Discussion 8

SI206 Discussion 8

Beautiful Soup

Beautiful Soup for scraping

To use the Beautiful Soup module for scraping, you need to create the Beautiful Soup object. There are 3 steps to it:

1. Create a variable that stores the url of website
2. Get the data from the url i.e. `r = requests.get(url)`
3. Create a soup object using the data i.e.
   `soup = BeautifulSoup(r.text, ‘html.parser’)`

Things to keep in mind with BeautifulSoup

1. soup.find(‘tag’) will return the first tag that matches

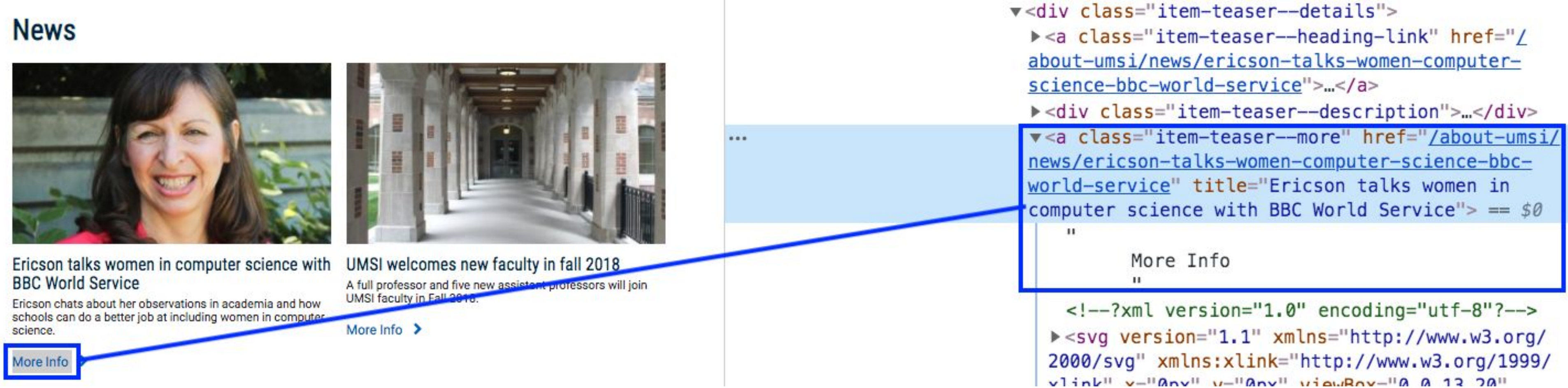
2. soup.find_all(‘tag’) will return a list of all the tags that match

3. You can use find and find_all on the tag objects to find children tags!

4. Use the tag_object.attrs to obtain a dictionary of the attributes in a tag object

5. Use the tag_object.get(attr_name) to get a specific attribute

Getting info from a single tag



News

Ericson talks women in computer science with BBC World Service

Ericson chats about her observations in academia and how schools can do a better job at including women in computer science.

More Info

UMSI welcomes new faculty in fall 2018

A full professor and five new assistant professors will join UMSI faculty in Fall 2018.

