

Web Scrapping

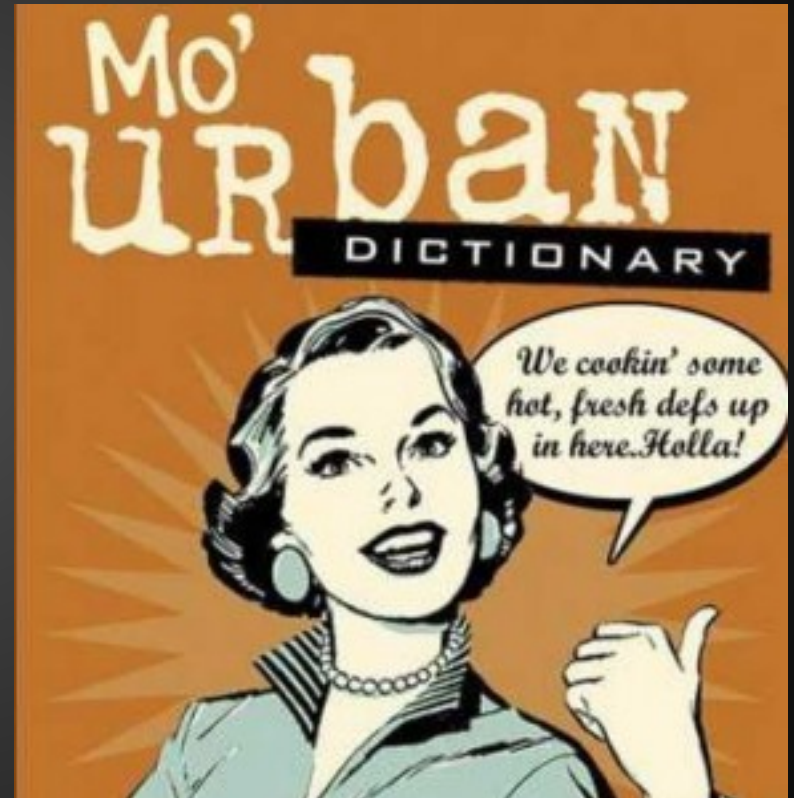
Scraping steps

Steps:

1. Figure out the URL(s) to scrape
2. Write code to download the HTML given the URL
3. Find unique HTML tags surrounding your target information
 - a. `Here's the title!`
4. Write code using string manipulation to extract relevant data
5. Display and/or store scraped information

Case Study: Urban Dictionary

- We will write a scraper for Urban Dictionary
- Given a term inputted by our user, print the first definition and example found on Urban Dictionary for that term



Step 1: Figure out URLs

- First thing we do is visit the website!
- Click a few definitions on urbandictionary.com
- Try to find a pattern in the URLs
 - Find the “minimum URL” that will give you a certain page

Figure Out the URL

The screenshot shows a web browser window with the URL `www.urbandictionary.com/define.php?term=antistalking&defid=7230994`. The page features the Urban Dictionary logo, a search bar with 'antistalking' entered, and a navigation menu. The main content area displays the definition of 'antistalking' as the 'word of the day' for October 21, 2013. It includes a description, an example sentence, and a list of related terms. A sidebar on the left lists trending and alphabetical words. At the bottom, there is a 'Random Word' button.

Urban Dictionary: antistalk x

www.urbandictionary.com/define.php?term=antistalking&defid=7230994

Subscribe Feedback Like 2.6m Follow 181K followers English

look up any word, like purp:
antistalking search

word of the day categories favorites **dictionary** game thesaurus names media store add blog

random A B C D E F G H I J K L M N O P Q R S T U V W X Y Z # new tv

trending
ratchet
swag
cenosillicaphobia
hipster
douchebag
molly
purp
hashtag
paratrooping
yolo

categories
gaming
sports
food
sex
tv
film
celebrities
military
music
weather
insults

alphabetical
antispamdexing
antispaming
Anti-Spaniel
anti spick
antispire
Antisplash
anti-splash paper
Anti-Spoiler
Anti-spooning
ANTI-SPORTIST
anti-spouse
anti spunkieside
anti square
Anti Squirrel Coalition
anti-stalker

antistalking

1. antistalking

word of the day: October 21, 2013
Methodically learning another person's routine in order to systematically avoid them.
Bradley is antistalking his boss today because he hates the bastard.
anti-stalking anti stalking stalking stalker antistalker
by Clearly Oct 13 add a video
746 up, 163 down

