**import** **urllib**

thisurl = "http://www-rohan.sdsu.edu/~gawron/index.html"

handle = urllib.urlopen(thisurl)

html\_gunk = handle.read()

**Using HTMLParser**

Once you have your hands on some raw html downloaded from the web, you are probably interested in parsing it, using an HTML parser, and then extracting some piece of content you know is available.

**import** **HTMLParser**

**import** **urllib**

urlText = []

*#Define HTML Parser*

**class** **parseText**(HTMLParser.HTMLParser):

**def** handle\_data(self, data):

**if** data != '**\n**':

urlText.append(data)

*#Create instance of HTML parser*

lParser = parseText()

thisurl = "http://www-rohan.sdsu.edu/~gawron/index.html"

*#Feed HTML file into parser*

lParser.feed(urllib.urlopen(thisurl).read())

lParser.close()

**for** item **in** urlText:

**print** item

# Simple scraping

In the previous sections we introduced the two basic steps of web-scraping: downloading some data and parsing it into text. The downside of doing it the way we did it there was that we threw away all the HTML and ended up with pure strings. But in many cases that will give you far more than you want. For example, you will get all the ad copy, or, more likely, all the code (in various web frameworks) that loads the ad copy. Sifting through this to find what you want can be quite tedious and, if significant amounts of data are being downloaded, it can be impossible.

In this section we’ll have a look at a slightly different approach, using an HTML parser to give you a tree that preserves the structure of the original HTML, and then searching that tree for the data you want. For this approach, we’ll use a slightly easier to use HTML parser, provided by the lxml module. We’ll start by using an example from the Python for Data Analysis book. Here’s the code for that example:

**from** **lxml.html** **import** parse

**from** **urllib2** **import** urlopen

**from** **pandas.io.parsers** **import** TextParser

**def** \_unpack (row,kind='td'):

elts = row.findall('.//*%s*' % kind)

**return** [val.text **for** val **in** elts]

**def** parse\_options\_data (table):

rows = table.findall('.//tr')

header = \_unpack(rows[0],kind='th')

data = [\_unpack(r) **for** r **in** rows[1:]]

**return** TextParser(data,names=header).get\_chunk()

**if** \_\_name\_\_ == '\_\_main\_\_':

*#parsed = parse('http://finance.yahoo.com/q/op?s=AAPL+Options')*

*#parsed = parse('http://www-rohan.sdsu.edu/~gawron')*

*#parsed = parse('http://www.lajollasurf.org/cgi-bin/plottide.pl')*

url = 'http://www.ezfshn.com/tides/usa/california/san*%20d*iego'

parsed = parse(url)

*#id="ctl00\_ctl00\_Content\_MCC\_RadDatePicker\_calendar\_Top"*

doc = parsed.getroot()

links = doc.findall('.//a')

links\_sub\_list = links[15:20]

lnk = links\_sub\_list[0]

sample\_url = lnk.get('href')

sample\_display\_text = lnk.text\_content()

tables = doc.findall('.//table')

*## Look at tables, find a table of interest*

*#puts = tables[9]*

*## Ditto*

*#calls = tables[13]*

dt = tables[0]

rows = dt.findall('.//tr')

headers = \_unpack(rows[0],kind='th')

row\_vals = \_unpack(rows[1],kind='td')

*#call\_data = parse\_options\_data(calls)*

tide\_data = parse\_options\_data(dt)

**print** tide\_data[:10]

# Web Crawling

In this section we look at some basic operations to implement in a web crawler.

We start with the simple case of a very special format, RSS feeds, and the process of downloading clean text from an RSS feed.

