Web2RSS with Perl

$펄마니아->{김기석}

http://code.google.com/p/web2rss/

```
svn checkout http://web2rss.googlecode.com/svn/trunk/ web2rss
```

Web2RSS ?


Korean Perl Workshop 2008
Rising Perl
펄마니아 소개
회원 목록
문서 모음
프로젝트 소개
NEW Trac & Wiki
자유 게시판
P&S게시판
언어 펄
윈도우 펄
CGI
데이터베이스
mod_perl&Mason
업로드 자료실
관련 링크
아이콘
RSS PERLMANIA
SUB BLOGLINES
추가 HanRSS
0 000000 464946
Perlmania lang BBS
No
3296 특정 작업 수행 이후
3295 Re: 특정 작업 수
3294 Re: Re: 특정 작
3293 변수만 저장된 File을
3292 Re: 변수만 저장된
3291 Re: 변수만 저장된
3290 Re: Re: 변수만
3289 Re: Re: Re: 변
3288 Re: Re: Re: 변
3287 Re: Re: Re: 변
3286 Re: Re: Re: 변
3285 정규표현식 패턴 문
3284 Re: 정규표현식 패
3283 Re: Re: 정규표현
3282 WRFILE로 파일핸들
3281 예 뭔가를 빠뜨렸
3280 Re: 예 뭔가를 빠
3279 예외처리(try/catch)
<?xml version="1.0" encoding="euc-kr" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://purl.org/rss/1.0/"
xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:taxo="http://purl.org/rss/1.0/modules/taxonomy/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:syn="http://purl.org/rss/1.0/modules/syndication/" xmlns:admin="http://webns.net/mvcb/">
<channel rdf:about="http://www.perlmania.or.kr">
<title>PERLMANIA IS NOW SUFFERING FROM THE ENCODING CHANGE OF XML::RSS!!</title>
<link>http://www.perlmania.or.kr</link>
<description>Perlmania 최근 게시물</description>
<items>
<rdf:Seq>
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=windows&article=326" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6588" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6587" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=lang&article=3510" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6586" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6585" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6584" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=DB&article=325" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6583" />
<rdf:li rdf:resource="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=free&article=6582" />
</rdf:Seq>
</items>
</channel>
<item rdf:about="http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=windows&article=326">
<title>윈도우에서 mod_perl설정 질문입니다..</title>
<link>http://www.perlmania.or.kr/bbs/bbs.html?mode=read&table=windows&article=326</link>
<description>
<![CDATA[
CSV파일을 파싱하여 DB에 저장하는 펄스크립트 CGI를 짰습니다.
웹에서 파일을 업로드하고 위의 역할을 하는 CGI를 구동시키면
처음 좀 실행되다가 멈추고 한 10분에 나머지 작업을 합니다.
작업관리자의 CPU 사용율을 보나 MySQL 모니터로 쿼리 날리는것을 봐도
그냥
]]>
</description>
<dc:creator>
<![CDATA[외계인]]>
</dc:creator>
<dc:date>2008-08-22T07:43:45+09:00</dc:date>
</item>

왜 RSS 로?

RSS Reader
Flash(Flex)
Silverlight
gadget
rss2email
rss2js
rss2html
rss2json
....

Web | RSS > rss.xml

1. HTML 소스 가져오기
2. HTML 소스에서 원하는 부분을 저장
3. RSS 포멧으로 출력

use LWP::Simple;

