These slides: **bit.ly/yaydhsi**
(Copy & paste things from here)

Web scraping 201

with wget, Python, and BeautifulSoup

June 15, 2017
DHSI unconference half-a-workshop facilitated by
Robin Davis • @robincamille

↓ See notes in Google Slides for more info ↓

THE BIG IDEA

Web pages are structured documents.

Use that structure to pluck out the bits you're interested in.

These slides: **bit.ly/yaydhsi**
(Copy & paste things from here)

```
<details>
<summary>

DHSI Classes in Session


(click for details and locations)
</summary>
<br>
<li>25. [Foundations] Intro to Computation for Literary Criticism (<a href="https://www.uvic.ca/home/about/
<li>26. [Foundations] Developing a Digital Project (With Omeka) (<a href="https://www.uvic.ca/home/about/ca
<li>27. [Foundations] Models for DH at Liberal Arts Colleges (&amp;amp; 4 yr Institutions) (<a href="https://ww
<li>28. [Foundations] Introduction to Javascript and Data Visualization (<a href="https://www.uvic.ca/home/
<li>29. Wrangling Big Data for DH (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/cle.php">C
<li>30. Stylometry with R: Computer-Assisted Analysis of Literary Texts (<a href="https://www.uvic.ca/home/
<li>31. Sounds and Digital Humanities (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/mac.ph
<li>32. Digital Humanities Pedagogy: Integration in the Curriculum (<a href="https://www.uvic.ca/home/about
<li>34. Creating LAMP Infrastructure for Digital Humanities Projects (<a href="https://www.uvic.ca/home/abo
<li>35. Understanding Topic Modeling (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/mac.php
<li>36. Palpability and Wearable Computing (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/m
<li>37. Building a Professional Identity and Skillset in the Digital Humanities (<a href="https://www.uvic.
<li>38. Digital Editing with TEI: Critical, Documentary and Genetic Editing (<a href="https://www.uvic.ca/h
<li>40. Understanding Digital Video (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/mac.php"
<li>41. Beyond TEI: Metadata for Digital Humanities (<a href="https://www.uvic.ca/home/about/campus-info/ma
<li>42. Extracting Cultural Networks from Thematic Research Collections (<a href="https://www.uvic.ca/home/
<li>43. Digital Public Humanities (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/mac.php">M
<li>44. Using Fedora Commons / Islandora (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/hsd
<li>45. Practical Software Development for Nontraditional Digital Humanities Developers (<a href="https://w
<li>46. Documenting Born Digital Creative and Scholarly Works for Access and Preservation (<a href="https:/
<li>47. An Introduction to Computational Humanities: Mining, Machine Learning and Future Challenges (<a hre
<li>48. Games for Digital Humanists (<a href="https://www.uvic.ca/home/about/campus-info/maps/maps/dsb.php"
<li>49. Introduction to XSLT for Digital Humanists (<a href="https://www.uvic.ca/home/about/campus-info/map
<br><br>
</details>
```

Sample workflow...

1. Save HTML files of a bunch of web pages
2. Extract just the bits you're interested in, like the "raw text" of the main content
3. Save extracted bits to .txt files
4. Do DH–y things

~ Live demo ~

Get ready to install 3 things!!!!

(These instructors are for Mac, sorry everybody else — you can do the same things but installation will be different)

Install Homebrew

Open Terminal

Paste this in:

/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"

What is? Homebrew helps you install software from the command line (aka Terminal). You'll need it for the next step. And it's just great to have.

Install wget

In Terminal, type this:

brew install wget

(You may have to type **sudo brew install wget** if you get errors about permissions stuff)

What is? wget is a very basic & very effective web scraping tool used on the Command Line. It will download webpages on command.

Install Beautiful Soup

In Terminal, type this:

pip install beautifulsoup4

(You may have to type **sudo pip install beautifulsoup4** if you get errors about permissions stuff)

What is? Beautiful Soup is a Python library that processes the text of web pages. Example of use: Input: HTML file you just downloaded; output: just the text inside tags

Let’s scrape some web!!!

Scrape the DHSI course list page with wget

In Terminal, type this:

wget -E http://www.dhsi.org/courses.php

Extract course abstracts from the HTML

Open IDLE or your fave Python editor, paste this in, and run it:

```
from bs4 import BeautifulSoup
with open('courses.php.html') as doc:
    text = BeautifulSoup(doc,"html.parser")

    summaries = text.find_all('details')
    for s in summaries[:10]:
    print s.get_text()
```

See more code at bit.ly/yaydeer

Python Shell

```
>>>

Constance Crompton, Lee Zickel, and Emily C Murphy


[Please click for course details.]




[This offering is now full (11 January)]


For those new to the field, this is an introduction to the theory and practice o
f encoding electronic texts for the humanities.

  This workshop is designed for individuals who are contemplating

   embarking on a text-encoding project, or for those who would like

   to better understand the philosophy, theory, and practicalities

   of encoding in XML (Extensible Markup Language) using the Text

   Encoding Initiative (TEI) Guidelines. No prior experience with

   XML is assumed, but the course will move quickly through the

   basics.



This is a hands-on course.  Consider this offering in complement with, and / or
```

Next steps

Instead of printing to the screen, you might output the extracted text bits into a .txt file so you can play with it later.

Use BeautifulSoup and/or regexes to zero in on certain text on complex web pages (e.g., the first italicized word in each paragraph).