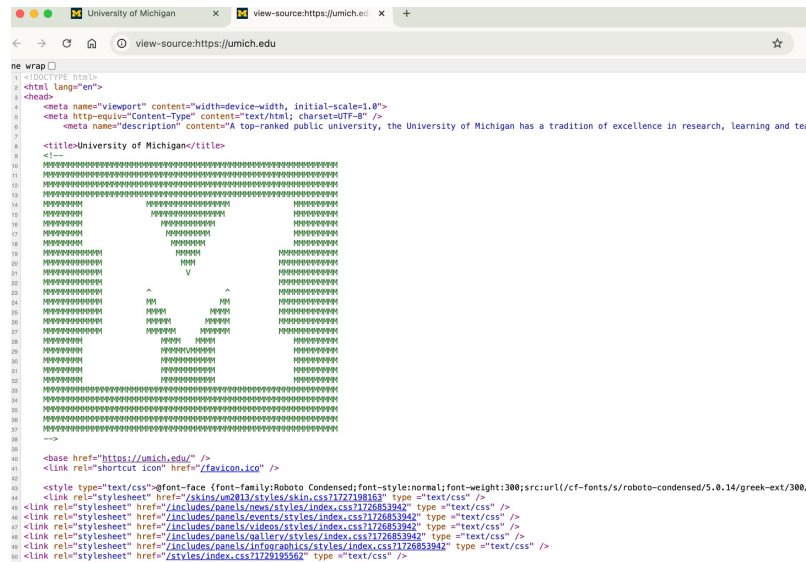


Discussion 10: More on Beautiful Soup

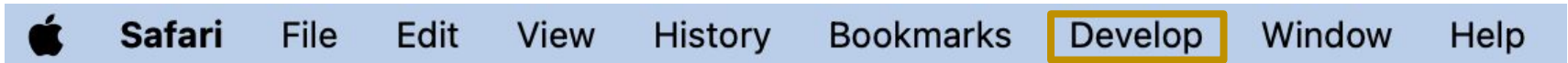
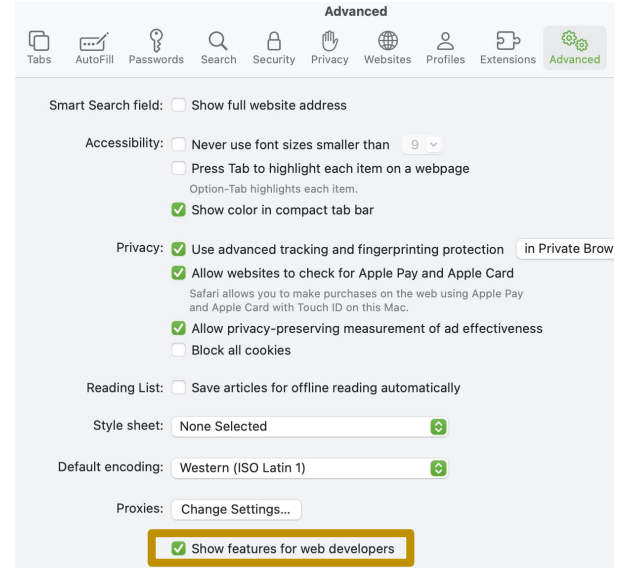
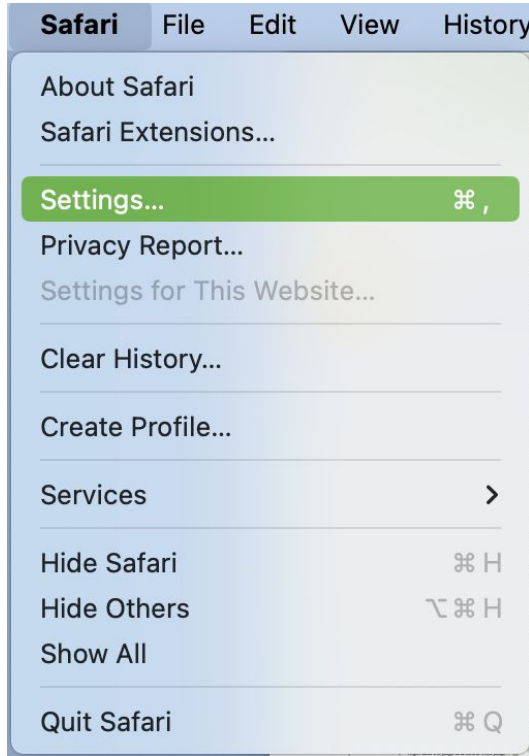
View page's source code in Google Chrome

- Open the page's source code view:
 - **Windows:**
 - i. Right-click the page and select View page source (or View source).
 - ii. Alternatively, you can press Ctrl + U on your keyboard.
 - **Mac:**
 - i. Right-click the page and select View page source (or View source).
 - ii. Press Cmd + Option + U on your keyboard.
- Open the developer tools:
 - Windows: Press the F12 key on your keyboard.
 - Mac: Press Cmd + Option + I on your keyboard.