Random Word

Figure Out the URL

 www.urbandictionary.com/define.php?term=antistalking

Figure Out the URL

The screenshot shows a web browser window with the URL `www.urbandictionary.com/define.php?term=Barack%20Obama`. The browser's address bar and tabs are visible at the top. The Urban Dictionary website has a blue header with the 'urban DICTIONARY' logo. A search bar contains 'Barack Obama'. Below the header is a navigation bar with links like 'word of the day', 'categories', 'favorites', 'dictionary', 'game', 'thesaurus', 'names', 'media', 'store', 'add', and 'blog'. A secondary navigation bar shows an alphabet index from 'random' to 'tv'. On the left side, there are lists for 'trending' words, 'categories', and 'see also' related terms. The main content area displays two definitions for 'Barack Obama'. The first definition includes a description, a quote, tags, and user interaction statistics. The second definition is a satirical paragraph. A 'Random Word' button is located between the two definitions.

Urban Dictionary

look up any word, like [cenosillicaphobia](#):
Barack Obama search

word of the day categories favorites **dictionary** game thesaurus names media store add blog

random A B C D E F G H I J K L M N O P Q R S T U V W X Y Z # new tv

trending
ratchet
swag
cenosillicaphobia
hipster
douchebag
molly
purp
hashtag
paratrooping
yolo

categories
gaming
sports
food
sex
tv
film
celebrities
military
music
weather
insults

see also
obama
president
barack
democrat
politics
john mccain
black
hillary clinton
liberal
nigger
sarah palin
election
socialist

1. Barack Obama

The first half-white president.

I bet you only thought of Barack Obama as black, didn't you? Shame.

president 43 black white mixed race
by PlutoRoman Jun 13 add a video

5363 up, 1863 down

Random Word

2. Barack Obama

A puppet picked by the Global Elite to fastened the process of globalization. He is on the Council of Foreign Relations which advocates world government throught its think tank. It published Foreign Affairs magazine. He is not the "Radical Ismalist anti-christ" that ignorant right-wingers claim and is definately not the "most radical democrat" that will end the Iraq occupation. He is but another pawn in place to perpetuate the right and left mirage in US domestic politics. His foreign policy is just as aggressive and detrimental as any other puppets.

Ignorant people will vote for him because his empty rhetorical speeches promise change (while he lives his life by the status quo) that he will never institute. He is an uncle tom in every sense. Any black man voting for him should follow him closely before selling out.

Figure Out the URL



www.urbandictionary.com/define.php?term=Barack%20Obama

Step 2: Download the HTML

```
import urllib2
url = 'http://www.urbandictionary.com/define.php?term=antistalking'
html = urllib2.urlopen(url).read()
print html
```

Note that this will only work in the terminal!

How can we allow user input?

```
import urllib2
user_word = raw_input('Define a word: ')
url = 'http://www.urbandictionary.com/define.php?term=' + user_word
html = urllib2.urlopen(url).read()
print html
```

Step 3: Find unique HTML tags

- View page source (or look at the html you printed to the terminal)
- Find the first definition / example on the page
 - Use Ctrl-F to find specific words!
- Look for surrounding tags/patterns
 - If we're lucky they'll be unique

Find Unique HTML Tags

The screenshot shows a web browser window with the URL `www.urbandictionary.com/define.php?term=antistalking&defid=7230994`. The browser's address bar and tabs are visible at the top. The Urban Dictionary logo is on the left, and a search bar is on the right. Below the logo, there are navigation links for 'word of the day', 'categories', 'favorites', 'dictionary', 'game', 'thesaurus', 'names', 'media', 'store', 'add', and 'blog'. The 'dictionary' link is highlighted. Below these links, there is a row of letters from A to Z, with 'A' highlighted. The main content area shows the definition of 'antistalking' as the 'word of the day' for October 21, 2013. The definition is: 'Methodically learning another person's routine in order to systematically avoid them.' Below the definition, there is a quote: 'Bradley is antistalking his boss today because he hates the bastard.' and a list of related terms: 'anti-stalking', 'anti stalking', 'stalking', 'stalker', and 'antistalker'. The definition is by 'Clearly' on 'Oct 13'. There are 746 upvotes and 163 downvotes. A 'Random Word' button is at the bottom. On the left side, there is a sidebar with 'trending' words and 'categories'. The 'antistalking' category is highlighted at the bottom of the sidebar.

