus Torvald

SYMBOLS

A Amplitude

& Propositional logic symbol

a Filter Coefficient

$\mathcal{B}$ Number of Beats

INTRODUCTION

ROLLY MAULANA AWANGGA, S.T., M.T.

Informatics Research Center
Bandung, Jawa Barat, Indonesia

Pada era disruptif saat ini. git merupakan sebuah kebutuhan dalam sebuah organisasi pengembangan perangkat lunak. Buku ini diharapkan bisa menjadi penghantar para programmer, analis, IT Operation dan Project Manajer. Dalam melakukan implementasi git pada diri dan organisasinya.

Rumusnya cuman sebagai contoh aja biar keren[1].

$$ABC\mathcal{D}\mathcal{E}\mathcal{F}\alpha\beta\Gamma\Delta\sum_{def}^{abc} \tag{I.1}$$

BAB 1

PERMULAAN

Pada bab ini, kita akan menjalankan program Python yang pertama yaitu, hello_world.py. Tahap pertama, kita akan memastikan apakah Python sudah terinstall dengan benar. Kita juga akan menginstall teks editor untuk menulis program Python.

1.1 Menyiapkan *Programming Environment*

Di sini, kita akan melihat dua versi utama Python yang saat ini digunakan dan menguraikan langkah-langkah untuk menyiapkan Python pada sistem.

1.1.1 Python 2 dan Python 3

Saat ini, telah tersedia dua versi Python: Python 2 dan Python yang lebih baru, yaitu Python 3. Mengapa? Karena setiap bahasa pemrograman berevolusi ketika ide dan teknologi baru muncul atau berkembang, dan tentu saja para pengembang bahasa Python terus membuat Python agar lebih fleksibel dan kuat. Sebagian besar perubahan yang dilakukan, berkembang sedikit demi sedikit secara teratur dan hampir tidak

terlihat, tetapi dalam beberapa kasus kode yang ditulis untuk Python 2 mungkin tidak berjalan dengan baik pada sistem yang menggunakan Python 3.

1.1.2 Menjalankan Kode

Python dilengkapi dengan *interpreter* yang berjalan di jendela terminal, yang memungkinkan kita untuk mencoba kode Python tanpa harus menyimpan dan menjalankan seluruh program. Contohnya adalah seperti pada

Listing 1.1 Kode Pada Jendela Terminal

```
>>> print("Hello World!")
Hello World!
```

Pada baris pertama, adalah baris yang kita tuliskan sendiri perintahnya, lalu bisa dieksekusi dengan cara menekan tombol enter. Sebagian besar contoh yang akan ditampilkan pada buku ini akan dijalankan melalui teks editor, karena kita akan menulis kode pada teks editor tersebut. Setiap kali Anda melihat tiga buah tanda tanda panah sebelah kiri seperti pada listing 1.1 itu artinya kita sedang menggunakan jendela terminal. Listing 1.1 menunjukan program sederhana yang umum dilakukan *programmer* pada awal pembelajaran. Jika kode berjalan dengan sesuai, maka apapun program yang telah dibuat menggunakan Python bisa berjalan dengan sempurna.

1.1.3 Python pada Berbagai Sistem Operasi

Python adalah bahasa pemrograman lintas platform, yang artinya berjalan pada semua sistem operasi, seperti Windows, Linux, dan OS X. Program Python apa pun bisa dijalankan pada seluruh sitem yang telah terpasang Python sebelumnya. Namun, cara untuk mengatur Python pada setiap sistem operasi akan berbeda. Di bagian ini kita akan belajar cara men-*set up* Python dan menjalankan program Hello World pada setiap sistem operasi.Yang pertama adalah, kita akan memeriksa apakah Python telah ter-*install* dengan benar atau belum. Kemudian, menginstal teks editor sederhana dan menyimpan file Python kosong bernama hello_world.py. Tahap terakhir, yaitu menjalankan program Hello World.