## Newsfeeds

**import** **os**

**import** **sys**

**import** **feedparser**

**from** **BeautifulSoup** **import** BeautifulStoneSoup

**from** **nltk** **import** clean\_html

**import** **urllib**

*#import HTMLParser*

**def** cleanHtml(html):

**return** BeautifulStoneSoup(clean\_html(html),

convertEntities=BeautifulStoneSoup.HTML\_ENTITIES).contents[0]

**def** get\_feedparser\_feed(FEED\_URL):

fp = feedparser.parse(FEED\_URL)

**if** fp **and** fp.entries **and** fp.entries[0]:

print "Fetched *%s* entries from '*%s*'" % (len(fp.entries), fp.feed.title)

**else**:

print 'No entries parseed!'

sys.exit()

**return** fp

**def** get\_blog\_posts(fp):

**global** feed\_dict, blog\_posts

blog\_posts = []

**for** e **in** fp.entries:

**try**:

content = e.content[0]

**except** AttributeError:

content = e.summary\_detail

feed\_dict = {'title': e.title,

'content': cleanHtml(content.value),

'link': e.links[0].href}

blog\_posts.append(feed\_dict)

**return** blog\_posts

FEED\_URL = 'http://feeds.feedburner.com/oreilly/radar/atom'

fp = get\_feedparser\_feed(FEED\_URL)

blog\_posts = get\_blog\_posts(fp)

Fetched 15 entries **from** 'O'Reilly Radar - Insight, analysis, **and** research about emerging technologies'

## Cleaning HTML

The example above involved cleaning some HTML. Here are some of the steps.

e = fp.entries[0]

content = e.content[0]

raw\_html = content.value

c\_html = cleanHtml(raw\_html)

Each feed object is a json format representation of the content: a Python dict with string keys and dictionary, list, and string values. For example :samp: e:

{'author': u'Matthew Gast',

'author\_detail': {'name': u'Matthew Gast'},

'authors': [{'name': u'Matthew Gast'}],

'content': [{'base': u'<http://radar.oreilly.com/2014/02/bluetooth-low-energy-in-public-spaces.html>',

'language': None,

'type': u'text/html',

'value': u'<p>I&#8217;ve been thinking a lot about the <a href="<http://en.wikipedia.org/wiki/Bluetooth_low_energy>">new low-energy form of Bluetooth</a>xa0(BLE) recently, with an eye toward thinking about ways it can be used.xa0The core advantages the protocol has over other similar standards is that it&#8217;s optimized for lower data rates, and extremely long battery life. While we may complain about how much energy a Wi-Fi device uses, it&#8217;s acceptable to charge your phone once a day. If we could eliminate the need to recharge, what lower-data rate applications could we build?</p>n<p>The most obvious application of something like BLE is that it communicates over a shorter range, and therefore, can provide precise location information. Companies like <a href="<http://euclidanalytics.com/>">Euclid Analytics</a> measure foot traffic by using Wi-Fi signals, so the precision of the location is fairly rough. BLE devices have a smaller operating range, and thus would be able to provide information on what aisle a person is in instead of a broad area of the store.xa0(And yes, there are obvious privacy concerns here, especially given that many users tend to accept all the privileges requested by an app running on their phone, which might make BLE-enabled location personally identifiable.)<span id="more-58546"></span></p>n<p>An alternative way of using BLE to support purchasing is that it allows locations to describe themselves.xa0Say, for example, that my favorite pizza place has put in a BLE &#8220;beacon&#8221; that announces itself to the world, and they have created an app for my smartphone that lets me order. I place the order for a pizza with my app. When I walk into the store, my phone is listening for the BLE beacon, wakes up, and uses a network connection (either Wi-Fi or 4G) to tell the store computer that I have arrived to pick up order number whatever, and it charges a credit card on file to give me an electronic receipt.xa0Charging could be done through a merchant account, or even PayPal or Square, which would make it easier for a small take-out restaurant to perform this task.xa0At a busy pizza place, not having to handle payments can speed up the line substantially.xa0(Privacy is a substantially smaller concern in this case because there isn&#8217;t much to analyze that the company would not have already known from their cash register.) In this case, BLE is used to wake up a higher-speed interface so that the higher-speed, and likely higher-power, interface can stay asleep for much of the time.</p>n<p>Many busy restaurants hand out wireless discs to tell you when your table is ready.xa0In a slightly different version of the BLE-triggered payment, when you walk into a restaurant, your phone could notice it is in a restaurant due to the presence of a BLE beacon.xa0At this point, an app on the phone could automatically check you in with the host.xa0Conceivably, an app could monitor on-time performance of seating for the restaurant&#8217;s management.</p>n<p>A more generic form of queue management would be to monitor when you enter a line and when you exit by using BLE devices at the entry and exit points of the queue.xa0Banks, for instance, might be interested in using queue analytics to determine the right number of tellers to staff.xa0As a frequent traveler, real-time measurements of wait times at security checkpoints would be fabulous as well.</p>n<p>BLE is exciting because it&#8217;s a relatively inexpensive technology that can allow applications to gather highly detailed information about the physical world. Apps can learn about where they are and what they are near without needing to rely on a massive GPS database, and mobile devices can gather data that today is too labor-intensive to create and too difficult to report in real time.</p>n<hr />n<p><em>If you are interested in the collision of hardware and software, and other aspects of the convergence of physical and digital worlds, subscribe to the <a href="<http://www.oreilly.com/solid/solid-newsletter.csp>">free Solid Newsletter</a> u2014 and to learn more about the Solid Conference coming to San Francisco in May, visit <a href="<http://solidcon.com/solid2014>">the Solid website</a>.</em></p>n<div class="feedflare">n<a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:V_sGLiPBpWU>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?i=SAF1pnI2Uvg:tCKXPbLIfSI:V_sGLiPBpWU>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:yIl2AUoC8zA>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=yIl2AUoC8zA>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:JEwB19i1-c4>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?i=SAF1pnI2Uvg:tCKXPbLIfSI:JEwB19i1-c4>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:7Q72WNTAKBA>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=7Q72WNTAKBA>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:qj6IDK7rITs>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=qj6IDK7rITs>" /></a>n</div><img height="1" src="<http://feeds.feedburner.com/~r/oreilly/radar/atom/~4/SAF1pnI2Uvg>" width="1" />'}],

