Mining Wikipedia for Awesome Data

Neil Crosby

What's this about then?

- There's loads of groovy content on Wikipedia[citation needed].
- You are lazy.
- You want groovy content on your site.

Wikipedia has an API

- Who knew?
- http://en.wikipedia.org/w/api.php

API has lots of options

Param	Values	What does it do?
format	php, json, TODO	Output format.
redirects	0, 1	Redirect to good pages.
rvsection	0, 1, 2, 3, etc	Page section to get data for.
action	query, parse	API method.

Getting WikiText? Easy

- http://en.wikipedia.org/w/api.php?format=php&action=query&titles=one+flew+over+the+cuckoo’s+nest&rvprop=content&prop=revisions&redirects=1

Searching? Harder

- Wikipedia doesn’t have a good search engine.

Use Yahoo! BOSS

- http://boss.yahooapis.com/ysearch/web/v1/site:en.wikipedia.org+one+flew+over+the+cuckoo’s+nest?appid=yourBOSSiD
- First result: http://en.wikipedia.org/wiki/One_Flew_Over_the_Cuckoo's_Nest_(film)

Then get WikiText

- http://en.wikipedia.org/w/api.php?format=php&action=query&titles=One_Flew_Over_the_Cuckoo's_Nest_(film)&rvprop=content&prop=revisions&redirects=1

The WikiText

'''''One Flew Over the Cuckoo's Nest''''' is a [[1975 in film|1975]] [[comedy-drama]] film [[film director|directed]] by [[Miloš Forman]]. The film is an adaptation of the 1962 novel ''[[One Flew Over the Cuckoo's Nest (novel)|One Flew Over the Cuckoo's Nest]]'' by [[Ken Kesey]]. The movie was the first to [[List of Big Five Academy Award winners and nominees|win all five]]...

But I wanted HTML!

- WikiText is no good for dumping into a website.

Another API call

- http://en.wikipedia.org/w/api.php?action=parse&format=php&text=returned+wiki+text
- Text will be big - do as a POST.

Wiki HTML!

```
<p><i><b>One Flew Over the Cuckoo's
Nest</b></i> is a <a href="/wiki/
1975_in_film" title="1975 in film">1975</a>
<a href="/wiki/Comedy-drama"
title="Comedy-drama">comedy-drama</a>
film <a href="/wiki/Film_director" title="Film
director">directed</a> by <a href="/wiki/
Milo%C5%A1_Forman" title="Miloš
Forman">Miloš Forman</a>. The film is an...
```

Reducing the HTML

- DOMDocument->loadHTML()
- DOMXPath->query() to get wanted nodes.
- DOMDocument->saveHTML()
- str_replace() away HTML boilerplate.

The Cuckoo Problem

- “One Flew Over the Cuckoo’s Nest”
- A book?
- A film?
- Depends on context.

The Cuckoo Solution

- Give context:
  - “one flew over the cuckoo’s nest book”
  - “one flew over the cuckoo’s nest movie”
- Yahoo! BOSS gives relevant result. Yay.

There’s still a problem...

- Sometimes you can give too much context.
- “wii fit” gets expected result.
- “wii fit electronics” returns “WiiMote”.
- Oh dear.

When is too much?

- Who knows?
- Just because an article exists for the basic term doesn’t mean it’s the right article.
- I’ve not solved this yet.

It’s all too complicated

- So don’t do it all.
- Use a library.
- http://thecodetrain.co.uk/code/wikislurp

Runs as a web service

- http://yoursite.com/wikislurp/?params=blah

What are the params?

Param	Meaning
secret	Your self-chosen appid.
query	What you'd like wiki info about.
context	A little bit of context.
section	Article section to retrieve. Zero indexed.
xpath	Specify the elements to return.
output	Serialised php or json.

What does it return?

- An array.
- Keys for “url”, “title” and “article”.

Why a webservice?

- You can’t abandon a function call in PHP.
- You can abandon a CURL call.
- If wikislurp takes too long, move on.

Kitten Break

There’s some code coming up, soz.

http://www.flickr.com/photos/gsx-r750/1475603952/

How to call WikiSlurp

- http://yoursite.com/wikislurp/?secret=YOUR+SECRET&query=one+flew+over+the+cuckoo’s+nest&context=book&xpath=/html/body/p[position()<=3]&section=0&output=json

And from PHP?

```
$s = curl_init();
curl_setopt($s,CURLOPT_URL, $url);
curl_setopt($s,CURLOPT_HEADER,false);
curl_setopt($s,
    CURLOPT_RETURNTRANSFER,1);
// wait 1 second, then abort
curl_setopt($s,CURLOPT_TIMEOUT,1);
$result = curl_exec($s);
curl_close( $s );
```

XPath?

Query	Gives You
//p	All <p>
/html/body/p	All <p> directly under <body>
/html/body/p[2]	2nd <p> directly...
/html/body/p[position()<=3]	First three <p> directly...

Oh noes, more XPath

Query	Gives You
/html/body/p[@class='fish']	All <p> with single class “fish”
/html/body/p[contains(concat(" ",@class," "), " fish ")]	All <p> with any class including “fish”

Phew.

Have another kitten.

http://www.flickr.com/photos/evapro/305689596/

Future Features

- Do something intelligent with context.
- Convert to HTML without an extra API call.
- Return proper error codes if things go wrong.

Where is this used?

- TheTenWordReview.com
- IsNeilAnnoyedBy.com

Questions?

- I will blog about this talk at The Code Train.
- No, really - I will.
- Download the slurpy source code from http://thecodetrain.co.uk/code/wikislurp
- Slides? http://icanhaz.com/wikislurpslides
- I was and am http://NeilCrosby.com/vcard