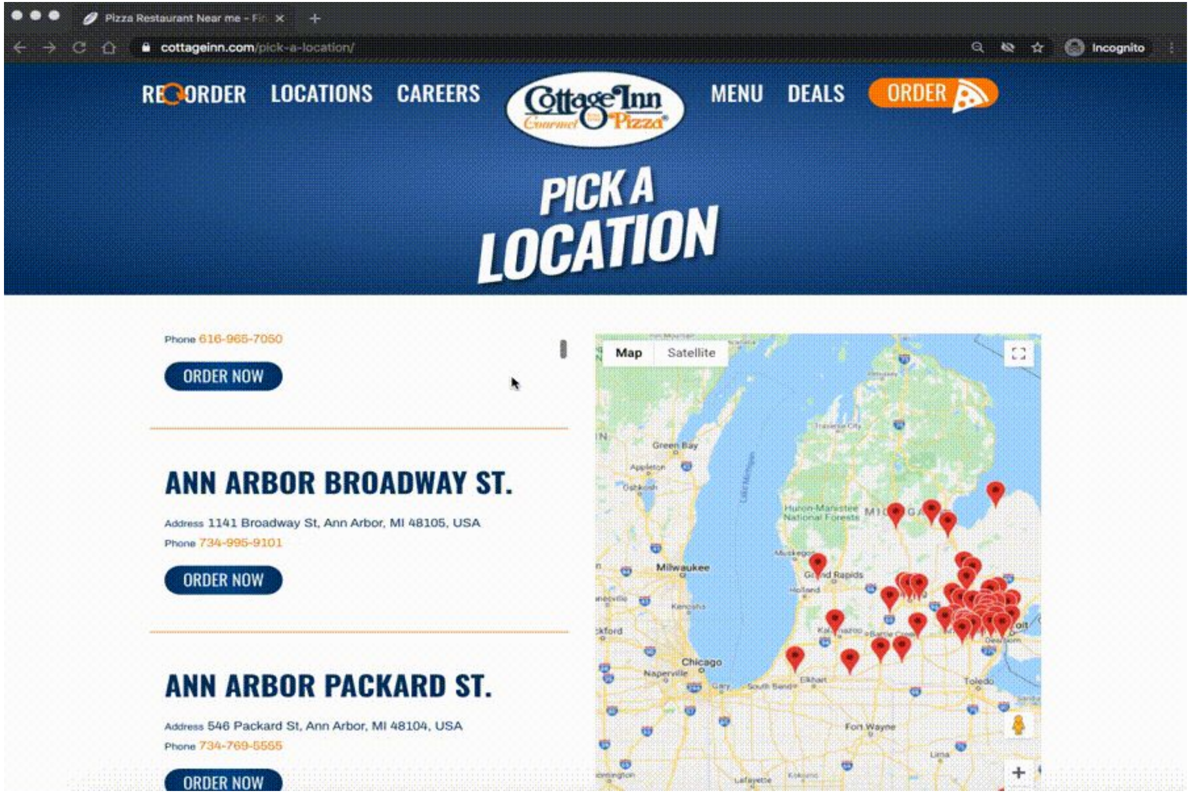
More Info ›

```
▼<div class="item-teaser--details">
  ▶<a class="item-teaser--heading-link" href="/
  about-umsi/news/ericson-talks-women-computer-
  science-bbc-world-service">…</a>
  ▶<div class="item-teaser--description">…</div>
...
  ▼<a class="item-teaser--more" href="/about-umsi/
  news/ericson-talks-women-computer-science-bbc-
  world-service" title="Ericson talks women in
  computer science with BBC World Service"> == $0
      "
          More Info
          "
      <!--?xml version="1.0" encoding="utf-8"?-->
    ▶<svg version="1.1" xmlns="http://www.w3.org/
    2000/svg" xmlns:xlink="http://www.w3.org/1999/
    xlink" x="0px" y="0px" viewBox="0 0 13 20"
```

Since that tag is the first of its type on the page, we can use the soup.find()

```
# Get first tag of a certain type from the soup
tag = soup.find('a', class_='item-teaser--more')
# Get info from tag
info = tag.get('href')
```

Getting info from all tags of a certain type



Phone 616-965-7050

ORDER NOW

ANN ARBOR BROADWAY ST.

Address 1141 Broadway St, Ann Arbor, MI 48105, USA

Phone 734-995-9101

ORDER NOW

ANN ARBOR PACKARD ST.