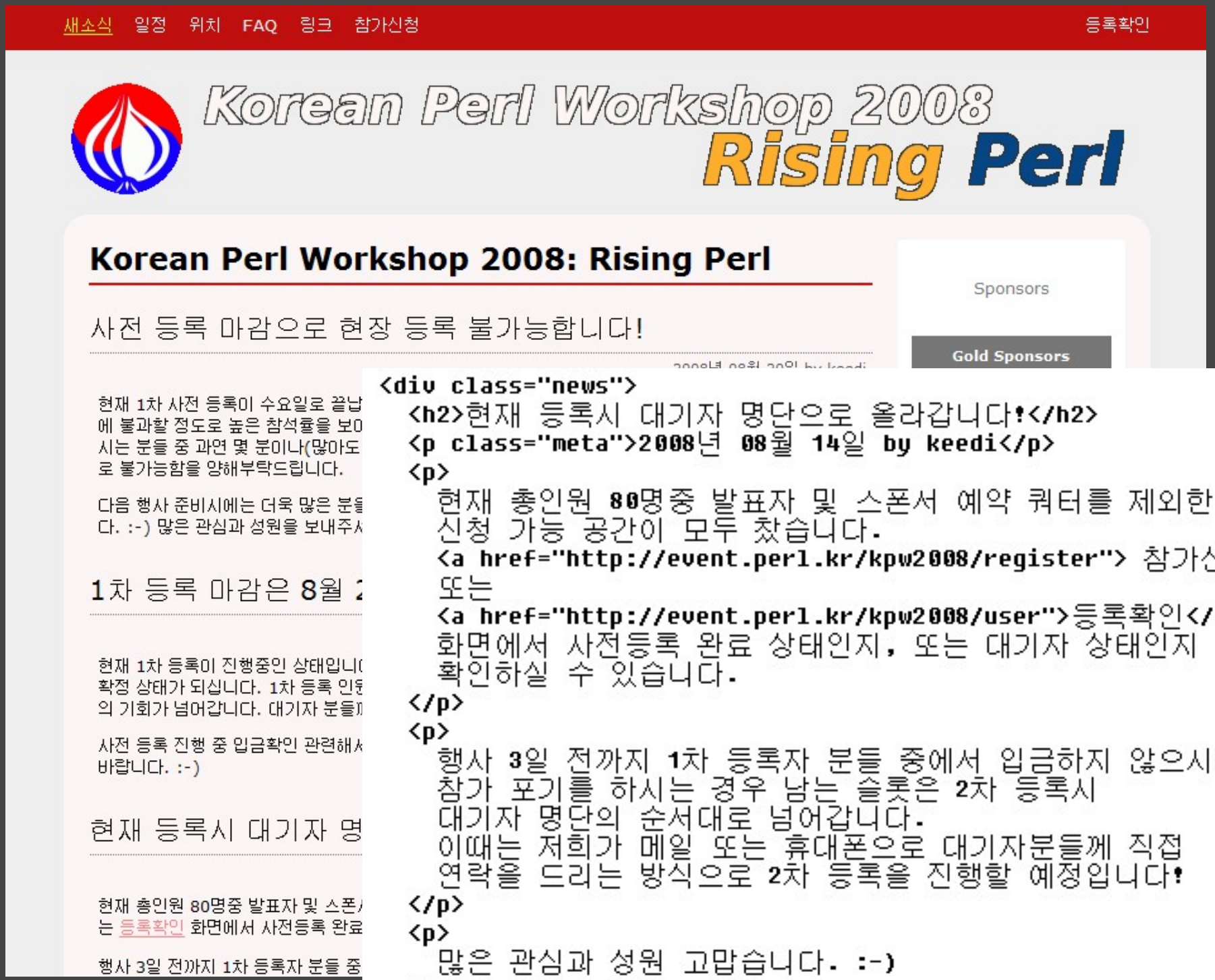
```
# HTML 소스 가져오기
my $html = get("http://event.perl.kr/kpw2008");

# 정규표현식으로 원하는 부분 저장 하기
my ( $rss_title ) = ( $html =~ m!<title>(.+?)</title>!i );
my ( $title, $desc ) = ( $html =~
      m!<div class="news"><h2>(.+?)</h2>(.+?)</div>!i );

# RSS 포멧으로 출력
$rss->save('rss.xml');
```

use LWP::Simple;

```
# HTML 소스 가져오기
my $html = get("http://event.perl.kr/kpw2008");

# 정규표현식으로 원하는 부분 저장 하기
my ( $rss_title ) = ( $html =~ m!<title>(.+?)</title>!i );
my ( $title, $desc ) = ( $html =~
      m!<div class="news"><h2>(.+?)</h2>(.+?)</div>!i );

# RSS 포멧으로 출력
$rss->save('rss.xml');
```

use XML::RSS;

```
my $rss = new XML::RSS();

$rss->channel(
    title => $title,
    link => $url,
);

$rss->add_item(
    title => $item->{title},
    link => $item->{link},
);

print $cgi->header("text/xml"), $rss->as_string;
```




```
<div class="news">
  <h2>현재 등록시 대기자 명단으로 올라갑니다!</h2>
  <p class="meta">2008년 08월 14일 by keedi</p>
  <p>
    현재 총인원 80명중 발표자 및 스폰서 예약 쿼터를 제외한
    신청 가능 공간이 모두 찼습니다.
    <a href="http://event.perl.kr/kpw2008/register"> 참가신청 </a>
    또는
    <a href="http://event.perl.kr/kpw2008/user">등록확인</a>
    화면에서 사전등록 완료 상태인지, 또는 대기자 상태인지
    확인하실 수 있습니다.
  </p>
  <p>
    행사 3일 전까지 1차 등록자 분들 중에서 입금하지 않으시거나
    참가 포기를 하시는 경우 남는 슬롯은 2차 등록시
    대기자 명단의 순서대로 넘어갑니다.
    이때는 저희가 메일 또는 휴대폰으로 대기자분들께 직접
    연락을 드리는 방식으로 2차 등록을 진행할 예정입니다!
  </p>
  <p>
    많은 관심과 성원 고맙습니다. :-)
  </p>
</div>
```