'dc\_type': u'text',

'feedburner\_origlink': u'<http://radar.oreilly.com/2014/02/bluetooth-low-energy-in-public-spaces.html>',

'guidislink': False,

'id': u'<http://radar.oreilly.com/?p=58546>',

'link': u'<http://feedproxy.google.com/~r/oreilly/radar/atom/~3/SAF1pnI2Uvg/bluetooth-low-energy-in-public-spaces.html>',

'links': [{'href': u'<http://feedproxy.google.com/~r/oreilly/radar/atom/~3/SAF1pnI2Uvg/bluetooth-low-energy-in-public-spaces.html>',

'rel': u'alternate',

'type': u'text/html'}],

'published': u'2014-02-28T12:00:37Z',

'published\_parsed': time.struct\_time(tm\_year=2014, tm\_mon=2, tm\_mday=28, tm\_hour=12, tm\_min=0, tm\_sec=37, tm\_wday=4, tm\_yday=59, tm\_isdst=0),

'summary': u'I&#8217;ve been thinking a lot about the new low-energy form of Bluetoothxa0(BLE) recently, with an eye toward thinking about ways it can be used.xa0The core advantages the protocol has over other similar standards is that it&#8217;s optimized for lower data &#8230;',

'summary\_detail': {'base': u'<http://feeds.feedburner.com/oreilly/radar/atom>',

'language': None,

'type': u'text/html',

'value': u'I&#8217;ve been thinking a lot about the new low-energy form of Bluetoothxa0(BLE) recently, with an eye toward thinking about ways it can be used.xa0The core advantages the protocol has over other similar standards is that it&#8217;s optimized for lower data &#8230;'},

'tags': [{'label': None,

'scheme': u'[http://radar.oreilly.com](http://radar.oreilly.com/)',

'term': u'Uncategorized'},

{'label': None, 'scheme': u'[http://radar.oreilly.com](http://radar.oreilly.com/)', 'term': u'BLE'},

{'label': None,

'scheme': u'[http://radar.oreilly.com](http://radar.oreilly.com/)',

'term': u'Bluetooth Low Energy'},

{'label': None, 'scheme': u'[http://radar.oreilly.com](http://radar.oreilly.com/)', 'term': u'Solid'},

