ScrapingForlap

24 May 2018 20:31

Installasi beatifull soup:

```
>pip install beautifulsoup4
```

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

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 SCRAPING Forlap UMS:

```
UMS_Forlap_link =
'https://forlap.ristekdikti.go.id/perguruantinggi/detail/NkQxMjQxNDItRTc5OC00RjYyLTg3NEItQ0U0MzVCNTQwOUYx'

from bs4 import BeautifulSoup4 as bs
import urllib3
L = urllib3.PoolManager()
html = L.request('GET', UMS_Forlap_link)

if html.status == 200:
      Ums_html = bs(html.data, 'html.parser')
      print( ums_html.pritify() )
```

Contoh output dari variabel "ums_html" :

```
  </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>)