Demo

step1.pl

```
my ( $title ) = ( $html =~ m!<title>(.+?)</title>!i );

my ( @item ) = ( $html =~ m!<div class="news">\s*<h2>(.+?)</h2>(.+?)\s*</div>!g );
```

```
my ( @matchs ) = ( $html =~ m!<td\s*valign=top><a\s*href="(http://.+?)"\s*id=.+?\s*target=nw>
(.+?)</a>\s*<br>\s*<font size=-1>\s*<font color=.+?>(.+?) -</font>\s*<nobr>.+?
</nobr>\s*</font>\s*<br>\s*<font size=-1>(.+?)</font>!gi );
```

```
my (@title, @link, @desc );

( @title ) = ( $html =~ m!<td>\s*<a class=bu\s+href=".+?" target="_blank">(.+?)</a>\s*<font class=date>!gi );

( @link  ) = ( $html =~ m!<td>\s*<a class=bu\s+href="(.+?)" target="_blank">.+?</a>\s*<font class=date>!gi );

( @desc  ) = ( $html =~ m!<td class=bk>(.+?)</td>\s*</tr>!gi );
```

LWP + Regex

좋은점

속도가 빠름

펄과 함께 시간을 많이 보낼 수 있음 (정규표현력 상승)

나쁜점

웹사이트 변경시 작동 안함 (봄,가을 개편)

6개월 뒤 본인도 해독 불가 (어쩌면)

http://search.cpan.org

HTML::TagParser

HTML Document parser with
DOM-like Methods

use HTML::TagParser;

```
my $dom = HTML::TagParser->new( $url );

my ( $title ) =
    $dom->getElementsByTagName("title" )->innerText;

my @a_tag_list = $dom->getElementsByTagName( "a" );
my $link = $a_tag_list[0]->attributes->{href};


# DOMLike Methods
$elm = $dom->getElementById( $id );
@list = $dom->getElementByName( $name );
@list = $dom->getElementByTagName( $tagname );
@list = $dom->getElementByAttribute( $attrname, $value );
```

WWW::Mechanize

Web browsing in Perl object

use WWW::Mechanize;

```
my $mech = WWW::Mechanize->new();
$mech->get( $url );

my $res = $mech->find_all_links(
    text_regex => qr/\[기사₩]/i,
    url_regex => qr/menu=viewbody/i,
);

$title = $mech->title(); # <title>

# <a> 리스트
$rss->add_item( title => $link->text(),
   link => $link->URI()->abs ) foreach my $link( @{$res} );
```


```
<table width="460" border="0" cellspacing="0" cellpadding="0">

      <TR>
       <TD ALIGN='left' class='te2'><img src='/img/_pt_wcw.gif'> 
            <a href='?menu=viewbody&number=13279&keyfield=&key=&page=1'>[오피셜] 아스날, 실베스트르 영입 완료</a>
       </TD>
  <td align='right' class='te2'>2008-08-21</td>
      </TR>
      <TR>
       <TD HEIGHT='2' colspan='2'></TD>
      </TR>
```

Demo

step2.pl
mech.pl

JEEN님 블로그에서,

Web::Scraper?

http://jeen.tistory.com/entry/Perl-%B0%A3%B4%DC%C7%D1-%C0%A5%C6%E4%C0%(

이빨까기인형

지역로그 | 태그 | 미디어로그 | 방명록

[Perl] 간단한 웹페이지 스크랩핑 – Web::Scraper

IT/Perl 2008/06/24 17:06

일반적으로 웹 페이지를 스크래핑할 때 정규표현식에 대한 이해는 필수입니다.
페이지를 싸그리 통채로 스크랩하는 거야 물론 간단하지만, 페이지의 일부분을 도려내고 싶을 때는 정규표현이 없으면 헤맬 수 밖에 없습니다.

