Web Scrapping for Dummies
It’s easier than you think

Michael Puthawala

Department of Mathematics
University of California, Los Angeles

April 12, 2017
Outline

Introduction

HTML

Steps to scrapping the web

Example 1

Example 2

Live Demo!
A very good tutorial

I used https://first-web-scraper.readthedocs.io when I wrote my first web scrapper.

This is an excellent tutorial written for people investigative reporters who haven’t necessarily ever used python before.
What is web scrapping?

It’s the process of taking data from the web, and recording it.
What is web scrapping?

It’s the process of taking data from the web, and recording it.

Without having to ask for it.
Motivating example 1

Text corpus
From Wikipedia, the free encyclopedia

It has been suggested that this article be merged into Corpus linguistics. (Discuss) Proposed since March 2015.

This article includes a list of references, but its sources remain unclear because it has insufficient inline citations.
Please help to improve this article by introducing more precise citations. (December 2009) (Learn how and when to remove this template message)

In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Overview

A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus).

Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora. There are two main types of parallel corpora...
Motivating example 2

<table>
<thead>
<tr>
<th>Date</th>
<th>Featured Artists</th>
<th>Venue</th>
<th>City</th>
<th>State</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>02/26/1970</td>
<td>Teagarden &amp; Van Winkle</td>
<td>Red Carpet Lounge</td>
<td>Detroit</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>02/26/1970</td>
<td>Josh White, Jr.</td>
<td>Pease Auditorium</td>
<td>Ypsilanti</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>02/26/1970</td>
<td>Savoy Brown, Renaissance</td>
<td>Eastern Theater</td>
<td>Detroit</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>The Stooges</td>
<td>Action House</td>
<td>Island Park</td>
<td>New York</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Rare Earth</td>
<td>The Delts</td>
<td>Hazellet</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Rare Earth</td>
<td>The Delts</td>
<td>Hazellet</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>MCS</td>
<td>Michigan Theater (D)</td>
<td>Detroit</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Front</td>
<td>Felt Forum</td>
<td>NYC</td>
<td>New York</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Commander Cody &amp; His Lost Planet Airmen</td>
<td>Family Dog</td>
<td>San Francisco</td>
<td>California</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>The Rationals</td>
<td>Sherwood Forest</td>
<td>Davison</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Teagarden &amp; Van Winkle</td>
<td>Red Carpet Lounge</td>
<td>Detroit</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Universal Family</td>
<td>Grandmother’s</td>
<td>Lansing</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Rare Earth</td>
<td>The Delts</td>
<td>Hazellet</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Chubby Checker</td>
<td>Coral Gables (EL)</td>
<td>East Lansing</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/04/1970</td>
<td>Universal Family</td>
<td>Grandmother’s</td>
<td>Lansing</td>
<td>Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>03/05/1970</td>
<td>Rare Earth</td>
<td>The Delts</td>
<td>Hazellet</td>
<td>Michigan</td>
<td>USA</td>
</tr>
</tbody>
</table>
Hypertext Markup Language (HTML) is the most common markup language on the internet. It is one of the languages that tell your browser what data is on a web page, and how to display it.

If you have ever browsed the internet, then you’ve come across HTML.
HTML is a markup language like Latex, so the HTML that makes up a page and the page itself are related, but often it's not at all obvious exactly how they are related.
Consider this web page, to its HTML code...
Let’s try and get a sense of what’s going on in the HTML with this very simple page.

This is a header

This is paragraph 1.

This is paragraph 2.

This is paragraph 3.

- First element in an unordered list.
- Second element in an unordered list.
  1. Ordered list element one.
  2. Ordered list element two.
  3. Ordered list element three.
- Third element in an unordered list.
Simple Page’s HTML

Here is the HTML that generated the previous webpage

```html
<!DOCTYPE html>
<html>
  <body>
    <h1>This is a header</h1>
    <p>This is paragraph 1.</p>
    <p>This is paragraph 2.</p>
    <p>This is paragraph 3.</p>
    <ul>
      <li>First element in an unordered list.</li>
      <li>Second element in an unordered list.</li>
      <ol>
        <li>Ordered list element one.</li>
        <li>Ordered list element two.</li>
        <li>Ordered list element three.</li>
      </ol>
      <li>Third element in an unordered list.</li>
    </ul>
  </body>
</html>
```
HTML forms a tree

