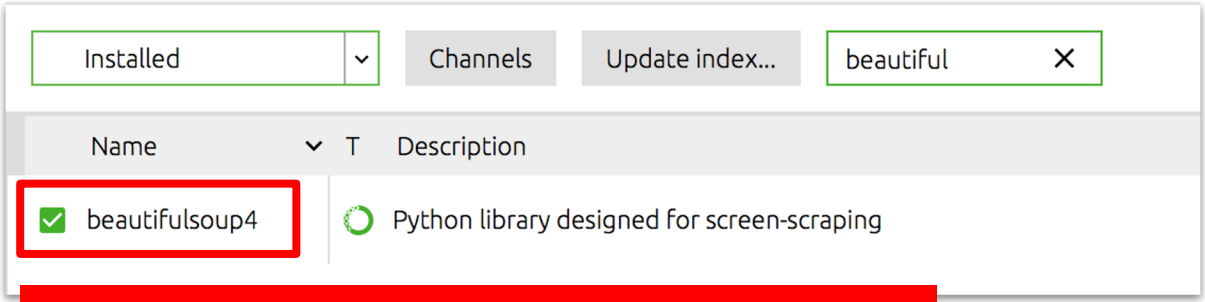
Web Scraping with Python

Setup prerequisite

1. Anaconda (Python 3)

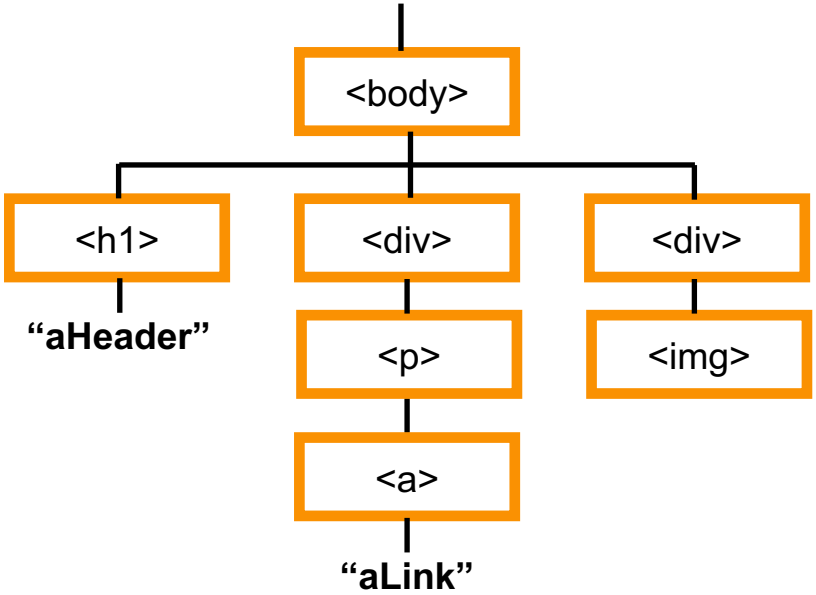
1. Jupyter Notebook or Jupyter Lab

1. BeautifulSoup Library
   -> Run from bs4 import BeautifulSoup



ตรวจสอบว่าได้ติดตั้ง **BeautifulSoup4** เรียบร้อยแล้ว

BeautifulSoup Primer

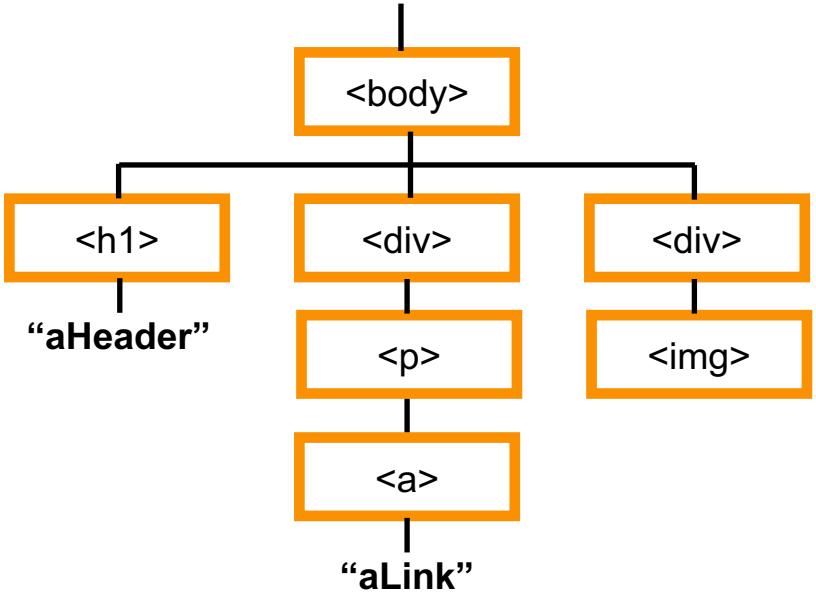



```
<body>
        <h1>aHeader</h1>
        <div class=”section1”>
                <p>
