Data Scraping

PS236B

March 3, 2010

Three Steps

1. Inspect.
2. Parse.
3. Extract.

Software

- Perl:
  - `WWW-mechanize` module
- Python
  - `BeautifulSoup` library
- R?

Scraping with R

- `readLines`
- XML and `rjson` packages
- `rcurl` package
- most important: `grep`, `sub`, and `gsub` commands

Easy Case

- Well formed HTML Table already in a rectangular format.
- Use the `readHTMLTable` command from the `XML` package to parse.
- Light use of regular expressions to extract.

Origin (UTC)	Present-day Country and link to Wikipedia article	Latitude	Longitude	Depth (km)	Magnitude	Secondary Effects	PDE Shaking Deaths	PDE Total Deaths	Utsu Total Deaths	EM-DAT Total Deaths	Other Source Deaths
1900-05-11 17:23	Japan	38.700	141.100	5	MJ 7.0						
1900-07-12 06:25	Turkey	40.300	43.100		UK 5.9				140		
1900-10-29 09:11	Venezuela	11.000	-66.000	0	MW 7.7						
1901-02-15 00:00	China	26.000	100.100	0	MS 6.5						
1901-03-31 07:11	Bulgaria	43.400	28.700		UK 6.4				4		
1901-08-09 09:23	Japan	40.500	142.500	35	MW 7.2	T					
1901-11-15 20:15	New Zealand	-43.000	173.000	0	MS 6.8				1		
1902-01-30 14:01	Japan	40.500	141.300	35	MS 6.9				1		

Wikipedia's Earthquake List

```
<table class="sortable wikitable">
<tr bgcolor="#CCCCCC">
<th width="22%">Origin (<a href="/wiki/UTC" title="UTC" class="mw-redirect">UTC</a>)</th>
<th width="35%">Present-day Country and link to Wikipedia article</th>
<th width="10%"><a href="/wiki/Latitude" title="Latitude">Latitude</a></th>
<th width="10%"><a href="/wiki/Longitude" title="Longitude">Longitude</a></th>
<th width="8%">Depth (<a href="/wiki/Km" title="Km" class="mw-redirect">km</a>)</th>
<th width="6%">Magnitude</th>
<th width="9%">Secondary Effects</th>
<th width="10%">PDE Shaking Deaths</th>
<th width="10%">PDE Total Deaths</th>
<th width="10%">Utsu Total Deaths</th>
<th width="10%">EM-DAT Total Deaths</th>
<th width="10%">Other Source Deaths</th>
</tr>
<tr>
<td>1900-05-11 17:23</td>
<td><a href="/wiki/Japan" title="Japan">Japan</a></td>
<td>38.700</td>
<td>141.100</td>
<td>5</td>
<td><a href="/wiki/Japan_Meteorological_Agency_seismic_intensity_scale" title="Japan Meteorological Agency seismic intensity
scale">MJ</a> 7.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
```

HTML CODE

gsub is your friend

- example: `gsub(".*([0-9]{1}.[0-9]{1}).*", earthquakes$magnitude, replacement="\\1")`
- . - matches any character
- * - matches zero or more copies
- () - copies everything between the parentheses.

gsub is your friend

- example: `gsub(".*([0-9]{1}.[0-9]{1}).*", earthquakes$magnitude, replacement="\\1")`

  - `[0-9] {1}.[0-9]{1}-` match a digit between 0 and 9 exactly once, followed by a period, followed by one digit between 0 and 9
  - `replacement="\\1"` - replace with the first match

Other useful regular expressions

- `^` - start of a string, `$` - end of a string
- `[0-9]{1,2}` - match any 1 **or** 2 digit sequence of numbers
- To match "meta-characters", such as " `. \ | ( ) [ { ^ $ * + ?`", you need to precede them by a "\", i.e. an escape character

Harder Case

- Does't follow web standards, so convenient XML and HTML parsers don't work.
- Use `readLines()` and lots of regular expressions to extract data.