Address 546 Packard St, Ann Arbor, MI 48104, USA

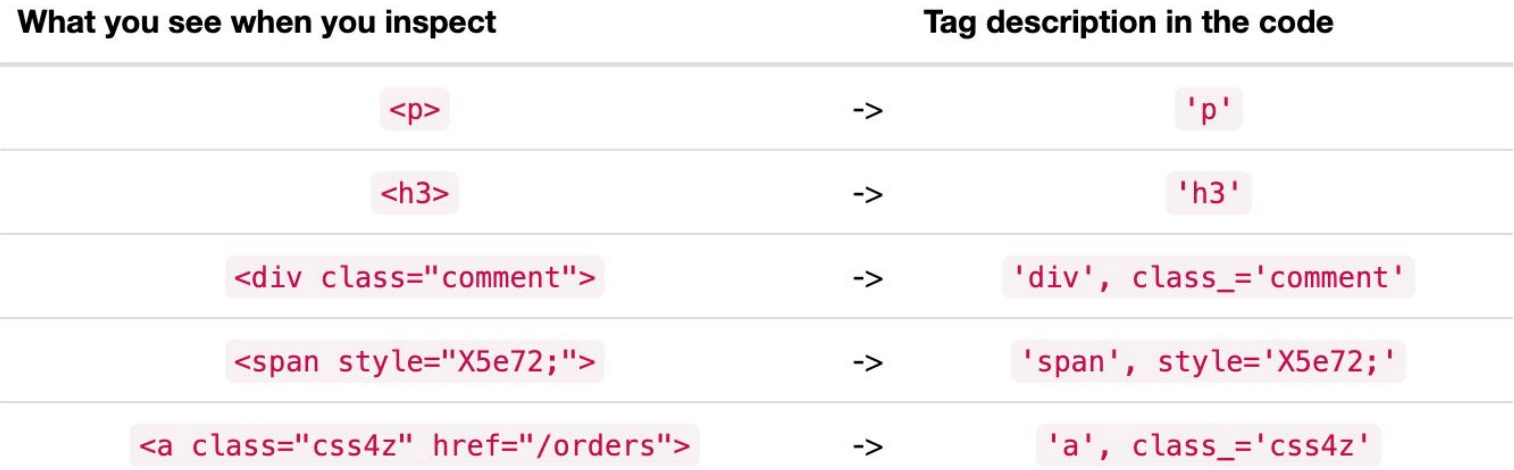
Phone 734-769-5555

ORDER NOW

We see that we need to get info from all the `h3` tags from the webpage. The *text* in those tags has the information we need!

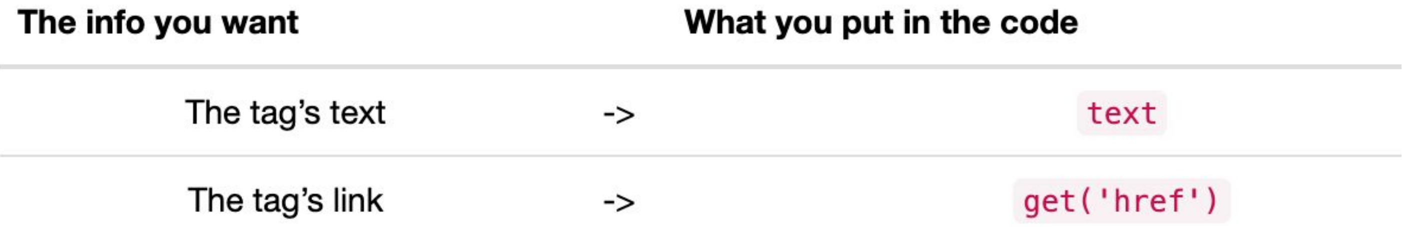
```
# Get all tags of a certain type from the soup
tags = soup.find_all('h3')
# Collect info from the tags
collect_info = []
for tag in tags:
    # Get info from tag
    info = tag.text
collect_info.append(info)
```

1. Find the tag description and use that as an argument in soup.find() or soup.find_all()



What you see when you inspect		Tag description in the code
`<p>`	->	`'p'`
`<h3>`	->	`'h3'`
`<div class="comment">`	->	`'div', class_='comment'`
`<span style="X5e72;">`	->	`'span', style='X5e72;'`
`<a class="css4z" href="/orders">`	->	`'a', class_='css4z'`