그리고 정규표현은 가독성이 단점입니다. 어중간한 정규표현이야 괜찮겠지만... 어떤 정규표현을 보고, 이게 뭐하는 것인지 알아내기란 참으로 힘든 일일 겁니다.

```
use LWP::Simple;

my $content = get("http://www.daum.net");
my ($title) = $content =~ /<title>(.+)<₩/title>/i;

print $title;
```

< 어중간하게 쉬운 정규표현의 예 : 웹페이지의 제목 을 얻어내는 스크립트 >

```
use Web::Scraper;
use URI;

my $html = scraper {
  process 'title', title => 'TEXT';
}->scrape(URI->new("http://www.daum.net"));

print $html->{title};
```

< 정규표현을 사용하지 않은 예 >

이런 간단한 예제에서는 Web::Scraper 가 지는 것은 당연하겠죠. :-)
차이가 있다면 문자코드입니다. 정규표현을 사용한 예제에서는 해당페이지에서 사용하고 있는 문자코드(다른 경우에는 euc-kr)로 결과가 나오게 되지만, Web::Scraper 를 사용했을 때는 모든 결과는 반드시 UTF-8 으로 됩니다.

http://www.slideshare.net/miyagawa/web-scraper-shibuyapm-tech-talk-8?src=emb

slideshare

Home | My Slidespace | Upload | Community | Tags | Widgets

Latest | Most Viewed | Most Embedded | Featured | Most Favorited | Most Downloaded | Slidecasts

It's free! Upload

Easy Web Data Scraping
Easily Scrap Web Data in Minutes. Download Now & Get a Free Trial.
www.AutomationAnywhere.com/scrape

Yangshuo Li River Retreat
New hotel on the Li River with great river views
www.li-river-retreat.com

FileMaker XML
FileMaker XML/XSLT Experts Products, Consultation, Training
www.chapsoft.com

Practical Web Scraping with Web::Scraper

Tatsuhiko Miyagawa
miyagawa@gmail.com

Six Apart, Ltd. / Shibuya Perl Mongers
Shibuya.pm Tech Talks #8

Web Scraper Shibuya.pm tech talk #8

From miyagawa, 10 months ago

6250 views | 1 comment | 4 favorites | 151 downloads | 13 embeds (Stats)

Share
Favorite
+ Group / Event
Download file

Tags
shibuya.pm perl webscraper [perl][miyagawa]

Groups / Events
TechPresentations
Shibuya Perl Mongers
Perl

Embed <div style="width:425px;text-alig
options

More by user | Related slideshows

Perl 5.10 for People Who Aren't Totall...
45935 views

Perl で入門テキストマイニン...
14384 views

Perl Sucks - and what to do about

All Comments (1) | Comments on Slide 1

Comments 1-1 of 1

guest64adc4 said 10 months ago (slide 1)
Good work.

use Web::Scraper;

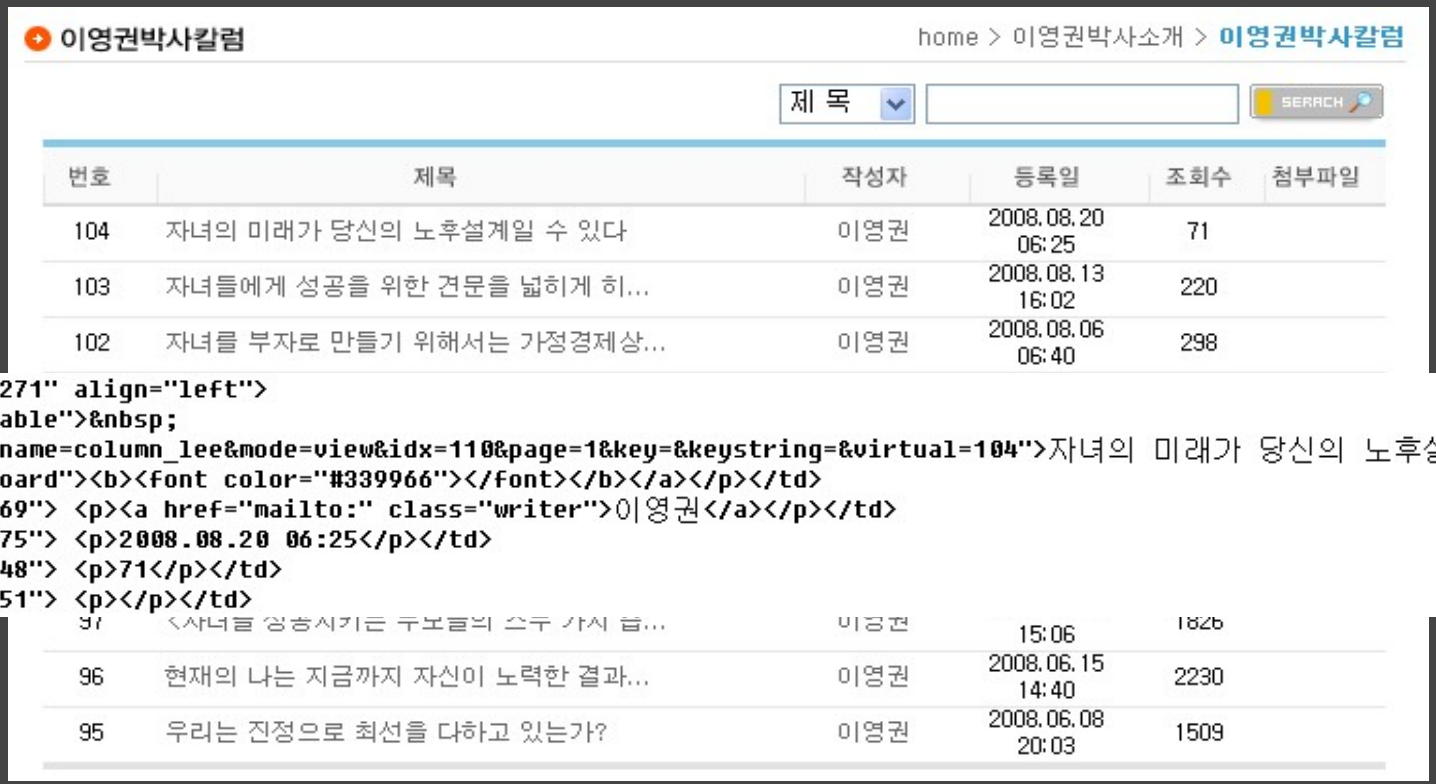
```
# <td class='te2'><a href='@href'>TEXT</a></td>
my $scrp = scraper {
    process "td.te2>a",
    "items[]" => {
        'title' => 'TEXT',
        'link' => '@href',
    };
};

my $res = $scrp->scrape( URI->new($url) );

foreach my $item ( @{ $res->{items} } ) {
    title => $item->{title},
    link => $item->{link},
}
```

