

R-Ladies lightning talks

Budapest, 2017. 12. 12.

**Webscraping – egy egyszerű példán
keresztül**

Készítette: Téglás Szilvia

Az alapprobléma

www.erzsebetvaros.hu/uvegzszeb


ÜVEGZSEB - AZ ÖNKORMÁNYZATI VAGYONHOZ KÖTŐDŐ EGYES SZERZŐDÉSEK

Keresés:

Megbízó:

Megbízott:

Szerződéskötés_dátuma:

KERESÉS 

TALÁLATOK : 914 DB

➤ Megbízó: Budapest Főváros VII. kerület Erzsébetváros Önkormányzata
Megbízott: Erzsébetvárosi Média Nonprofit Kft.
Szerződés száma:
Szerződéskötés dátuma: 0000.00.00.
Szerződéskötés értéke: 27.300.000

➤ Megbízó: Budapest Főváros VII. kerület Erzsébetváros Önkormányzata
Megbízott: Budapesti Rendőr-főkapitányság
Szerződés száma:
Szerződéskötés dátuma: 0000.00.00.
Szerződéskötés értéke: 4.850.000

A jó hír



www.erzsebetvaros.hu/uvegzs...oldal/1/866

ÜVEGZSEB - AZ ÖNKORMÁNYZATI VAGYONHOZ KÖTÖDŐ EGYES SZERZŐDÉSEK

Keresés:

Megbízó:

Összes

Megbízott:

Összes

Szerződéskötés_dátuma:

Összes

KERESÉS



KIVÁLASZTOTT SZERZŐDÉS

Szerződés száma:

Szerződéskötés dátuma: 0000.00.00.

SZERZŐDŐ FELEK

Megbízó

Megbízott

Budapest Főváros VII. kerület Erzsébetváros Önkormányzata

Erzsébetvárosi Média Nonprofit Kft.

Közzététel dátuma: 2015.05.04.

Szerződés időtartam:

Szerződés típusa: támogatási

Szerződés értéke: 27.300.000

Szerződés tárgya: végelszámolási eljárással összefüggő kiadások fedezésére



VISSZA

A jó hír közelebbről



view-source:www.erzsebetvaros.hu/uvegzseb/oldal/1/866

```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2 <html xmlns="http://www.w3.org/1999/xhtml">
3
4 <head>
5   <title>Erzsébetváros</title>
6   <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
7   <meta name="description" content="mivel foglalkozik az oldal" />
8   <meta name="keywords" content="keresőszavak" />
9   <meta name="robots" content="index,follow" />
10  <meta name="author" content="Web Maker - Design Kft." />
11
12  <link type="text/css" rel="stylesheet" href="/css/reset.css" />
13  <link type="text/css" rel="stylesheet" href="/css/style.css" />
14  <link type="text/css" rel="stylesheet" href="/css/menu.css" />
15  <link type="text/css" rel="stylesheet" href="/css/orbit.css" />
16  <link type="text/css" rel="stylesheet" href="/css/slide.css" />
17  <!--[if IE]>
18  <style type="text/css">
19    .timer { display: none !important; }
20    div.caption { background:transparent; filter:progid:DXImageTransform.Microsoft.gradient(startColorstr=#99000000,endColorstr=#99000000);zoom: 1; }
21  </style>
22  <![endif]-->
23  <link rel="shortcut icon" href="/images/favicon.ico" />
24
25  <script type="text/javascript" src="/js/jquery.1.3.2.min.js" ></script>
26  <script type="text/javascript" src="/js/functions.js" ></script>
27  <script type="text/javascript" src="/js/jquery.orbit.min.js" ></script>
28  <script type="text/javascript" src="/js/slide.js" ></script>
29
30  <script type="text/javascript" src="/swfobject.js"></script>
31
32  <script type="text/javascript">AC_FL_RunContent = 0;</script>
33  <script src="/AC_RunActiveContent.js" type="text/javascript"></script>
34
35
36
37  <script type="text/javascript" src="/js/jquery.ui.core.js"></script>
38  <script type="text/javascript" src="/js/jquery.ui.datepicker.js"></script>
39
40  <link rel="stylesheet" type="text/css" href="/shadowbox/shadowbox.css" />
41  <script type="text/javascript" src="/shadowbox/shadowbox.js"></script>
42  <script type="text/javascript">
43
```

Szerződéskötés dátuma:0000.00.00.

SZERZŐDŐ FELEK

	Megbízott
ormányzata	Erzsébetvárosi Média Nor

Back Alt+Left Arrow

Forward Alt+Right Arrow

Reload Ctrl+R

Save as... Ctrl+S

Print... Ctrl+P

Cast...

