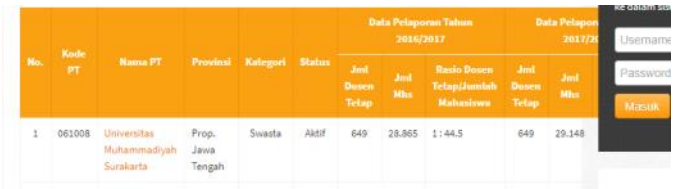
Web Scrap

10 May 2018 13:47

Web scrap adalah proses mengekstrak data dari sebuah website, menggunakan program

Studi kasus (tugas terakhir):
Mengekstrak data profile mahasiswa dari website
http://forlap.ristekdikti.go.id
Karakteristik website forlap:

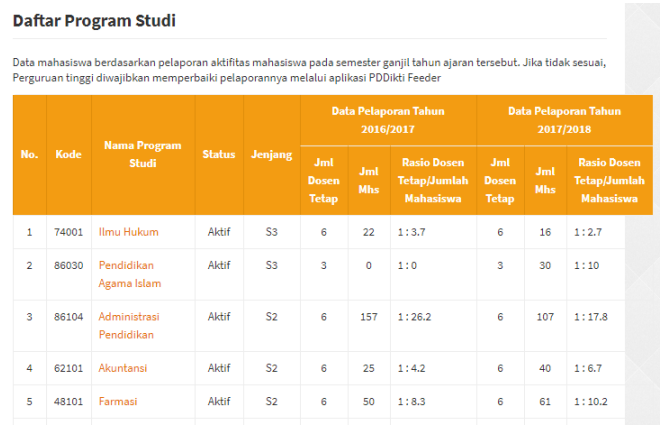
1. Link yang mengarah ke profile mahasiswa bersifat konstan/tidak berubah, selama NIM dari mahasiswa tidak berubah.
2. Link ke arah profile mahasiswa dapat dicari berawal dari link universitas.
3. Link Universitas di dalamnya mengandung link ke semua program studi (fakultas).



No.	Kode PT	Nama PT	Provinsi	Kategori	Status	Data Pelaporan Tahun 2016/2017			Data Pelapor 2017/2	
						Jml Dosen Tetap	Jml Mhs	Rasio Dosen Tetap/Jumlah Mahasiswa	Jml Dosen Tetap	Jml Mhs
1	061008	Universitas Muhammadiyah Surakarta	Prop. Jawa Tengah	Swasta	Aktif	649	28.865	1 : 44.5	649	29.148

Username
Password
Masuk

https://forlap.ristekdikti.go.id/perguruantinggi/detail/NkQxMjQxNDItRTc5OC00RjYyLTg3NEItQ0U0MzVCNTQwOUYx
Contoh daftar program studi yang terdpat di link forlap UMS (tidak semua ditampilkan)



Daftar Program Studi

Data mahasiswa berdasarkan pelaporan aktifitas mahasiswa pada semester ganjil tahun ajaran tersebut. Jika tidak sesuai, Perguruan tinggi diwajibkan memperbaiki pelaporannya melalui aplikasi PDDikti Feeder

No.	Kode	Nama Program Studi	Status	Jenjang	Data Pelaporan Tahun 2016/2017			Data Pelaporan Tahun 2017/2018		
					Jml Dosen Tetap	Jml Mhs	Rasio Dosen Tetap/Jumlah Mahasiswa	Jml Dosen Tetap	Jml Mhs	Rasio Dosen Tetap/Jumlah Mahasiswa
1	74001	Ilmu Hukum	Aktif	S3	6	22	1 : 3.7	6	16	1 : 2.7
2	86030	Pendidikan Agama Islam	Aktif	S3	3	0	1 : 0	3	30	1 : 10
3	86104	Administrasi Pendidikan	Aktif	S2	6	157	1 : 26.2	6	107	1 : 17.8
4	62101	Akuntansi	Aktif	S2	6	25	1 : 4.2	6	40	1 : 6.7
5	48101	Farmasi	Aktif	S2	6	50	1 : 8.3	6	61	1 : 10.2

4. Di dalam link program studi terdaftar daftar link profile mahasiswa dan profile dosen.
   Contoh:

Dasar-dasr HTML:

Struktur dasar HTML:

```
<!DOCTYPE html>
<html>
<head>
<!-- meta data -->
<title>Web Scrap</title>


<link rel="stylesheet" href="/w3css/4/w3.css">

<style>

#mytabel {
   background: yellow;
}



</style>

</head>
<body>
<!-- isi dokumen html di sini -->
<h3>Belajar web scrap</h3>
<p> <strong>Paragrap baru</strong>, <br>
untuk ganti baris paksa</p>

<h1>Structur TABEL dalam dokumen HTML</h1>

<table id="mytabel" class="list">
<tr>
     <td>NIM</td><td>Nama</td>
</tr>
</table>


</body>
```

```
</html>
```

Program web scrape menggunakan Python:
Referensi uptodate:
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Buku dapat didownload dari http://gen.lib.rus.ec
Key word: "Getting Started with Beautiful Soup"

Installasi beatifull soup:

>pip install beautifulsoup4

Download tool untuk parser (xml dan html)

```
>pip install lxml       ==> untuk parser format XML
>pip install html5lib   ==> untk parser format html
```

>pip install urllib3 : untuk request dokument html dari link

Contoh menggunakan beautiful soup:
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Contoh data :

```
  </span>
 </div>
 <table class="table1">
  <tr class="text-warning bold">
   <td class="tright" width="200px">
    Status PT
   </td>
   <td class="tcenter" width="30px">
    :
   </td>
   <td>
    Aktif
   </td>
  </tr>
  <tr>
   <td class="tright">
    Perguruan Tinggi
   </td>
   <td class="tcenter">
    :
   </td>
   <td>
    Universitas Muhammadiyah Surakarta
   </td>
  </tr>
  <tr>
   <td class="tright">
    Tanggal Berdiri
   </td>
```

Untuk mengambil data nama perguruan tinggi: lokasinya adalah terletak dalam tabel dengan class='table1', baris ke dua (<tr>), kolom ke tiga (<td>)