Urban Dictionary: antistalk x

www.urbandictionary.com/define.php?term=antistalking&defid=7230994

look up any word, like purp:

antistalking search

word of the day categories favorites **dictionary** game thesaurus names media store add blog

random **A** B C D E F G H I J K L M N O P Q R S T U V W X Y Z # new tv

trending
ratchet
swag
cenosillicaphobia
hipster
douchebag
molly
purp
hashtag
paratrooping
yolo

categories
gaming
sports
food
sex
tv
film
celebrities
military
music
weather
insults

alphabetical
antispamdexing
antispaming
Anti-Spaniel
anti spick
antispire
Antisplash
anti-splash paper
Anti-Spoiler
Anti-spooning
ANTI-SPORTIST
anti-spouse
anti spunkieside
anti square
Anti Squirrel Coalition
anti-stalker

antistalking

1. antistalking

word of the day: October 21, 2013

Methodically learning another person's routine in order to systematically avoid them.

Bradley is antistalking his boss today because he hates the bastard.

anti-stalking **anti stalking** **stalking** **stalker** **antistalker**

by **Clearly** Oct 13 [add a video](#)

746 up, 163 down

Random Word

Find Unique HTML Tags

The screenshot shows a web browser window with multiple tabs open. The active tab is 'Urban Dictionary: antistalk'. The address bar shows the URL 'www.urbandictionary.com/define.php?term=antistalking'. The page features the Urban Dictionary logo, a search bar with 'antistalking' entered, and a navigation menu. The main content area displays the definition of 'antistalking' as the 'word of the day' for October 21, 2013. The definition includes a description, a quote, and a list of related terms. A context menu is open over the page, showing options like 'Back', 'Forward', 'Reload', 'Save As...', 'Print...', 'Translate to English', 'View Page Source', 'View Page Info', 'Reload Frame', 'View Frame Source', 'View Frame Info', 'JSONView', and 'Inspect Element'. The page also has a sidebar with 'trending' and 'categories' lists, and a 'Random Word' button.

www.urbandictionary.com/define.php?term=antistalking

look up any word, like swag:
antistalking search

word of the day categories favorites dictionary game thesaurus names media store add blog

random A B C D E F G H I J K L M N O P Q R S T U V W X Y Z # new tv

trending
ratchet
swag
cenosillicaphobia
hipster
douchebag
molly
purp
hashtag
paratrooping
yolo

categories
gaming
sports
food
sex
tv
film
celebrities
military
music
weather
insults

alphabetical
antispamdexing
antispaming
Anti-Spaniel
anti spick
antispire
Antisplash
anti-splash paper
Anti-Spoiler
Anti-spooning
ANTI-SPORTIST
anti-spouse
anti spunkieside
anti square

1. antistalking
word of the day: October 21, 2013
Methodically learning another person's routine in order to systematically avoid them.
Bradley is antistalking his boss today because he hates the bastard.
[anti-stalking](#) [anti stalking](#) [stalking](#) [stalker](#) [antistalker](#)
by [Clearly](#) Oct 13 [add a video](#)
795 up, 171 down

Random Word

Back
Forward
Reload
Save As...
Print...
Translate to English
View Page Source
View Page Info
Reload Frame
View Frame Source
View Frame Info
JSONView
Inspect Element