Demo

step3.pl

" theres more than one page
to web2rss it. "

RSS로 만들고 싶은 웹페이지가 또 있다!




```
<td width="271" align="left">
<p class="table"> 
<a href="?tname=column_lee&mode=view&idx=110&page=1&key=&keystring=&virtual=104">자녀의 미래가 당신의 노후설계일 수 있다</a>
<a class="board"><b><Font color="#339966"></font></b></a></p></td>
<td width="69"> <p><a href="mailto:" class="writer">이영권</a></p></td>
<td width="75"> <p>2008.08.20 06:25</p></td>
<td width="48"> <p>71</p></td>
<td width="51"> <p></p></td>
```

%rss_id

```
my %rss_id = (

    soccerline =>
    {
        url => "http://www.soccerline.co.kr/news_list/index.php?menu=main",
        dsl => "td.te2>a",
        reg => "menu=viewbody",
    },

    bestmento =>
    {
        url => "http://www.bestmentorclub.org/information/colum.html",
        dsl => "p.table>a",
        reg => "colum.html",
    },
);

my $url = $rss_id{soccerline}->{url};
my $res = $scrp->scrape( URI->new($url) );
```

YAML

YAML Ain't Markup Language

use YAML; #YAML Ain't Markup Language

```
#my $rss_id = YAML::LoadFile('Conf/rss_list.yaml');
my $rss_id = YAML::Load(<<'....');
---
soccerline:
    url: http://www.soccerline.co.kr/news_list/
    dsl: td.te2>a
    reg: menu=viewbody
....

my $url = $rss_id->{soccerline}->{url};
my @key = keys %{$rss_id->{soccerline}};
```

웹서버 부하?

Cache::File
CGI::Cache

tie %hash, 'Cache::File'

%hash처럼 쓰는 Cache::File 모듈

tie %hash, 'Cache::File', \%opt;

```
tie %cached_rss, 'Cache::File', {
    cache_root => '/tmp',
    default_expires => '1 hour',
};

$cached_rss{$key}  = $data;  # 데이터 저장
$data =  $cached_rss{$key};  # 데이터 가져오기
if  ( defined $cached_rss{$key} ){
    # 캐싱되어있음
    print $cgi->header(), $cached_rss{$key}; exit;
}
else {
    # $cached_rss{$key} 에 결과 저장
}
```

Demo

step4.pl

Web2RSS with CPAN

and...

RSS 2 JSON

use JSON;

RSS 2 Javascript

use XML::RSS::JavaScript;

RSS 2 DeepZoom ?

use DeepZoom; #작업중

Perl  CPAN

Perl  RSS

감사합니다.