Adatkinyerés weblapokból

Gergi Miklós
BME Matematika Intézet

2008. szeptember 17.

Mivel csináljuk?

- grep, tr, sed, awk

Mivel csináljuk?

- grep, tr, sed, awk
- HTML::Parser, HTML::TreeBuilder, and HTML::Element

Mivel csináljuk?

- grep, tr, sed, awk
- HTML::Parser, HTML::TreeBuilder, and HTML::Element
- Ruby Hpricot

Hpricot

- XPath selector
- CSS selector
- jQuery

MNB – html

```
<tr class="MNBInternet_FS MNBInternet_DgRowBG2 MNBInternet_DgText2">
  <td class="MNBInternet_DgBorder" align="Center">
    <span class="MNBInternet_Heading4">CHF</span></td>
  <td align="center" class="MNBInternet_DgBorder">svájci frank</td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">1</span></td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">149,00 </span></td>
  <td class="MNBInternet_DgBorder"> </td>
  <td class="MNBInternet_DgBorder" align="Center">
    <span class="MNBInternet_Heading4">NZD</span></td>
  <td align="center" class="MNBInternet_DgBorder">új-z. dollár</td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">1</span></td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">111,51 </span></td>
</tr>
```

MNB – html

```
<tr class="MNBInternet_FS MNBInternet_DgRowBG2 MNBInternet_DgText2">
  <td class="MNBInternet_DgBorder" align="Center">
    <span class="MNBInternet_Heading4"> CHF </span></td>
  <td align="center" class="MNBInternet_DgBorder">svájci frank</td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">1</span></td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5"> 149,00  </span></td>
  <td class="MNBInternet_DgBorder"> </td>
  <td class="MNBInternet_DgBorder" align="Center">
    <span class="MNBInternet_Heading4">NZD</span></td>
  <td align="center" class="MNBInternet_DgBorder">új-z. dollár</td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">1</span></td>
  <td class="MNBInternet_DgBorder" align="Right">
    <span class="MNBInternet_Heading5">111,51 </span></td>
</tr>
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
puts mnb.at("span[text()='CHF']").html.strip
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
puts mnb.at("span[text()='CHF']/../..").html.strip
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
puts mnb.at("span[text()='CHF']/../../td[4]").html.strip
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
puts mnb.at("span[text()='CHF']/../../td[4]/span").html.strip
```

MNB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

mnb = Hpricot(open('http://mnb.hu/engine.aspx?page=napiarfolyamok'))
puts mnb.at("span[text()='CHF']/../../td[4]/span").html.strip
puts mnb.at("span[text()='NZD']/..~td~td~td/span").html.strip
```

IMDB – html

```
<table class="cast">
  <tr class="odd">
    <td class="hs">
      <a href="/name/nm0000136/" ... >
        <img src= ... >
      </a>
      <br>
    </td>
    <td class="nm">
      <a href="/name/nm0000136/">Johnny Depp</a>
    </td>
    <td class="ddd"> ...  </td>
    <td class="char">
      <a href="/character/ch0001274/">Jack Sparrow</a>
    </td>
  </tr>
  <tr class="even">
... ... ...
```

IMDB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

doc = Hpricot(open("http://www.imdb.com/title/tt0325980/"))
```

IMDB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

doc = Hpricot(open("http://www.imdb.com/title/tt0325980/"))
doc.search("table[@class='cast']//td[@class='nm']")
```

IMDB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

doc = Hpricot(open("http://www.imdb.com/title/tt0325980/"))
doc.search("table.cast//td.nm")
```

IMDB – ruby

```
#!/usr/bin/ruby
require 'hpricot'
require 'open-uri'

doc = Hpricot(open("http://www.imdb.com/title/tt0325980/"))
doc.search("table.cast//td.nm").map{|x| puts x.at("a").html}
```