HTML has a natural tree structure, and this fact will help us down the line when we are trying to navigate this tree to scrape data.
Steps to making an automated web scraper

- Retrieve the web page from a given source.
Steps to making an automated web scraper

- Retrieve the web page from a given source.
- Parse that code of the web page.
Steps to making an automated web scraper

- Retrieve the web page from a given source.
- Parse that code of the web page.
- Extract the important data from that page.
Steps to making an automated web scraper

- Retrieve the web page from a given source.
- Parse that code of the web page.
- Extract the important data from that page.
- Record the data.
Steps to making an automated web scraper

- Retrieve the web page from a given source.
- Parse that code of the web page.
- Extract the important data from that page.
- Record the data.
- Move to a different page and repeat.
Retrieving the HTML

This step is actually amazingly easy if you use the python module request, and don’t have to do any fancy authentication/logging on.
Retrieving the HTML

This step is actually amazingly easy if you use the python module request, and don’t have to do any fancy authentication/logging on.

Demo Time!
import requests

print('Requesting the HTML form: ' + 'https://en.wikipedia.org/wiki/HTML')
HTML_code = requests.get('https://en.wikipedia.org/wiki/HTML')
print('Saving HTML')

f = open('a_webpage.html', 'w+b')
f.write(HTML_code.text.encode('ascii', 'ignore'))
f.close()
Parsing the HTML

This step is extremely difficult, as it requires expert knowledge of HTML, an awareness of 'best practices,' and how to gracefully recover in the face of terribly written HTML.
Parsing the HTML

This step is extremely difficult, as it requires expert knowledge of HTML, an awareness of 'best practices,' and how to gracefully recover in the face of terribly written HTML.

This step would be extremely difficult if not for a very, VERY useful python package BeautifulSoup.

BeautifulSoup takes the response from requests, and converts it into a tree that is relatively easy to traverse.
Extracting the data

This part you really do have to do yourself, but BeautifulSoup makes it bearable.
Extracting the data

This part you really do have to do yourself, but BeautifulSoup makes it bearable. The difficulty of this step depends on
Extracting the data

This part you really do have to do yourself, but BeautifulSoup makes it bearable. The difficulty of this step depends on

- How many disparate pieces of data you want to scrape.
Extracting the data

This part you really do have to do yourself, but BeautifulSoup makes it bearable. The difficulty of this step depends on

- How many disparate pieces of data you want to scrape.
- How regular the formatting is on the source.
Extracting the data

This part you really do have to do yourself, but BeautifulSoup makes it bearable. The difficulty of this step depends on

- How many disparate pieces of data you want to scrape.
- How regular the formatting is on the source.
- If the data is embedded in the HTML, or is instead contained within, for example, a Javascript script running on the site.
Recording the Data

This last step is surprisingly difficult and time consuming.

It’s easy to scrape text very simple text, but what if:
Recording the Data

This last step is surprisingly difficult and time consuming.

It’s easy to scrape text very simple text, but what if:

- The source contains non-UTF-8 characters? (e.g. é, 😊).
Recording the Data

This last step is surprisingly difficult and time consuming.

It’s easy to scrape text very simple text, but what if:

- The source contains non-UTF-8 characters? (e.g. é, 😄).
- What of the source contains hypertext? (e.g. 'Check out the colors on http://bit.ly/2h80Lb6').
Recording the Data

This last step is surprisingly difficult and time consuming.

It’s easy to scrape text very simple text, but what if:

- The source contains non-UTF-8 characters? (e.g. é, 😊).
- What of the source contains hypertext? (e.g. 'Check out the colors on http://bit.ly/2h80Lb6').
- The data includes important non-text elements (e.g. pictures, videos, etc...).
Finding the Next Source

For all of the applications that I have tried this step has been pretty easy (e.g. find the next page in the table) but it could easily be quite difficult (e.g. many different links to the same source, lots of irrelevant/useless sources.)
Last quarter I worked on a project where I had to collect data from http://theconcertdatabase.com/concerts.
Checklist

- Is it HTML easy to get?
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
- Is the data easy to record? Mostly.
- Is it easy to move from one page of data to another? Yes.
Checklist

