

# How can I extract the data out of a typical html day/time schedule?

Asked 16 years, 3 months ago   Modified 7 years ago   Viewed 299 times



I'm trying to write a parser to get the data out of a typical html table day/time schedule (like [this](#)).

0



I'd like to give this parser a page and a table class/id, and have it return a list of events, along with days & times they occur. It should take into account rowspans and colspans, so for the linked example, it would return



```
{:event => "Music With Paul Ray", :times => [T 12:00am - 3:00am, F 12:00am - 3:00am]}, etc.
```

I've sort of figured out a half-executed messy approach using ruby, and am wondering how you might tackle such a problem?

html

regex

Share

Improve this question

Follow

edited Dec 16, 2017 at 23:30



eLRuLL

18.8k ● 9 ● 77 ● 104

asked Sep 23, 2008 at 3:14



Jeff

## 4 Answers

Sorted by: Highest score (default)



The best thing to do here is to use a HTML parser. With a HTML parser you can look at the table rows programmatically, without having to resort to fragile regular expressions and doing the parsing yourself.

2



Then you can run some logic along the lines of (this is not runnable code, just a sketch that you should be able to see the idea from):



```
for row in table:
  i = 0
  for cell in row: # skipping row 1
    event = name
    starttime = row[0]
    endtime = table[ i + cell.rowspan + 1 ][0]
```

```
print event, starttime, endtime
i += 1
```

Share Improve this answer Follow

answered Sep 23, 2008 at 6:54



Jerub

42.6k ● 15 ● 75 ● 90



0



This is what the program will need to do:

1. Read the tags in (detect attributes and open/close tags)
2. Build an internal representation of the table (how will you handle malformed tables?)
3. Calculate the day, start time, and end time of each event
4. Merge repeated events into an event series

That's a lot of components! You'll probably need to ask a more specific question.

Share Improve this answer Follow

answered Sep 23, 2008 at 3:57



I answer wrong - have fun

2,811 ● 2 ● 26 ● 36



0



Use <http://www.crummy.com/software/BeautifulSoup/> and that task should be a breeze.

Share Improve this answer Follow

answered Sep 23, 2008 at 7:06



Hannes Ovrén

21.8k ● 9 ● 68 ● 76



0



As said, using regexes on HTML is generally a bad idea, you should use a good parser.

For validating XHTML pages, you can use a simple XML parser which is available in most languages. Alas, in your case, the given page doesn't validate (W3C's markup validation service reports 230 Errors, 7 warning(s)!

For generic, possibly malformed HTML, there are libraries to handle that (kigurai recommends BeautifulSoup for Python, I know also TagSoup for Java, there are others).

Share Improve this answer Follow

answered Sep 26, 2008 at 16:15



PhiLho

41.1k ● 6 ● 99 ● 136

---