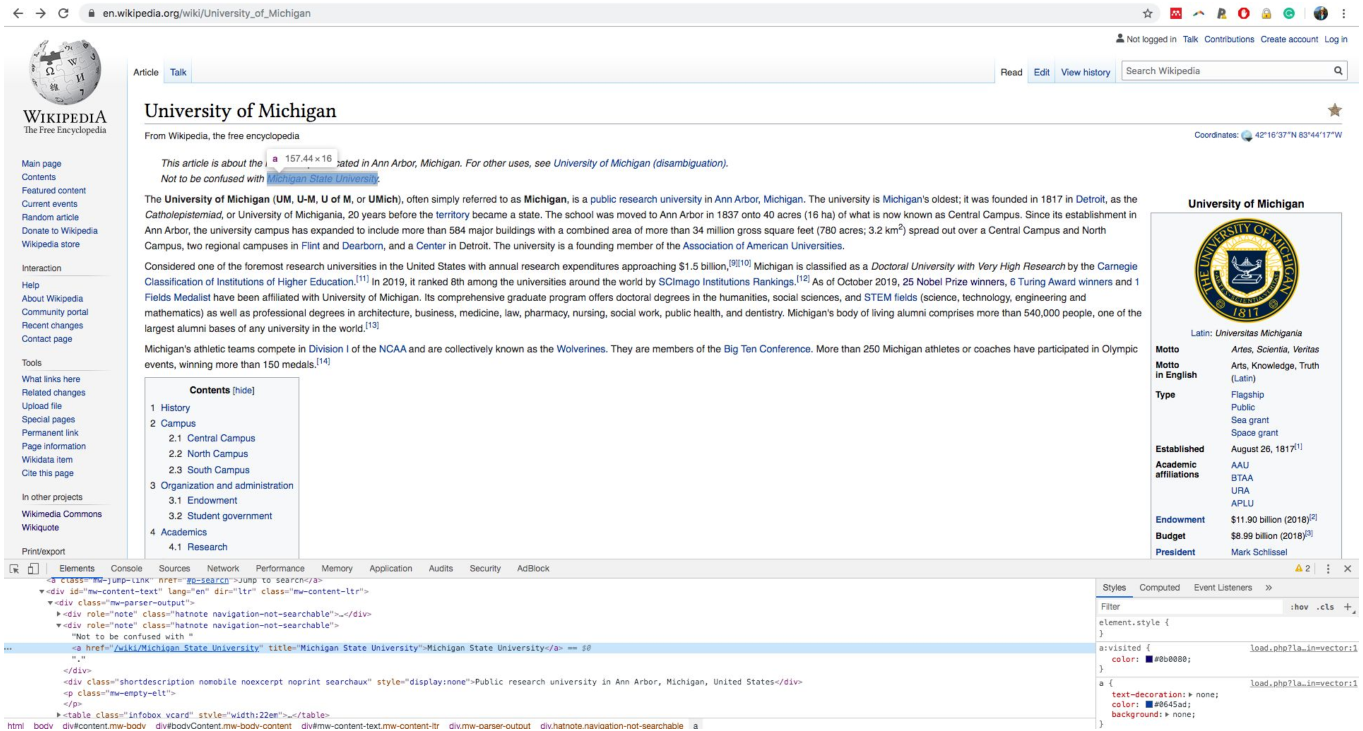
2. Determine if you want to get text from a tag, or a link from a tag



The info you want		What you put in the code
The tag's text	->	`text`
The tag's link	->	`get('href')`

Use Dev Tools!

Right click on an element you want to know more about and choose ‘Inspect’.



University of Michigan

From Wikipedia, the free encyclopedia

Coordinates: 42°16′37″N 83°44′17″W

This article is about the … ated in Ann Arbor, Michigan. For other uses, see University of Michigan (disambiguation).

Not to be confused with Michigan State University.

The **University of Michigan** (**UM**, **U-M**, **U of M**, or **UMich**), often simply referred to as **Michigan**, is a public research university in Ann Arbor, Michigan. The university is Michigan's oldest; it was founded in 1817 in Detroit, as the *Catholepistemiad*, or University of Michigania, 20 years before the territory became a state. The school was moved to Ann Arbor in 1837 onto 40 acres (16 ha) of what is now known as Central Campus. Since its establishment in Ann Arbor, the university campus has expanded to include more than 584 major buildings with a combined area of more than 34 million gross square feet (780 acres; 3.2 km^2) spread out over a Central Campus and North Campus, two regional campuses in Flint and Dearborn, and a Center in Detroit. The university is a founding member of the Association of American Universities.