Find Unique HTML Tags

```
275 href="/define.php?term=antistereotypical">antistereotypical</a></li><li><a href="/define.php?term=anti-stiffy">anti-stiffy</a></li><li><a href="/define.php
276 term=antistimulatingtothemaxcoreness">antistimulatingtothemaxcoreness</a></li><li><a href="/define.php?term=AntiStrachanTwat">AntiStrachanTwat</a></li><li>
277 term=Anti%20Straight">Anti Straight</a></li><li><a href="/define.php?term=Antistrophe">Antistrophe</a></li><li><a href="/define.php?term=Anti-Stubby">Anti-Stubby</a></li><li><a
278 href="/define.php?term=Anti-Sue">Anti-Sue</a></li><li><a href="/define.php?term=anti-sunist">anti-sunist</a></li><li><a href="/define.php?term=Anti%20Superdry%20Faggot">Anti Superdry Faggot</a>
279 </li><li><a href="/define.php?term=Antisupporters">Antisupporters</a></li><li><a href="/define.php?term=Antiswag">Antiswag</a></li><li><a href="/define.php?term=Anti%20Swag">Anti Swag</a></li>
280 <li><a href="/define.php?term=Anti-Swag">Anti-Swag</a></li><li><a href="/define.php?term=Antiswagg">Antiswagg</a></li><li><a href="/define.php?term=anti-swaggasaurus%20rex">anti-swaggasaurus
281 rex</a></li><li><a href="/define.php?term=Anti-Swag Wall">Anti-Swag Wall</a></li><li><a href="/define.php?term=anti-swapatunity">anti-swapatunity</a></li><li><a href="/define.php?term=Anti-
282 SxE">Anti-SxE</a></li>
283 </ul>
284 </div>
285 </div>
286 </div>
287 <div class='span6'>
288 <div id='content'>
289 <!-- google_ad_section_start --><div id='entries'>
290 <div class='word' data-defid='7230994'>
291 <a class="index" href="http://antistalking.urbanup.com/7230994">1.</a>
292 <span>
293 antistalking
294 </span>
295 </div>
296 <div class='text' colspan='2' id='entry_7230994'>
297 <div class='daily_date'>
298 word of the day: October 21, 2013
299 </div>
300 <div class="definition">Methodically learning another person's routine in order to systematically avoid them.</div><div class="example">Bradley is antistalking his boss today because he
301 hates the bastard.</div>
302 <div class='zazzle_links'>
303 <a class="add_to_list" data-defid="7230994" href="#">mark as favorite</a>
304 <a class="zazzle_link" href="/products.php?term=antistalking&defid=7230994"><span class="zazzle_link_text">buy antistalking mugs &amp; shirts</span></a>
305 </div>
306 <div class='greenery'>
307 <span class='tags'>
308 <a href="/define.php?term=anti-stalking">anti-stalking</a>
309 <a href="/define.php?term=anti%20stalking">anti stalking</a>
310 <a href="/define.php?term=stalking">stalking</a>
311 <a href="/define.php?term=stalker">stalker</a>
312 <a href="/define.php?term=antistalker">antistalker</a>
313 </span>
314 <br>
315 by <a class="author" href="/author.php?author=Clearly">Clearly</a>
316 <span class='date'>
317 Oct 13
318 </span>
319 <a href="/video.php?word=antistalking&defid=7230994">add a video</a>
320 <a class="add_an_image" href="/images.new.php?word=antistalking">add an image</a>
321 </div>
322 <div class='tools' id='tools_7230994'>
323 <span class='status'></span>
324 <span class='thumbs'></span>
325 <div class='addthis_toolbox' style='padding-left: 10px'>
326 <a class="addthis button twitter" href="http://www.addthis.com/bookmark.php?v=300&winname=addthis&pub=ra-50dc926d011f6845&source=tbx-300&lng=en-
327
```

Find Unique HTML Tags

The screenshot shows a web browser window with the URL `www.urbandictionary.com/define.php?term=Barack%20Obama`. The browser's address bar and tabs are visible at the top. The Urban Dictionary website has a blue header with the 'urban' logo and a search bar containing 'Barack Obama'. Below the header is a navigation bar with links like 'word of the day', 'categories', 'favorites', 'dictionary', 'game', 'thesaurus', 'names', 'media', 'store', 'add', and 'blog'. A secondary navigation bar shows the alphabet from 'random' to 'tv'. On the left side, there are sections for 'trending' words (ratchet, swag, etc.), 'categories' (gaming, sports, etc.), and 'see also' (obama, president, etc.). The main content area displays the definition for 'Barack Obama'.

1. Barack Obama

The first half-white president.