1.1.3.1 Python pada Windows Pertama, periksa apakah Python telah terpasang pada sistem atau belum, dengan cara buka *Command Prompt* dengan memasukkan atau mengetikan kata "cmd" pada menu Start. Pada jendela terminal, ketikan python dalam huruf kecil. Jika mendapatkan prompt Python (`>>>`), Python telah terpasang pada sistem. Namun, jika melihat pesan kesalahan yang mengatakan bahwa python adalah bukan perintah yang dikenali. Jika demikian, unduh Python *Installer* untuk Windows. Buka `http://python.org/downloads/`. Akan tersedia dua pilihanl, satu untuk mengunduh Python 3 dan satu lagi untuk mengunduh Python 2. Klik tombol Python 3, lalu secara otomatis mulai mengunduh *Installer*. Setelah *Installer* berhasil terunduh, jalankan *Installer*. Pastikan mencentang opsi *Add Python to PATH*, agar lebih mudah untuk mengkonfigurasi sistem dengan benar.

Sekarang kita akan memulai untuk mencoba apakah Python sudah terpasang dengan benar atau belum, dengan cara membuka *Command Prompt* lalu mengetik **python**. Jika mendapatkan prompt Python (>>>), Python telah terpasang pada sistem dengan berhasil.

Listing 1.2 Python Yang Telah Terpasang

```
C:\> python
Python 3.5.0 (v3.5.0:374f501f4567,
Mar 19 2019, 22:15:05)
[MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license"
for more information.
>>>
```

1.1.4 Mengatasi Masalah Umum Pada Python

Jika tidak dapat menjalankan hello_world.py, berikut adalah beberapa solusi yang dapat gunakan atau dapat dicoba:

- Ketika sebuah program mengandung kesalahan yang cukup besar, Python akan selalu menampilkan *traceback*. Python melihat melalui file dan mencoba melaporkan atau memberitahu masalahnya. *Traceback* bisa memberi petunjuk tentang masalah apa yang menghambat program berjalan.
- Ingat bahwa sintaks sangat penting dalam pemrograman, tanda kutip yang tidak cocok, atau tanda kurung yang tidak cocok dapat menghambat program berjalan.
- Mulai lagi dari awal. tidak perlu men-*uninstall* apa pun, tetapi cobal untuk menghapus file hello_world.py dan membuatnya lagi dari awal.
- Minta orang lain untuk mengikuti langkah-langkah dalam bab ini, bisa di komputer mana saja, dan perhatikan apa yang dilakukan dengan hati-hati.
- Temui orang yang cukup familiar dengan Python dan minta mereka membantu memecahkan masalahnya.

BAB 2

MENGENAL WEB SCRAPING

2.1 Web Scraping

2.1.1 Apa Itu Web Scraping?

Secara teori, *web scraping* adalah praktik mengumpulkan data melalui cara apa pun selain program yang berinteraksi dengan API (atau melalui manusia menggunakan browser). Cara ini paling umum dilakukan dengan menulis program otomatis yang membutuhkan server web, meminta data (biasanya dalam bentuk HTML, Json, dan file lain yang terdiri dari halaman web), dan kemudian mem-*parsing* data tersebut untuk mengekstrak informasi yang diperlukan. Sedangkan pada praktiknya,*web scraping* mencakup beragam teknik dan teknologi pemrograman, seperti analisis data dan keamanan informasi.

2.1.2 Mengapa Web Scraping?

Ada banyak alasan mengapa web scraping semakin dibutuhkan pada abad ke 21 ini. Dengan semakin berkembangnya data, jumlah data yang tersedia mungkin sudah tidak terhitung lagi. Bayangkan jika kita membutuhkan data-data itu lalu Anda harus

mengumpulkan dan menyimpan jutaan data dalam satu file. Teknik web scraping bisa membantu kita untuk mengumpulkan data dengan lebih cepat dan otomatis, semua akan berjalan lancar selama server masih berfungsi.