Considered one of the foremost research universities in the United States with annual research expenditures approaching $1.5 billion,[9][10] Michigan is classified as a *Doctoral University with Very High Research* by the Carnegie Classification of Institutions of Higher Education.[11] In 2019, it ranked 8th among the universities around the world by SCImago Institutions Rankings.[12] As of October 2019, 25 Nobel Prize winners, 6 Turing Award winners and 1 Fields Medalist have been affiliated with University of Michigan. Its comprehensive graduate program offers doctoral degrees in the humanities, social sciences, and STEM fields (science, technology, engineering and mathematics) as well as professional degrees in architecture, business, medicine, law, pharmacy, nursing, social work, public health, and dentistry. Michigan's body of living alumni comprises more than 540,000 people, one of the largest alumni bases of any university in the world.[13]

Michigan's athletic teams compete in Division I of the NCAA and are collectively known as the Wolverines. They are members of the Big Ten Conference. More than 250 Michigan athletes or coaches have participated in Olympic events, winning more than 150 medals.[14]

Contents [hide]

1 History
2 Campus
2.1 Central Campus
2.2 North Campus
2.3 South Campus
3 Organization and administration
3.1 Endowment
3.2 Student government
4 Academics
4.1 Research

University of Michigan

Latin: *Universitas Michigania*

Motto	*Artes, Scientia, Veritas*
Motto in English	Arts, Knowledge, Truth (Latin)
Type	Flagship Public Sea grant Space grant
Established	August 26, 1817[1]
Academic affiliations	AAU BTAA URA APLU
Endowment	$11.90 billion (2018)[2]
Budget	$8.99 billion (2018)[3]
President	Mark Schlissel

```
    <a class="mw-jump-link" href="#p-search">Jump to search</a>
  ▼<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
    ▼<div class="mw-parser-output">
      ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
      ▼<div role="note" class="hatnote navigation-not-searchable">
          "Not to be confused with "
...       <a href="/wiki/Michigan_State_University" title="Michigan State University">Michigan State University</a> == $0
          "."
        </div>
        <div class="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">Public research university in Ann Arbor, Michigan, United States</div>
        <p class="mw-empty-elt">
        </p>
      ▶<table class="infobox vcard" style="width:22em">…</table>
```

```
element.style {
}
a:visited {                    load.php?la…in=vector:1
   color: ■#0b0080;
}
a {                            load.php?la…in=vector:1
   text-decoration: ▶ none;
   color: ■#0645ad;
   background: ▶ none;
}
```

Scraping Wikipedia

We will use BeautifulSoup to get some data from https://en.wikipedia.org/wiki/University_of_Michigan

Task 1: Create a BeautifulSoup object

Task 2: Get the URL that links to list of Nobel Prize Winners. The clickable link can be found in the second paragraph on the page.

HINT: You will have to add https://en.wikipedia.org to the URL retrieved using BeautifulSoup

Scraping Wikipedia

Task 3: Get the details from the box titled "College/school founding". Get all the college/school names and the year they were founded and organize the same into key-value pairs.

Organize the details into a dictionary as shown below:

```
{'College of Literature, Science, and the Arts': '1841',
 'School of Medicine': '1850',
 .
 .
 'School of Kinesiology': '1984'}
```

APPENDIX - Tips

1. We can filter tags by their attributes by passing additional arguments to the *find()* or *find_all()* methods. For instance, if I only want to get *a* tags that link to Google, I could do:
    a. `soup.find_all('a', href='https://www.google.com')`
2. Remember that you need to use *class_* instead of *class* in *find* or *find_all* because class is a reserved word in Python.
3. When trying to decide how you want to grab a particular tag, remember that in HTML a *class* is typically assigned to multiple tags while an *id* is unique.
    a. Sometimes a tag may have multiple classes separated by a space. Do not treat these all as one class.