I bet you only thought of Barack Obama as black, didn't you? Shame.

president 43 black white mixed race

by PlutoRoman Jun 13 add a video

5363 up, 1863 down

Random Word

2. Barack Obama

A puppet picked by the Global Elite to fastened the process of globalization. He is on the Council of Foreign Relations which advocates world government throught its think tank. It published Foreign Affairs magazine. He is not the "Radical Ismalist anti-christ" that ignorant right-wingers claim and is definately not the "most radical democrat" that will end the Iraq occupation. He is but another pawn in place to perpetuate the right and left mirage in US domestic politics. His foreign policy is just as aggressive and detrimental as any other puppets.

Ignorant people will vote for him because his empty rhetorical speeches promise change (while he lives his life by the status quo) that he will never institute. He is an uncle tom in every sense. Any black man voting for him should follow him closely before selling out.

Find Unique HTML Tags

```
Roll</a></li><li><a href="/define.php?term=Barack%20N%20Roll">Barack N Roll</a></li><li><a href="/define.php?term=Baracko">Baracko</a></li><li class="active">
href="/define.php?term=Barack%20Obama">Barack Obama</a></li><li><b><a href="/define.php?term=Barack%20%5BObama%5D">Barack [Obama]</a></li><li><b><a
term=Barack%20%27Bama">Barack O%39;Bama</a></li><li><a href="/define.php?term=Barack%20Obamas">Barack Obamas</a></li><li><a href="/define.php?term=Barack%20Obama%27s%20Assworm">Barack
Obama%39;s Assworm</a></li><li><a href="/define.php?term=Barack%20%27bamasaur">Barack O%39;bamasaur</a></li><li><a href="/define.php?term=Barack%20Obama%27s%20fault">Barack Obama%39;s
fault</a></li><li><a href="/define.php?term=Barack%20Obama%20Syndrome">Barack Obama Syndrome</a></li><li><a href="/define.php?term=Barack%20Obama%20Weed">Barack Obama Weed</a></li><li><a
href="/define.php?term=Barack%20obameter">barack obameter</a></li><li><a href="/define.php?term=Barack%20Obhudda">Barack Obhudda</a></li><li><a href="/define.php?term=Barack-O-Bomb">Barack-O-
Bomb</a></li><li><a href="/define.php?term=Barack%20Obonga">Barack Obonga</a></li><li><a href="/define.php?term=barackocracy">barackocracy</a></li><li><a href="/define.php?
term=Barack%20%27Got%20Em">Barack O%39;Got Em</a></li><li><a href="/define.php?term=Barackolypse">Barackolypse</a></li><li><a href="/define.php?term=barackonaut">barackonaut</a></li><li><a
href="/define.php?term=Barackonomics">Barackonomics</a></li><li><a href="/define.php?term=Barackool">Barackool</a></li><li><a href="/define.php?term=barack-o-pop">barack-o-pop</a></li><li><a
href="/define.php?term=Barack%20Osama">Barack Osama</a></li><li><b><a href="/define.php?term=Barack%20out%20with%20your%20cock%20out">Barack out with your cock out</a></li><li><a
href="/define.php?term=Barackracy">Barackracy</a></li><li><a href="/define.php?term=barackrisky">barackrisky</a></li><li><a href="/define.php?term=Barack%20Roll">Barack Roll</a></li>
</ul>
</div>
</div>
</div>
<div class='span6'>
<div id='content'>
<!-- google_ad_section_start --><div id='entries'>
<div class='word' data-defid='5031878'>
<a class='index' href='http://barack-obama.urbanup.com/5031878'>1.</a>
<span>
Barack Obama
</span>
</div>
<div class='text' colspan='2' id='entry_5031878'>
<div class='definition'>The first half-white president.</div><div class='example'>I bet you only thought of Barack Obama as black, didn't you? Shame.</div>
<div class='zazzle_links'>
<a class='add_to_list' data-defid='5031878' href='#'>mark as favorite</a>
<a class='zazzle_link' href='/products.php?term=Barack%20Obama&defid=5031878'><span class='zazzle_link_text'>buy barack obama mugs & shirts</span></a>
</div>
<div class='greenery'>
<span class='tags'>
<a href="/define.php?term=president">president</a>
<a href="/define.php?term=43">43</a>
<a href="/define.php?term=black">black</a>
<a href="/define.php?term=white">white</a>
<a href="/define.php?term=mixed%20race">mixed race</a>
</span>
<br>
by <a class='author' href='/author.php?author=PlutoRoman'>PlutoRoman</a>
<span class='date'>
Jun 13
</span>
<a href="/video.php?word=Barack+Obama&defid=5031878'>add a video</a>
<a class='add_an_image' href='/images.new.php?word=Barack+Obama'>add an image</a>
</div>
<div class='tools' id='tools_5031878'>
<span class='status'></span>
<span class='thumbs'></span>
<div class='addthis_toolbox' style='padding-left: 10px'>
<a class='addthis_button_twitter' href='http://www.addthis.com/bookmark.php?v=300&winname=addthis&pub=ra-50dc926d011f6845&source=tbx-300&lng=en-US&s=twitter&url=http%3A%2F%2Fbarack-obama.urbanup.com%2F5031878&title=Urban%20Dictionary%3A%20beat%20time' target='_blank'></a>
```