- Is it HTML easy to get? Yup.
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse?
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular?
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
- Is the data easy to record?
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
- Is the data easy to record? Mostly.
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
- Is the data easy to record? Mostly.
- Is it easy to move from one page of data to another?
Checklist

- Is it HTML easy to get? Yup.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Yes, extremely so.
- Is the data easy to record? Mostly.
- Is it easy to move from one page of data to another? Yes.
Pseudocode

for i = 1..m
    request HTML from 'theconcertdatabase.com/concerts?field_date_value=&page='+i
    parse HTML
    find anything on that page that looks like a row in a table
    for each row in table:
        for each cell in row:
            record the contents of each cell

    record that row to an output file
09/29/2006, Kamelot, Harpos, Detroit, Michigan, USA
09/29/2006, Kasabian, "Saint Andrew's Hall", Detroit, Michigan, USA
09/29/2006, "Plain White T's", State Theater, Detroit, Michigan, USA
09/29/2006, Powerman 5000, Temple Club, Lansing, Michigan, USA
09/29/2006, The Mac Podz, Blind Pig, Ann Arbor, Michigan, USA
09/29/2006, Melvins, Magic Stick, Detroit, Michigan, USA
09/29/2006, Insane Clown Posse, The Intersection, Grand Rapids, Michigan, USA
09/29/2006, Mushroomhead, Machine Shop, Flint, Michigan, USA
09/30/2006, Melvins, The Intersection, Grand Rapids, Michigan, USA
09/30/2006, Allister, Shelter, Detroit, Michigan, USA
09/30/2006, ill Nio, Orbit Room, Grand Rapids, Michigan, USA
09/30/2006, Hank III, Machine Shop, Flint, Michigan, USA
09/30/2006, Veruca Salt, Hayloft Liquor Stand, Mount Clemens, Michigan, USA
10/01/2006, The String Cheese Incident, Michigan Theater (AA), Ann Arbor, Michigan, USA
10/02/2006, The Junior Varsity, Temple Club, Lansing, Michigan, USA
10/02/2006, Joseph Arthur, Magic Bag, Ferndale, Michigan, USA
10/03/2006, Everclear, Blind Pig, Ann Arbor, Michigan, USA
10/03/2006, The King Khan & BBQ Show, Kraftbrau Brewery, Kalamazoo, Michigan, USA
10/04/2006, Psychostick, Temple Club, Lansing, Michigan, USA
10/04/2006, Robin Trower, Royal Oak Music Theater, Royal Oak, Michigan, USA
10/04/2006, Yo La Tengo, Michigan Theater (AA), Ann Arbor, Michigan, USA
Another Instructive Example

For that same project, I had to collate the previously mentioned concert data with data which related an artist to their genre.
Another Instructive Example

For that same project, I had to collate the previously mentioned concert data with data which related an artist to their genre.

Where did the genre information come from? The fount of all knowledge:
Another Instructive Example

For that same project, I had to collate the previously mentioned concert data with data which related an artist to their genre.

Where did the genre information come from? The fount of all knowledge: wikipedia!
Kamelot

From Wikipedia, the free encyclopedia

For other uses, see Camelot (disambiguation).

Kamelot is an American power metal band from Tampa, Florida, formed by Thomas Youngblood and Richard Warner in 1991. The Norwegian vocalist Roy Khan joined for the album Siège Perilous, and shared songwriting credit with Youngblood until his departure in April 2011. On June 22, 2012, Youngblood announced on their website that their new vocalist is the Swedish singer Tommy Karevik, who was first featured on Kamelot's album Silverthorn as main vocalist, co-songwriter and lyricist.

As of 2015, Kamelot had released eleven studio albums (excluding two reissues), two live albums, one live DVD, and seventeen music videos. Their latest studio album is Haven, which was released May 2015 via Napalm Records.

Contents [hide]

1 Biography
   1.1 Early years (1988–2000)
   1.3 Ghost Opera, Poetry for the Poisoned and Khan's
Wikipedia has a List of Lists of Musicians that belong to a bunch of genres.

The plan is to use each one of the lists to generate a list of musicians which belong to a given genre.
Checklist

- Is it HTML easy to get?
  - Somewhat. Wikipedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse?
  - Yup, BeautifulSoup!
- Is the data regular?
  - Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.
- Is the data easy to record?
  - Somewhat. The data is just text in the HTML, but many bands use unusual characters.
- Is it easy to move from one page of data to another?
  - Fairly easy by using the list of lists page.
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse?