Efisiennya teknik web scraping ini juga membantu proses pengambilan data demi kebutuhan analisa. Karena web scraping membantu mengumpulkan semua data tanpa terlewat. Dengan begitu, Anda akan mendapat insight yang bernilai dengan lebih cepat. Kita juga bisa memanfaatkan web scraping untuk mengumpulkan data lain yang penting.

Selain di dunia bisnis, di dunia seni pun, web scraping telah diterapkan untuk proyek 2006 "We Feel Fine" oleh Jonathan Harris dan Sep Kamvar, men-*scrap* berbagai situs blog berbahasa Inggris untuk frasa yang dimulai dengan "I feel" atau "I am feeling" dan dengan data tersebut, bisa diolah menjadi visualisasi data bagaimana perasaan orang-orang di dunia setiap harinya. Terlepas dari bidang apapun, hampir selalu ada cara*web scraping* dapat memandu bisnis lebih efektif dan meningkatkan produktivitas.

2.1.3 Teknik web scraping

Karena web scraping sudah mulai familiar, maka banyak orang yang melakukannya karena beberapa kemudahan yang telah dijabarkan diatas, ada beberapa teknik automasi yang bisa kita lakukan untuk melakukan web scraping.

1. Parsing HTML

   Parsing HTML adalah salah satu teknik yang paling banyak digunakan dalam web parsing atau web scraping. Biasanya parsing HTML dilakukan menggunakan JavaScript dan menarget halaman HTML linear dan nested. Metode ini dapat mengidentifikasi script HTML dari websia. Script ini juga kemudian digunakan untuk mengekstraksi text, links, dan data.

2. Parsing DOM

   Content, style, dan XML structure didefinisikan dalam Document Object Model atau yang biasa disebut DOM. Beberapa programmer yang ingin mengetahui cara kerja sebuah internal web dan ingin mengekstrak *script* yang berjalan di dalamnya akan lebih memilih untuk melakukan *web scraping* melalui parsing DOM. Node dikumpulkan terlebih dahulu menggunakan DOM yang terlah diparsing dan XPath membantu mempermudah proses scraping.

3. XPath

   XML Path Language atau lebih familiar dengan istilah XPath adalah bahasa *query* yang bekerja pada dokumen Extensible Markup Language atau biasa disebut XML. Dokumen XML biasa disusun dengan *tree structure*, XPath bisa digunakan untuk menganalisa struktur dokumen dengan memilih nodes berdasarkan parameter yang telah tersedia. XPath juga lazim digunakan bersamaan dengan DOM parsing dalam mengesktrasi seluruh halaman website.

4. Google Docs

Salah satu produk Google yaitu Google Sheets juga ternyata bisa dipakai sebagai salah satu alat scraping,, dan ini adalah salah satu alat scraping yang cukup popoler karena cara penggunaannya yang tidak rumit dan tidak membutuhkan keahlian khusus. Pada Google Sheets sendiri, kita bisa memanfaatkan fungsi IMPORTXML untuk melakukan scraping data dari website. Selain itu, juga bisa menggunakan command ini untuk melihat apakah sebuah website aman dari tindakan scraping atau tidak.

2.1.4 Resiko dan Ancaman

Pada perspektif Bisnis, otomatisasi web seperti web scraping, juga memiliki dampak negatif. Yang berpengaruh disini adalah reputasi perusahaan, SOP, dan proses bisnis internal. Beberapa resiko yang mungkin akan timbul adalah:

1. Data statistik: Setiap *request* yang dilakukan oleh robot sangat kecil kemungkinannya akan tercatat pada laporan statistik, sehingga akan menyebabkan data analisis tersebut akan menjadi bias atau tidak akurat. Dan dengan ketidak akuratannya data statistik tersebut, tim bisnis marketing akan beresiko salah dalam contohnya menganalisa pasar, dan mengambil keputusan bisnis.

2. Bulk Order: Dengan otomatis, web robot memungkinkan untuk membuat random dan distributed order fiktif. Tujuannya bisa beragam, mungkin ingin merusak bisnis kompetitor dengan menghabiskan stock barang yang dimilikinya, atau hanya sekedar seseorang yang mengambil barang promosi dengan jumlah sangat banyak, dan menjuanya kembali dengan harga normal.