Step 4: Write code to extract data

- Given the HTML string, how do we pull out just the information between our target tags?
- Ex.
 - `html = 'This is a long block of text blah blah
<div class="definition">this part is important!</div>
more stuff we could care less about'`
- We need more string manipulation!

Advanced String Manipulation

- `string.find(s)` returns the index of `s` in the given string
 - index means the number of characters from the beginning of the string
 - `'hello'.find('h')` # returns 0
 - `'hello'.find('e')` # returns 1
 - `'hello'.find('l')` # returns 2
 - `'hello'.find('o')` # returns 4
 - `'hello'.find('hell')` # returns 0
 - `'hello'.find('llo')` # returns 3
 - `'hello'.find('b')` # returns -1

Advanced String Manipulation

- `string[a:b]` returns the substring of string from a (inclusive) to b (non-inclusive)
 - can also use `string[a:]` to go through end of word or `string [:b]` to go from beginning of word
 - `'happiness'[0:2]` # returns 'ha'
 - `'happiness'[1:4]` # returns 'app'
 - `'happiness'[-4:-1]` # returns 'nes'
 - `'happiness'[-4:]` # returns 'ness'
 - `'happiness'[:4]` # returns 'happ'

Advanced String Manipulation

- `string.replace(old, new)` replaces all occurrences of `old` with `new` in `string`
 - `"brown fox".replace("brown", "red")` # returns 'red fox'
 - `"brown fox".replace("fox", "cow")` # returns 'brown cow'
 - `"brown fox".replace(" ", ",")` # returns 'brown,fox'
 - `"brown fox".replace("o", "a")` # returns 'brawn fax'

Advanced String Manipulation

```
words = 'hello there'
print words.split(' ')           # ['hello', 'there']
print words.find('the')          # 6
print words[6]                   # 't'
print words[6:9]                  # 'the'
print words.replace('the',       # 'hello sartre'
                    'sart')
```

Where to look stuff up

- Documentation: <http://docs.python.org/2/library/string.html>
 - Ctrl-F and search for “string.find” in the docs
- Google!
 - ex. “how to split a string in python”

Step 4: Write code to extract data

```
html = 'This is a long block of text blah blah  
<div class="definition">this part is important!</div>  
more stuff we could care less about'  
start_idx = html.find('<div class="definition">')  
definition_substring = html[start_idx:]  
print definition_substring
```

What will this print?

```
<div class="definition">this part is important!</div> more  
stuff we could care less about'
```

Step 4: Write code to extract data

```
html = 'This is a long block of text blah blah  
<div class="definition">this part is important!</div>  
more stuff we could care less about'  
start_idx = html.find('<div class="definition">')  
start_idx = start_idx + len('<div class="definition">')  
definition_substring = html[start_idx:]  
print definition_substring
```

What will this print?

this part is important!</div> more stuff we could care less
about'

Step 4: Write code to extract data

```
html = 'This is a long block of text blah blah  
<div class="definition">this part is important!</div>  
more stuff we could care less about'  
start_idx = html.find('<div class="definition">')  
start_idx = start_idx + len('<div class="definition">')  
definition_substring = html[start_idx:]  
end_idx = definition_substring.find('</div>')  
definition_substring = definition_substring[:end_idx]  
print definition_substring
```

Step 5: Display or store information

- For now, just print it out!
- In upcoming weeks we will learn how to use Python to save and load text files for long-term use

Another Pop Quiz!