{'label': None, 'scheme': None, 'term': u'Uncategorized'},

{'label': u'BLE', 'scheme': None, 'term': u'ble'},

{'label': u'Bluetooth Low Energy',

'scheme': None,

'term': u'bluetooth-low-energy'},

{'label': u'Solid', 'scheme': None, 'term': u'solid'}],

'title': u'Bluetooth Low Energy in public spaces',

'title\_detail': {'base': u'<http://feeds.feedburner.com/oreilly/radar/atom>',

'language': None,

'type': u'text/html',

'value': u'Bluetooth Low Energy in public spaces'},

'updated': u'2014-02-28T12:55:53Z',

'updated\_parsed': time.struct\_time(tm\_year=2014, tm\_mon=2, tm\_mday=28, tm\_hour=12, tm\_min=55, tm\_sec=53, tm\_wday=4, tm\_yday=59, tm\_isdst=0)}

The content att of e is a list of dictionaries, each a json format obj. The value attribute contains the content string:

{'base': u'<http://radar.oreilly.com/2014/02/bluetooth-low-energy-in-public-spaces.html>',

'language': None,

'type': u'text/html',

'value': u'<p>I&#8217;ve been thinking a lot about the <a href="<http://en.wikipedia.org/wiki/Bluetooth_low_energy>">new low-energy form of Bluetooth</a>xa0(BLE) recently, with an eye toward thinking about ways it can be used.xa0The core advantages the protocol has over other similar standards is that it&#8217;s optimized for lower data rates, and extremely long battery life. While we may complain about how much energy a Wi-Fi device uses, it&#8217;s acceptable to charge your phone once a day. If we could eliminate the need to recharge, what lower-data rate applications could we build?</p>n<p>The most obvious application of something like BLE is that it communicates over a shorter range, and therefore, can provide precise location information. Companies like <a href="<http://euclidanalytics.com/>">Euclid Analytics</a> measure foot traffic by using Wi-Fi signals, so the precision of the location is fairly rough. BLE devices have a smaller operating range, and thus would be able to provide information on what aisle a person is in instead of a broad area of the store.xa0(And yes, there are obvious privacy concerns here, especially given that many users tend to accept all the privileges requested by an app running on their phone, which might make BLE-enabled location personally identifiable.)<span id="more-58546"></span></p>n<p>An alternative way of using BLE to support purchasing is that it allows locations to describe themselves.xa0Say, for example, that my favorite pizza place has put in a BLE &#8220;beacon&#8221; that announces itself to the world, and they have created an app for my smartphone that lets me order. I place the order for a pizza with my app. When I walk into the store, my phone is listening for the BLE beacon, wakes up, and uses a network connection (either Wi-Fi or 4G) to tell the store computer that I have arrived to pick up order number whatever, and it charges a credit card on file to give me an electronic receipt.xa0Charging could be done through a merchant account, or even PayPal or Square, which would make it easier for a small take-out restaurant to perform this task.xa0At a busy pizza place, not having to handle payments can speed up the line substantially.xa0(Privacy is a substantially smaller concern in this case because there isn&#8217;t much to analyze that the company would not have already known from their cash register.) In this case, BLE is used to wake up a higher-speed interface so that the higher-speed, and likely higher-power, interface can stay asleep for much of the time.</p>n<p>Many busy restaurants hand out wireless discs to tell you when your table is ready.xa0In a slightly different version of the BLE-triggered payment, when you walk into a restaurant, your phone could notice it is in a restaurant due to the presence of a BLE beacon.xa0At this point, an app on the phone could automatically check you in with the host.xa0Conceivably, an app could monitor on-time performance of seating for the restaurant&#8217;s management.</p>n<p>A more generic form of queue management would be to monitor when you enter a line and when you exit by using BLE devices at the entry and exit points of the queue.xa0Banks, for instance, might be interested in using queue analytics to determine the right number of tellers to staff.xa0As a frequent traveler, real-time measurements of wait times at security checkpoints would be fabulous as well.</p>n<p>BLE is exciting because it&#8217;s a relatively inexpensive technology that can allow applications to gather highly detailed information about the physical world. Apps can learn about where they are and what they are near without needing to rely on a massive GPS database, and mobile devices can gather data that today is too labor-intensive to create and too difficult to report in real time.</p>n<hr />n<p><em>If you are interested in the collision of hardware and software, and other aspects of the convergence of physical and digital worlds, subscribe to the <a href="<http://www.oreilly.com/solid/solid-newsletter.csp>">free Solid Newsletter</a> u2014 and to learn more about the Solid Conference coming to San Francisco in May, visit <a href="<http://solidcon.com/solid2014>">the Solid website</a>.</em></p>n<div class="feedflare">n<a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:V_sGLiPBpWU>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?i=SAF1pnI2Uvg:tCKXPbLIfSI:V_sGLiPBpWU>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:yIl2AUoC8zA>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=yIl2AUoC8zA>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:JEwB19i1-c4>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?i=SAF1pnI2Uvg:tCKXPbLIfSI:JEwB19i1-c4>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:7Q72WNTAKBA>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=7Q72WNTAKBA>" /></a> <a href="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?a=SAF1pnI2Uvg:tCKXPbLIfSI:qj6IDK7rITs>"><img border="0" src="<http://feeds.feedburner.com/~ff/oreilly/radar/atom?d=qj6IDK7rITs>" /></a>n</div><img height="1" src="<http://feeds.feedburner.com/~r/oreilly/radar/atom/~4/SAF1pnI2Uvg>" width="1" />'}