                        <a href=”#”>aLink</a>
                </p>
        </div>
        <div class=”section2”>
                <img src=”img.jpg/>
        </div>
</body>
```

Workshop 2.1 : BeautifulSoup

01-basic_beautifulsoup.ipynb

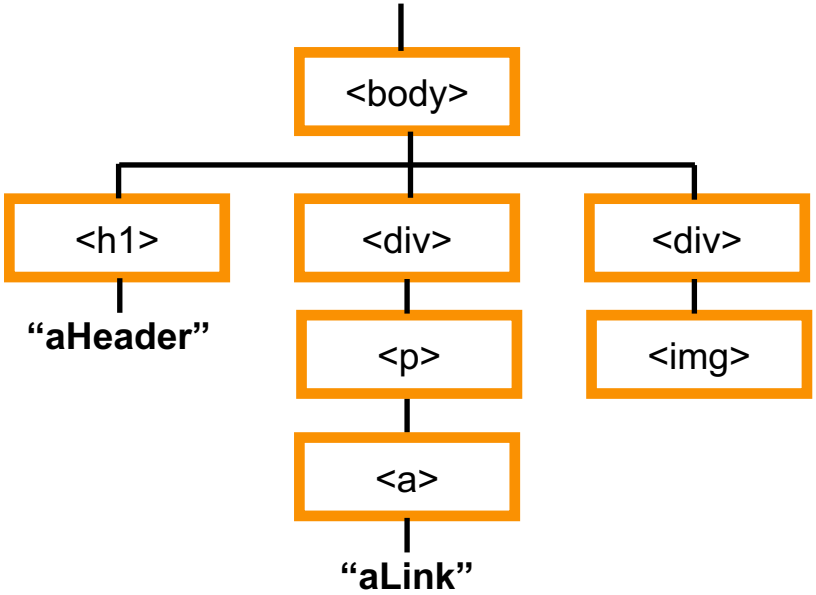
BeautifulSoup Primer : Find the target element




```
s = BeautifulSoup(html,’html.parser’)

s.body
```

BeautifulSoup Primer : Find the target element

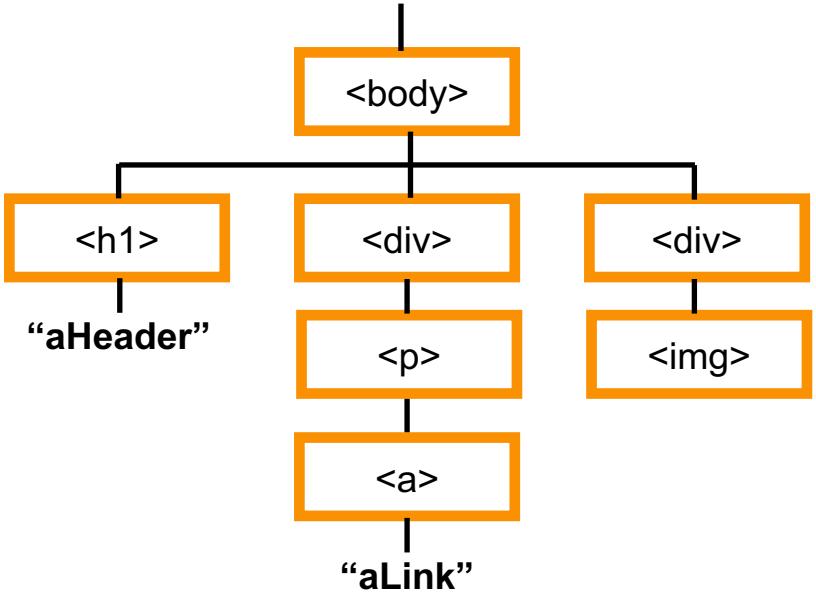


s.h1

or

s.find('h1')