Translate to English

View page source Ctrl+U

Inspect Ctrl+Shift+I

A lényeg



```
        </table>
      </form>
    </div>
  </div>
  <div class="centerTitle"><h1>KIVÁLASZTOTT SZERZŐDÉS</h1></div>
  <div class="centerContainer">
    <div class="contractsList">
      <table width='738' class='contractsBigRow'>
        <tr>
          <td class='contractsBigRow'>Szerződés száma:<strong></strong></td>
          <td class='contractsBigRow'>Szerződéskötés dátuma:<strong>0000.00.00.</strong></td>
        </tr>
        <tr>
          <td colspan='2'><h1>SZERZŐDŐ FELEK</h1></td>
        </tr>
        <tr>
          <td class='contractsHalf'><h2>Megbízó</h2></td>
          <td class='contractsHalf'><h2>Megbízott</h2></td>
        </tr>
        <tr>
          <td class='contractsHalf'><h3>Budapest Főváros VII. kerület Erzsébetváros Önkormányzata</h3></td>
          <td class='contractsHalf'><h3>Erzsébetvárosi Média Nonprofit Kft.</h3></td>
        </tr>
        <tr>
          <td colspan='2'><span>Közzététel dátuma:&nbsp;<strong>2015.05.04.</strong></span></td>
        </tr>
        <tr>
          <td colspan='2'><span>Szerződés időtartam:&nbsp;<strong></strong></span></td>
        </tr>
        <tr>
          <td colspan='2'><span>Szerződés típusa:&nbsp;<strong>támogatási</strong></span></td>
        </tr>
        <tr>
          <td colspan='2'><span>Szerződés értéke:&nbsp;<strong>27.300.000</strong></span></td>
        </tr>
        <tr>
          <td colspan='2'><span>Szerződés tárgya:&nbsp;<strong>végelszámolási eljárással összefüggő kiadások fedezésére</strong></span></td>
        </tr>
      </table>
    </div>
  </div>
```

A kód



```
1 setwd("D:/RLadies/lightningtalk_TSZ_20171212")
2 #install.packages("XML")
3 library(XML)
4 url<-"http://www.erzsebetvaros.hu//uvegzseb/oldal/1/5"
5 html<-htmlTreeParse(url, useInternalNodes=T)
6 x<-xpathSApply(html,"//strong",xmlvalue)
7 y<-xpathSApply(html,"//h3",xmlvalue)
8 g<-list(url)
9 z<-append(g, y)
10 z2<-append(z, x)
11 z3<-as.data.frame(z2)
12 colnames(z3) <- c("honlap", "megbízó", "megbízott", "szerződés száma", "szerződéskötés dátuma", "közzététel dátuma",
13                  "szerződés időtartama", "szerződés típusa", "szerződés értéke", "szerződés tárgya")
14
15
16 for (i in 6:1200){
17   url<-paste("http://www.erzsebetvaros.hu//uvegzseb/oldal/1/", i, sep="")
18   html<-htmlTreeParse(url, useInternalNodes=T)
19   x<-xpathSApply(html,"//strong",xmlvalue)
20   y<-xpathSApply(html,"//h3",xmlvalue)
21   if(length(x) != 0) {
22     if (length(y) !=0)
23       g<-list(url)
24     z<-append(g, y)
25     z2<-append(z, x)
26     z3x<-as.data.frame(z2)
27     colnames(z3x) <- c("honlap", "megbízó", "megbízott", "szerződés száma", "szerződéskötés dátuma", "közzététel dátuma",
28                       "szerződés időtartama", "szerződés típusa", "szerződés értéke", "szerződés tárgya")
29     z3<-rbind(z3, z3x)
30   }
31   else
32     url
33 }
34
35 write.csv(z3, "7ker_uvegzseb_alapadatok.csv")
36 |
```

A honlap, ahonnan az adatot kinyerni szeretnénk

A kapcsolat felállítása a honlap és az R közt

A számunkra szükséges információk „jelölői”, az adatok kinyerése

Az adatok összefűzése és az oszlopok elnevezése, tárolás data.frame-ként (ez a kiírásnál lesz fontos)


Ciklus → végiglépeget a számozás alapján az összes honlapon, és összefűzi az összes információt

Ha egy honlap nem létezik, az ebből kiderül, és ezzel a kis feltétellel nem hibára fut, hanem továbblép

A végső, teljes adatbázis kiírata csv formátumban

A kód közelebbről



```
Console ~/ 
> library(XML)
> url<-"http://www.erzsebetvaros.hu//uvegzseb/oldal/1/5"
> html<-htmlTreeParse(url, useInternalNodes=T)
> x<-xpathSApply(html,"//strong",xmlValue)
> y<-xpathSApply(html,"//h3",xmlValue)
> g<-list(url)
> x
[1] ""
[2] "1988.11.10."
[3] "2009.06.11."
[4] "Határozatlan"
[5] "szolgáltatói"
[6] "15.420 e/év"
[7] "Erzsébet krt. 6. szám alatti napi takarítás, évi kétszeri nagytakarítás"
> y
[1] "Budapest Főváros VII. kerület Erzsébetváros Önkormányzata"
[2] "Takarító Szövetkezet"
> g
[[1]]
[1] "http://www.erzsebetvaros.hu//uvegzseb/oldal/1/5"
> |
```


Az eredmény



FILE										HOME										INSERT										PAGE LAYOUT										FORMULAS										DATA										REVIEW										VIEW									
<div><div><div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div>																																																																															

Köszönöm a figyelmet! ☺



Kérdések?