The raw\_html objects is just an HTML string: We convert this to c\_html using an important module called Beautiful Soup (BS). In the conversion, many things are done: For example, HTML code points are turned into unicode to give something close to a plain unicode string, links and images removed. The content string is left, but the object BS returns (c\_html) is called a soup object. It is not a simple string, but prints like one.

u'Iu2019ve been thinking a lot about the new low-energy form of Bluetooth xa0(BLE) recently, with an eye toward thinking about ways it can be used.xa0The core advantages the protocol has over other similar standards is that itu2019s optimized for lower data rates, and extremely long battery life. While we may complain about how much energy a Wi-Fi device uses, itu2019s acceptable to charge your phone once a day. If we could eliminate the need to recharge, what lower-data rate applications could we build? n The most obvious application of something like BLE is that it communicates over a shorter range, and therefore, can provide precise location information. Companies like Euclid Analytics measure foot traffic by using Wi-Fi signals, so the precision of the location is fairly rough. BLE devices have a smaller operating range, and thus would be able to provide information on what aisle a person is in instead of a broad area of the store.xa0(And yes, there are obvious privacy concerns here, especially given that many users tend to accept all the privileges requested by an app running on their phone, which might make BLE-enabled location personally identifiable.) n An alternative way of using BLE to support purchasing is that it allows locations to describe themselves.xa0Say, for example, that my favorite pizza place has put in a BLE u201cbeaconu201d that announces itself to the world, and they have created an app for my smartphone that lets me order. I place the order for a pizza with my app. When I walk into the store, my phone is listening for the BLE beacon, wakes up, and uses a network connection (either Wi-Fi or 4G) to tell the store computer that I have arrived to pick up order number whatever, and it charges a credit card on file to give me an electronic receipt.xa0Charging could be done through a merchant account, or even PayPal or Square, which would make it easier for a small take-out restaurant to perform this task.xa0At a busy pizza place, not having to handle payments can speed up the line substantially.xa0(Privacy is a substantially smaller concern in this case because there isnu2019t much to analyze that the company would not have already known from their cash register.) In this case, BLE is used to wake up a higher-speed interface so that the higher-speed, and likely higher-power, interface can stay asleep for much of the time. n Many busy restaurants hand out wireless discs to tell you when your table is ready.xa0In a slightly different version of the BLE-triggered payment, when you walk into a restaurant, your phone could notice it is in a restaurant due to the presence of a BLE beacon.xa0At this point, an app on the phone could automatically check you in with the host.xa0Conceivably, an app could monitor on-time performance of seating for the restaurantu2019s management. n A more generic form of queue management would be to monitor when you enter a line and when you exit by using BLE devices at the entry and exit points of the queue.xa0Banks, for instance, might be interested in using queue analytics to determine the right number of tellers to staff.xa0As a frequent traveler, real-time measurements of wait times at security checkpoints would be fabulous as well. n BLE is exciting because itu2019s a relatively inexpensive technology that can allow applications to gather highly detailed information about the physical world. Apps can learn about where they are and what they are near without needing to rely on a massive GPS database, and mobile devices can gather data that today is too labor-intensive to create and too difficult to report in real time. n n If you are interested in the collision of hardware and software, and other aspects of the convergence of physical and digital worlds, subscribe to the free Solid Newsletter u2014 and to learn more about the Solid Conference coming to San Francisco in May, visit the Solid website .'