#	Year	Campus	Name	Title	Base Pay	Overtime Pay	Extra Pay	Gross Pay
1.	2008	BERKELEY	TEDFORD , JEFF	HEAD COACH-INTERCOLG ATHLETICS	$225,000.02	$0.00	$2,117,314.50	**$2,342,314.52**
2.	2008	BERKELEY	MONTGOMERY , MICHAEL J.	HEAD COACH-INTERCOLG ATHLETICS	$183,712.47	$0.00	$734,849.97	**$918,562.44**
3.	2008	BERKELEY	BOYLE , JOANNE	HEAD COACH-INTERCOLG ATHLETICS	$239,048.34	$0.00	$396,033.02	**$635,081.36**
4.	2008	BERKELEY	ISAACS , ANDREW M	ADJ PROF-ACAD YR-BUS/ECON/ENG	$58,077.78	$0.00	$543,500.00	**$601,577.78**
5.	2008	BERKELEY	HO , TECK HUA	PROFESSOR-ACAD YR-BUS/ECON/ENG	$269,000.04	$0.00	$269,076.69	**$538,076.73**
6.	2008	BERKELEY	BARBOUR , ANNE SAUNDERS	HEAD COACH-INTERCOLG ATHLETICS	$265,575.00	$0.00	$187,435.28	**$453,010.28**
7.	2008	BERKELEY	BIRGENEAU , ROBERT J.	CHANCELLOR	$436,800.00	$0.00	$8,916.00	**$445,716.00**
8.	2008	BERKELEY	SOMERVILLE , CHRISTOPHER R.	DIRECTOR	$218,199.96	$0.00	$167,731.92	**$385,931.88**
9.	2008	BERKELEY	AUERBACH , ALAN J	MISCELLANEOUS	$270,300.00	$0.00	$90,100.00	**$360,400.00**
10.	2008	BERKELEY	CIGNETTI , FRANK J.	ASST COACH-INTERCOLG ATHLETICS	$166,174.40	$0.00	$185,852.50	**$352,026.90**
11.	2008	BERKELEY	FARBER , DANIEL	PROFESSOR-LAW SCHOOL SCALE	$264,999.96	$0.00	$86,283.30	**$351,283.26**
12.	2008	BERKELEY	IYER , GANESH	ASSOCIATE DEAN	$199,900.08	$0.00	$146,217.33	**$346,117.41**
13.	2008	BERKELEY	PATTERSON , DAVID A	PROFESSOR-ACAD YR-BUS/ECON/ENG	$259,400.04	$0.00	$86,466.66	**$345,866.70**
14.	2008	BERKELEY	BRAUN , BEN	HEAD COACH-INTERCOLG ATHLETICS	$52,000.00	$0.00	$292,015.08	**$344,015.08**
15.	2008	BERKELEY	RUBINSTEIN , MARK E	PROFESSOR-ACAD YR-BUS/ECON/ENG	$281,100.00	$0.00	$62,466.66	**$343,566.66**
16.	2008	BERKELEY	JONES , VAUGHAN FREDERIC	PROFESSOR - ACADEMIC YEAR	$267,500.04	$0.00	$75,848.55	**$343,348.59**
17.	2008	BERKELEY	CHATMAN , JENNIFER A	PROFESSOR-ACAD YR-BUS/ECON/ENG	$222,800.04	$0.00	$111,519.12	**$334,319.16**

UC Salaries

grep is your friend

- Use the grep command to break the html into manageable pieces.
- `entry.start <- grep("orowlerow", html)`
- Produces an index which indicates where each entry begins.
- Loop over these entries and extract the data as you go along.

Multiple pages?

- Best case scenario: the url is structured logically: "http://ucpay.globl.org/index.php?campus=berkeley"
  - Loop over the various URL's using the `paste()` and `readLines()`
- Other scenarios: `post` commands, `https`
  - Diagnose with Live HTTP Headers package
  - Use `rcurl`, `httpRequest` packages

API's

- **A**pplication **P**rogramming **I**nterface
- Increasingly common: Google Maps for geocoding, NY Times article search, campaign finance data, roll call data, Twitter, etc
- Apply for an API key