2.1.5 Pertahanan

Pertahanan yang paling efektif dalam menaggulangi bot adalah memblokir IP agar tidak dapat mengakses website lagi, atau bisa dengan cara me-*redirect*-nya ke halaman captcha, seperti yang dilakukan google.com apabila mereka mencurigai suatu IP Penggunaan Captcha sangat efektif dalam mendeteksi apakah yang mengakses website adalah user sesungguhnya atau hanya robot.

BAB 3

MEMBANGUN WEB SCRAPER

BAB 4

MEMBANGUN WEB SCRAPER

4.0.1 Struktur HTML

Sebelum kita melompat pada kode, mari kita bahas terlebih dahulu hal-hal mendasar dari HTML, dan beberapa aturan dalam *web scraping*

1. HTML TAG

Algorithm 4.1

```
<!DOCTYPE html>
<html>
    <head>
    </head>
    <body>
        <h1> First Scraping </h1>
        <p> Hello World </p>
    <body>
</html>
```

(a) `<! DOCTYPE html>`: Dokumen HTML harus dimulai dengan deklarasi tipe.

(b) Dokumen HTML terdapat di antara `<html>` dan `</html>`.

(c) Deklarasi meta dan skrip dari dokumen HTML berada di antara `<head>` dan `</head>`.

(d) Bagian yang terlihat dari dokumen HTML adalah antara tag `<body>` dan `</body>`.

(e) Judul judul didefinisikan dengan tag `<h1>` hingga `<h6>`.

(f) Paragraf didefinisikan dengan tag `<p>`.

Tag berguna lainnya termasuk `<a>` untuk hyperlink, `<table>` untuk tabel, `<tr>` untuk baris tabel, dan `<td>` untuk kolom tabel. Juga, tag HTML kadang-kadang datang dengan atribut id atau kelas. Atribut id menentukan id unik untuk tag HTML dan nilainya harus unik dalam dokumen HTML. Atribut kelas digunakan untuk menentukan gaya yang sama untuk tag HTML dengan kelas yang sama. Kita dapat menggunakan id dan kelas ini untuk membantu kita menemukan data yang kita inginkan.

2. Aturan Scraping

(a) Periksa terllebih dahulu Syarat dan Ketentuan situs web tujuan sebelum melakukan *scraping*. Berhati-hatilah untuk membaca pernyataan tentang penggunaan data secara legal.

(b) Jangan meminta data dari situs web terlalu agresif atau terlalu sering dengan program yang telah dibuat (juga dikenal sebagai spam), karena ini dapat merusak situs web. Pastikan program berperilaku dengan cara yang masuk akal (mis. Bertindak seperti manusia). Satu request untuk satu halaman web per detik adalah yang paling disarankan.

(c) *layout* situs web dapat berubah dari waktu ke waktu, jadi pastikan untuk mengunjungi kembali situs dan menulis ulang kode sesuai kebutuhan.

4.0.2 BeautifulSoup

Beautiful Soup adalah *library* Python untuk menarik atau mengambil data dari file HTML dan XML. Hal ini bisa bekerja dengan parser apapun untuk menavigasi, mencari, dan memodifikasi pohon parse. Ini biasanya menghemat waktu menulis kode dan hari kerja. BeautifulSoup juga mencoba memahami yang tidak masuk akal; sedikit membantu memformat ulang dan mengatur web yang berantakan dengan memperbaiki HTML yang kacau dan menyajikan kepada kita struktur dalam Python yang mudah dibaca dan mewakili struktur XML. **Menginstal BeautifulSoup** Karena *library* BeautifulSoup bukan Python *library* default, maka harus terlebih dahulu diinstal. Kita akan menggunakan *library* BeautifulSoup versi 4 (BS4).

untuk Linux:

$ sudo apt-get install python-bs4

dan untuk Mac:

$ sudo easy_install pip

Menginstal *package* di Windows hampir sama dengan proses untuk Mac dan Linux. Unduh rilisan terakhir BeautifulSoup 4, navigasikan ke direktori tempat megekstrak lalu jalankan:

instal setup.py python

BeautifulSoup sekarang akan dikenali sebagai *library* Python di sistem. Kita masih bisa menguji ini dengan membuka terminal Python dan mengimportnya:

Algorithm 4.2

```
$ python
bs4 import from BeautifulSoup
```

Menjalankan BeautifulSoup Objek yang paling umum digunakan denga library BeautifulSoup adalah, objek BeautifulSoup itu sendiri.