What we get from c\_html is a kind of BeautifulSoup object object called a **NavigableString**.

type(c\_html)

BeautifulSoup.NavigableString

dir(c\_html)

['BARE\_AMPERSAND\_OR\_BRACKET',

'XML\_ENTITIES\_TO\_SPECIAL\_CHARS',

'XML\_SPECIAL\_CHARS\_TO\_ENTITIES',

'\_\_add\_\_',

'\_\_class\_\_',

'\_\_contains\_\_',

'\_\_delattr\_\_',

'\_\_dict\_\_',

'\_\_doc\_\_',

'\_\_eq\_\_',

'\_\_format\_\_',

'\_\_ge\_\_',

'\_\_getattr\_\_',

'\_\_getattribute\_\_',

'\_\_getitem\_\_',

'\_\_getnewargs\_\_',

'\_\_getslice\_\_',

'\_\_gt\_\_',

'\_\_hash\_\_',

...

'fetchNextSiblings',

'fetchParents',

'fetchPrevious',

'fetchPreviousSiblings',

'find',

'findAllNext',

'findAllPrevious',

'findNext',

'findNextSibling',

'findNextSiblings',

'findParent',

'findParents',

'findPrevious',

'findPreviousSibling',

'findPreviousSiblings',

'format',

'index',

'insert',

'isalnum',

'isalpha',

'isdecimal',

...

'zfill']

## Beautiful Soup

The Beautiful Soup module is an HTML parser like the one provided by Python’s HTMLParser module, but it is very powerful and flexible.

**from** **BeautifulSoup** **import** BeautifulSoup

html = "<html><p>Para 1<p>Para 2<blockquote>Quote 1<blockquote>Quote 2"

soup = BeautifulSoup(html)

print soup.prettify()

<html>

<p>

Para 1

</p>

<p>

Para 2

<blockquote>

Quote 1

<blockquote>

Quote 2

</blockquote>

</blockquote>

</p>

</html>

That document isn’t valid HTML, but it’s not too bad either. Here’s a really horrible document. Among other problems, it’s got a <FORM> tag that starts outside of a <TABLE> tag and ends inside the <TABLE> tag. (HTML like this was found on a website run by a major web company.)

html = """

<html>

<form>

<table>

<td><input name="input1">Row 1 cell 1

<tr><td>Row 2 cell 1

</form>

<td>Row 2 cell 2<br>This</br> sure is a long cell

</body>

</html>"""

print BeautifulSoup(html).prettify()

<html>

<p>

Para 1

</p>

<p>

Para 2

<blockquote>

Quote 1

<blockquote>

Quote 2

</blockquote>

</blockquote>

</p>

</html>

## International Commercial Crime Services Weekly Piracy Report (parsing example)

**import** **urllib2**

**from** **BeautifulSoup** **import** BeautifulSoup

page = urllib2.urlopen("http://www.icc-ccs.org/piracy-reporting-centre/live-piracy-report")

soup = BeautifulSoup(page)

**for** incident **in** soup('td'):

**try**:

incident['class']

**except** KeyError:

**continue**

**if** incident['class'] == u'jos\_fabrik\_icc\_ccs\_piracymap2012\_\_\_narrations fabrik\_element':

where, linebreak, what = incident.contents[:3]

print where.strip()

print what.strip()

print

01.03.2014: 0040 LT: Posn: 22:14N – 091:44E, Chittagong Anchorage, Bangladesh.

Four robbers armed with knives boarded an anchored bulk carrier from the stern. Duty crew spotted the robbers, raised the alarm and escaped as the robbers chased them. All crew mustered and vessel reported to the Bangladesh Coast Guard who sent out a patrol boat. The robbers managed to escape and further checking around the vessel found nothing stolen.