View page's source code in Safari

- Click *Safari* in the top menu.
 - Select *Settings*.
 - Select *Advanced*.
 - Select the *Show features for web developers* checkbox.
-
- Open the relevant page of your site in Safari.
 - Click *Develop* in the top menu.
 - Select *Show Page Source*



Find a tag, then find all of a tag type in that tag

```
<html>
<body>
  <div id="main-content">
    <p>This is a paragraph.</p>
    <a href="https://example.com/page1">Link 1</a>
    <a href="https://example.com/page2">Link 2</a>
  </div>
  <div id="footer">
    <a href="https://example.com/footer-link">Footer Link</a>
  </div>
</body>
</html>
```

Step 2

Step 3

Step 1: Parse the HTML

```
BeautifulSoup(html_content,
               'html.parser')
```

Step 2: Find the specific <div> by id













```
soup.find('div',
          id='main-content')
```

Step 3: Find all <a> tags within this <div> links

```
.find_all('a')
```

Step 4: Fill it into a dictionary

You'll be working with a list of National Historic Landmarks in Michigan on Wikipedia

[5] ⚙	Landmark name ⚙	Image ⚙	Date designated ^[6] ⚙	Location ⚙	County ⚙	Description ⚙
1†	Bay View	  More images	December 23, 1987 (#72000613 )	Bear Creek  45°23′08″N 84°55′49″W	Emmet	Established in 1876 as a Methodist camp meeting , this romantically -planned campground was converted to an independent chautauqua in 1885, a role it served until 1915. These two uniquely American community forms are exemplified in this extensive and well-preserved complex. ^[7]
2†	Calumet Historic District	  More images	March 28, 1989 (#89001097 )	Calumet  47°17′45″N 88°27′14″W	Houghton	Covering the industrial, commercial and residential districts of the Calumet and Hecla Mining Company operating area, Calumet focuses on the influence, innovations and longevity of the Michigan copper industry . ^[8]
3#	<i>City of Milwaukee</i> (Great Lakes Car Ferry)	  More images	December 14, 1990 (#90002221 )	Manistee  44°15′34″N 86°18′58″W	Manistee	Between 1931 and 1982, the <i>City of Milwaukee</i> served as a car ferry across Lake Michigan. She is the only pre-1940 Great Lakes car ferry still in existence. ^[9]

Task 1:

Create a **beautifulsoup object** with data from the Wikipedia page

Task 1 is under the main() function

```
def main():  
    #TASK 1: GET DATA FROM WIKIPEDIA  
    url = 'https://en.wikipedia.org/wiki/List\_of\_National\_Historic\_Landmarks\_in\_Michigan'
```

Task 2:

Create a **nested dictionary** of information about the historic landmarks

- The keys in the outer dictionary will be the landmark names
- The keys in the inner dictionary will be all of the columns in the table besides the image column
 - You'll have the following keys: date designated, location, county, description

A correctly formatted **dictionary** for the Bay View National Historic Landmark should look like this:

```
{'Bay View': {'date designated': 'December 23, 1987(#72000613)', 'location': 'Bear Creek 45°23'08"N 84°55'49"W\uffeff / \uffeff45.3855555555555555°N 84.930277777777779°W\uffeff / 45.3855555555555555; -84.930277777777779\uffeff (Bay View)', 'county': 'Emmet', 'description': 'Established in 1876 as a Methodist camp meeting, this romantically-planned campground was converted to an independent chautauqua in 1885, a role it served until 1915. These two uniquely American community forms are exemplified in this extensive and well-preserved complex.[7]'}, 'Calumet Historic District': {'date designated': 'March 28, 1989(#89001097)', 'location': 'Calumet 47°17'45"N 88°27'14"W\uffeff / \uffeff47.295833°N 88.453889°W\uffeff / 47.295833; -88.453889\uffeff (Calumet Historic District)', 'county': 'Houghton', 'description': 'Covering the industrial, commercial and residential districts of the Calumet and Hecla Mining Company operating area, Calumet focuses on the influence, innovations and longevity of the Michigan copper industry.[8]'}, 'City of Milwaukee (Great Lakes Car Ferry)': {'date designated': 'December 14, 1990(#90002221)', 'location': 'Manistee 44°15'34"N 86°18'58"W\uffeff / \uffeff44.259324°N 86.316018°W\uffeff / 44.259324; -86.316018\uffeff (City of Milwaukee (Great Lakes Car Ferry))', 'county': 'Manistee', 'description': 'Between 1931 and 1982, the City of Milwaukee served as a car ferry across Lake Michigan. She is the only pre-1940 Great Lakes car ferry still in existence.[9]'}, 'Cranbrook': {'date designated': 'June 29, 1989(#73000954)', 'location': 'Bloomfield Hills 42°34'23"N 83°14'57"W\uffeff / \uffeff42.573055555555555°N 83.249166666666667°W\uffeff / 42.573055555555555; -83.249166666666667\uffeff (Cranbrook)
```

```
{'date designated': 'December 23, 1987(#72000613)', 'location': 'Bear Creek 45°23'08"N 84°55'49"W\uffeff / \uffeff45.385555555555555°N 84.930277777777779°W\uffeff / 45.3855555555555555; -84.930277777777779\uffeff (Bay View)', 'county': 'Emmet', 'description': 'Established in 1876 as a Methodist camp meeting, this romantically-planned campground was converted to an independent chautauqua in 1885, a role it served until 1915. These two uniquely American community forms are exemplified in this extensive and well-preserved complex.[7]}'
```

Task 3:

Use regular expressions to find proper noun phrases in the description

- A proper noun phrase is just multiple consecutive words that start with capital letters
- e.g. “Lake Michigan” or “Michigan State University” but not “University of Michigan,” because “of” is not capitalized

This is a difficult regex. We highly recommend using **regex101** to help you

Task 3:

This sample output shows

- 1) the description for the Grand Hotel
- 2) correct output from `get_proper_noun_phrases`

```
Built in the late 19th century, this white clapboard structure is one of the few extant large  
wood-framed hotels of the era. Situated on a bluff overlooking Lake Huron, it has been called "the  
American dream of "a summer place.""[24]  
['Lake Huron']
```