

Wanderings with hpricot: A Screen Scraping Appetizer

Brian DeLacey

Boston Ruby Group, March 13, 2007
Revised: April 5, 2007
Kendall Square

Hpricot



- “A Fast, Enjoyable HTML Parser for Ruby”
- GoTo
 - <http://code.whytheluckystiff.net/hpricot/>
 - \$ gem install hpricot
 - The usual...
- Amazing software
- Darn good documentation too...
 - <http://code.whytheluckystiff.net/doc/hpricot/>

CSS Selectors

- See Challenges, Examples, and more...
 - <http://code.whytheluckystiff.net/hpricot/wiki/HpricotChallenge>
- <http://code.whytheluckystiff.net/hpricot/wiki/SupportedCssSelectors>



Hpricot Performance

“Here’s a benchmark parsing the Boing Boing home page fifty times. It’s a good page to test because it’s big and there’s some bogus end tags and old-style tables and break tags.” [_why](#)

	user	system	total	real
hpricot:	10.515625	0.000000	10.515625	(10.610571)
scrap:	32.546875	0.093750	32.640625	(32.923535)
htree:	56.609375	0.023438	56.632812	(57.096530)
rubyfulsoup:	29.289062	0.046875	29.335938	(29.586510)
mechanize: (*)	148.132812	1.101562	149.234375	(150.621922)
htmltok: (*)	19.632812	0.007812	19.640625	(19.795446)

(*) These libs are a bit more primitive, focusing only on reading documents, no calls are given for modifying documents.

mofu



“mofo is a microformat parser for Ruby based on Hpricot. It’s got a nice little DSL for defining microformats and currently supports hCard, hCalendar, hReview, hEntry, xoxo, rel-tag, and rel-bookmark. There may be a few kinks, but hey, it’s new.”