28.02.2014: 1300 LT: Posn: 22:33N – 062:44E, Around 40nm SE off Gwadar, Pakistan.

A bulk carrier underway was chased by a skiff for approximately four hours. The vessel took evasive measures as per BMP4, reported to UKMTO and headed toward the Pakistani coast for assistance. The Pakistani navy deployed a naval asset which located the skiff and detained the suspected pirates.

24.02.2014: 2245 LT: Posn: 22:15.8N - 091:43.2E, Chittagong Anchorage, Bangladesh.

Ten robbers in an unlit wooden boat armed with knives approached an anchored chemical tanker. Two robbers boarded the tanker using grappling hooks and stole ship’s stores and property. The duty A/B noticed the robbers and informed the bridge. Alarm raised, ships whistle sounded and crew rushed to the location. Seeing the alert crew, the robbers jumped overboard with the stolen items and escaped in their boat with their accomplices.

20.02.2014: 0150 LT: Posn: 04:54.0S - 011:49.2E, Pointe Noire Anchorage, The Congo.

Robbers boarded an anchored supply ship using a piece of rope. They stole ship’s properties and escaped when the duty crew spotted them.

20.02.2014: 1140 LT :Posn: 21:00N - 091:37E, Around 25nm off coastline, Bangladesh.

A tug towing a general cargo vessel underway noticed five fishing boats approaching the general cargo vessel. Two fishing boats came alongside and pirates boarded the vessel and were seen lowering the ship's property and stores. At the time of the incident the vessel under tow was not manned as it was underway for scrap.

19.02.2014: 0445 LT: Posn: 03:57N – 005:18E, 26nm SW of Pennington Oil Terminal, Nigeria.

Six pirates in a small boat approached a tanker under way and tried to hook on a boarding ladder. Alarm raised and vessel immediately started taking evasive manoeuvres. The pirates tried to hook on the ladder several times at different positions along the port and starboard quarters. The on board armed security team fired warning shots resulting in the pirates aborting the attempt and moving away.

06.02.2014: 0630 LT: Posn: 01:05N – 103:33E, Singapore Straits.

Seven robbers armed with knives boarded a container ship under way, entered the engine room and tied up the electrical officer. They then stole the engine spares as well as the electrical officers mobile phone. The electrician managed to untie himself and informed the bridge. Ship’s alarm raised and distress message sent out. The robbers escaped with stolen ship’s spares.

06.02.2014: 0615 LT: Posn: 01:03N – 103:36E, Singapore Straits.

Five robbers armed with knives boarded a general cargo ship under way, entered the engine room and aggressively approached the duty crew who immediately left the engine room and informed the bridge. Alarm raised, all crew mustered on the bridge and SSAS activated. Later a complete search of the vessel was carried out.

14.02.2014: 2030 LT: Posn: 05:59.9S – 106:55.6E, Jakarta Roads, Indonesia.

Duty A/B on routine rounds on board an anchored container ship noticed an unlit small wooden boat leaving the stern of the ship. The A/B immediately informed the bridge and the Master raised the alarm. On searching the vessel it was found that engine room stores had been stolen.

06.02.2014: 1055 LT: Posn: 04:01N – 005:01E, Around 75nm WSW of Brass, Nigeria.

Eight armed pirates in a speed boat chased a chemical tanker underway. The vessel raised alarm, made evasive manoeuvres, sent distress message and activated the SSAS alert. The pirates manoeuvred alongside the vessel, and boarded using a long ladder. The crew cut off the power in the ship and retreated into the citadel. After around five hours the crew emerged and noticed the pirates had used sledge hammers to break into stores and cabins. Ship's communication equipment was also destroyed. The crew managed to start the emergency generators and other necessary machinery, informed the owners and sailed the vessel to Lagos.

page = urllib2.urlopen("http://www.icc-ccs.org/piracy-reporting-centre/live-piracy-report")

soup = BeautifulSoup(page)

L = soup('td')

print L[0]

print

print L[2]

<td colspan="3">

<div class="fabrikNav"></div> </td>

<td class="jos\_fabrik\_icc\_ccs\_piracymap2012\_\_\_attack\_no fabrik\_element">

034-14 </td>