This is CSCI P-14300.

Web Scraping

what?

useful?

procedure

1: get raw html from url

2: process html

1: get raw html from url

requests

```
$ pip install requests
$ python
>>> import requests
>>> url = "https://cs50.github.io/summer/syllabus"
>>> print requests.get(url)
<Response [200]>
>>> print requests.get(url).text
<!DOCTYPE html>
...
```

2: process html

beautiful soup

```
$ pip install bs4
$ python
>>> from bs4 import BeautifulSoup
>>> url = "https://cs50.github.io/summer/syllabus"
>>> raw_html = requests.get(url).text
>>> soup = BeautifulSoup(raw_html)
```

tasks

find all links

```
>>> for link in soup.find_all(‘a’, href=True):
>>>     print(link.get('href'))
/summer
http://www.summer.harvard.edu/
...
```

get text from paragraphs

```
>>> for para in soup.find_all(‘p’, href=True):
>>>     print(para.get_text('href'))
Introduction to Web Programming

Harvard Summer School
...
```

let’s get started!

```
$ git clone git@github.com:daveyproctor/web-scraping-seminar.git
```

This is CSCI P-14300.

if git clone didn’t work:

```
$ wget https://summer50-1-1-dwp7.c9users.io/web-scraping-seminar.zip
```