Listing 4.1 Contoh Sederhana

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen("http://examplesite.com/")
bsObj = BeautifulSoup(html.read())
print(bsObj.h1)
```

outputnya akan menjadi:

`<h1>Hello<h1>`

Seperti pada listing 4.1, mengimpor library urlopen dan memanggil `html.read ()` untuk mendapatkan konten HTML dari halaman tersebut. Konten HTML ini kemudian diubah menjadi objek BeautifulSoup, dengan struktur berikut:

```
html-><html><head>...</head><body>...</body></html>
head-><head><title>Contoh..<title></head>
title-><title>Halaman Contoh</title>
body-><body><h1>contoh</h1><div>Ini Halam...</div></body>
h1 -><h1>Hello</h1>
div-><div>Ini Halaman contoh</div>
```

Perhatikan bahwa tag `<h1>` yang telah diekstrak dari halaman website dua lapisan jauh ke dalam struktur objek BeautifulSoup `(html -> body -> h1)`. Namun, jika hanya ingin mengambil dari objek secara langsung, bisa memanggil tag h1 dengan cara:

```
bsObj.h1
```

Atau bisa juga dengan salah satu cara di bawah ini

```
bsObj.html.body.h1
bsObj.body.h1
```

```
bsObj.html.h1
```

Pada bab selanjutnya, kita akan mempelajari lebih dalam tentang fungsi-fungsi dari beautifulsoup, semoga dengan sedikit contoh pada bab ini akan memberi gambaran untuk bab-bab selanjutnya.

BAB 5

BEAUTIFULSOUP

Find () dan findAll () adalah dua fungsi BeautifulSoup yang akan sering digunakan. Dengan kedua fungsi ini, kita dapat dengan mudah mem-*filter* halaman HTML untuk menemukan daftar tag yang diinginkan, atau satu tag, berdasarkan berbagai atribut mereka. Kedua fungsi ini sangat mirip:

```
findAll (tag, atribut, rekursif, teks, batas, kata kunci)
find (tag, atribut, rekursif, teks, kata kunci)
```

Dalam banyak kesempatan, 95% dari waktu kita saat menulis kode, kita akan menggunakan dua argumen pertama, yaitu tag dan attribut. Tag digunakan untuk mendeklarasikan sebuah objek. Misalnya, yang berikut ini akan mengembalikan daftar semua argumen tag dalam dokumen: 1

```
.findAll ({"h1", "h2", "h3", "h4", "h5", "h6"})
```

Argumen atribut mengambil library Python dari atribut dan mencocokkan tag yang ada padai salah satu dari atribut tersebut. Misalnya, fungsi berikut akan me-*return* tag hijau dan merah di dokumen HTML:

```
.findAll ("span", {"class": "green", "class": "red"})
```

Argumen rekursif adalah boolean. Seberapa dalam kita ingin masuk ke dalam sebuah struktur HTML? Jika rekursif diset ke True, fungsi findAll melihat ke *children*, dan *children's children*, untuk tag yang cocok dengan parameter. Jika itu salah, ia

hanya akan melihat tag paling atas dalam dokumen. Secara default, findAll bekerja secara rekursif (rekursif diset ke True);

DAFTAR PUSTAKA

1. R. Awangga, “Sampeu: Servicing web map tile service over web map service to increase computation performance,” in *IOP Conference Series: Earth and Environmental Science*, vol. 145, no. 1. IOP Publishing, 2018, p. 012057.

Index