Yup, BeautifulSoup!
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse? Yup, BeautifulSoup!

- Is the data regular? Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.
- Is the data easy to record? Somewhat. The data is just text in the HTML, but many bands use unusual characters.
- Is it easy to move from one page of data to another? Fairly easy by using the list of lists page.
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular?
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.
- Is the data easy to record?
Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.

Is it HTML easy to parse? Yup, BeautifulSoup!

Is the data regular? Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.

Is the data easy to record? Somewhat. The data is just text in the HTML, but many bands use unusual characters.
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.
- Is the data easy to record? Somewhat. The data is just text in the HTML, but many bands use unusual characters
- Is it easy to move from one page of data to another?
Checklist

- Is it HTML easy to get? Somewhat. Wikiedia is open source, but we have to do some work to find the correct pages.
- Is it HTML easy to parse? Yup, BeautifulSoup!
- Is the data regular? Only somewhat. Most artists have info cards, and most info cards list the relevant information, but not always, and the biographical information listed about each artist changes.
- Is the data easy to record? Somewhat. The data is just text in the HTML, but many bands use unusual characters
- Is it easy to move from one page of data to another? Fairly easy by using the list of lists page.
Pseudocode

given the URL of one of the lists, and a starting & ending bands
list_soup = requested & parsed webpage
list_of_band_URLs = extract_wikipedia_URLs(list_soup)
pruned_list = prune_list(list_of_band_URLs, first_band_name, last_band_name)
for each band_URL in pruned_list:
    band_soup = request and parse band_URL
    info_card = find_info_card(band_soup)
    if (info_card):
        for each row in info_card:
            scrape contents of row
    else:
        continue
<table>
<thead>
<tr>
<th>Jazz Artists</th>
<th>Subgenre1</th>
<th>Subgenre2</th>
<th>Subgenre3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Joplin</td>
<td>Ragtime</td>
<td>March</td>
<td>Waltz</td>
</tr>
<tr>
<td>Buddy Bolden</td>
<td>Jazz</td>
<td>Blues</td>
<td>Swing</td>
</tr>
<tr>
<td>Duke Ellington</td>
<td>Swing</td>
<td>Orchestral</td>
<td>Swing</td>
</tr>
<tr>
<td>Louis Armstrong</td>
<td>Dixieland</td>
<td>Jazz</td>
<td>Swing</td>
</tr>
<tr>
<td>Joe Venuti</td>
<td>Jazz</td>
<td>Big Band</td>
<td>Swing</td>
</tr>
<tr>
<td>Earl Hines</td>
<td>Jazz</td>
<td>Jazz</td>
<td>Swing</td>
</tr>
<tr>
<td>Fats Waller</td>
<td>Dixieland</td>
<td>Swing</td>
<td>Big Band</td>
</tr>
<tr>
<td>Count Basie</td>
<td>Jazz</td>
<td>Swing</td>
<td>Continental</td>
</tr>
<tr>
<td>Stephane Grappelli</td>
<td>Jazz</td>
<td>Swing</td>
<td>Big Band</td>
</tr>
<tr>
<td>Benny Goodman</td>
<td>Jazz</td>
<td>Swing</td>
<td>Big Band</td>
</tr>
<tr>
<td>Art Tatum</td>
<td>Jazz</td>
<td>Stride</td>
<td>Swing</td>
</tr>
<tr>
<td>Sun Ra</td>
<td>Avant-garde</td>
<td>Free Jazz</td>
<td>Jazz Fusion</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q: How Bad Could Reddit Comments Be Anyways?

Did you know that every Reddit comment that you’ve ever made is saved permanently?
Q: How Bad Could Reddit Comments Be Anyways?

Did you know that every Reddit comment that you’ve ever made is saved permanently?
Well it does, and it’s surprisingly easy to find them, and record them.
Q: How Bad Could Reddit Comments Be Anyways?

Did you know that every Reddit comment that you’ve ever made is saved permanently? Well it does, and it’s surprisingly easy to find them, and record them. Without your permission and without you ever knowing.
A: Turns Out Pretty Bad

Let’s see just how easy it is.
But Don’t Take My Word for it

Demo Time!