BeautifulSoup Primer : Find the target element

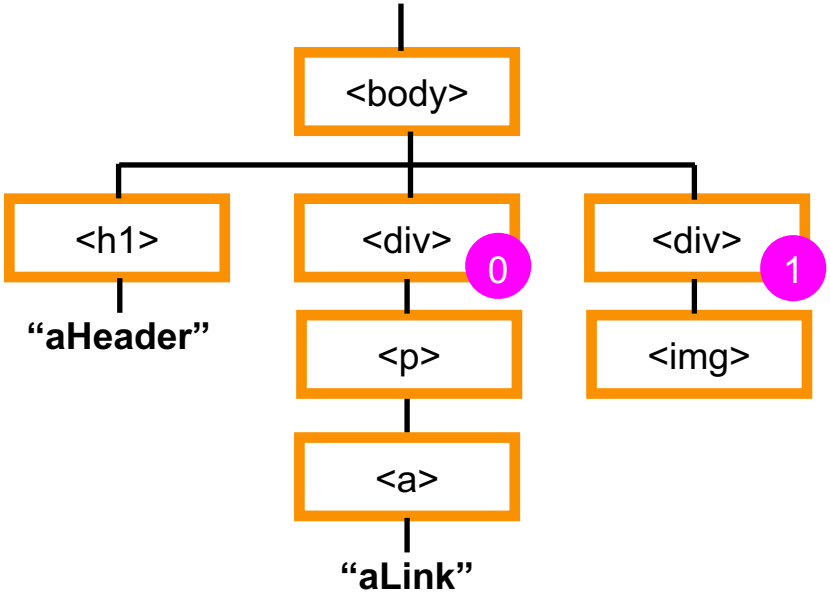


s.div

or

s.find(‘div’)

BeautifulSoup Primer : Find the target element

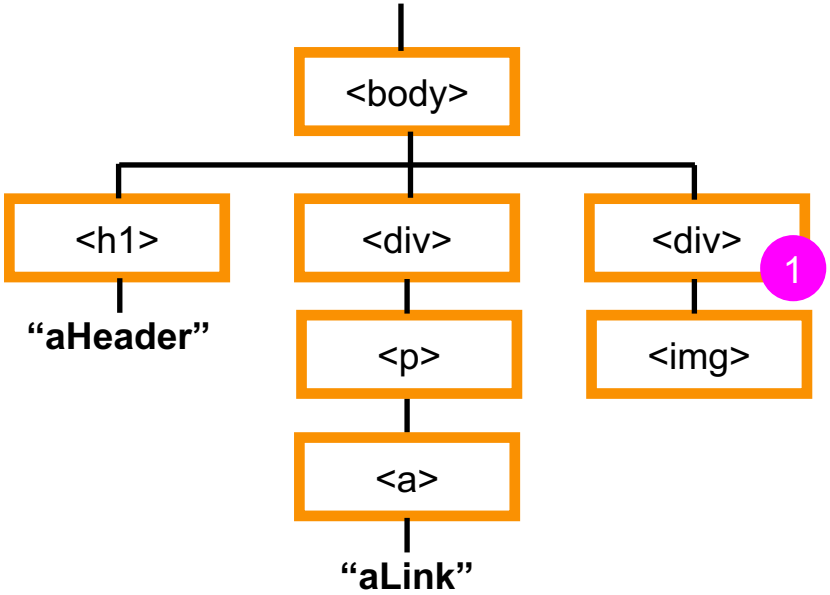


s.('div')

or

s.find_all('div')

BeautifulSoup Primer : Find the target element

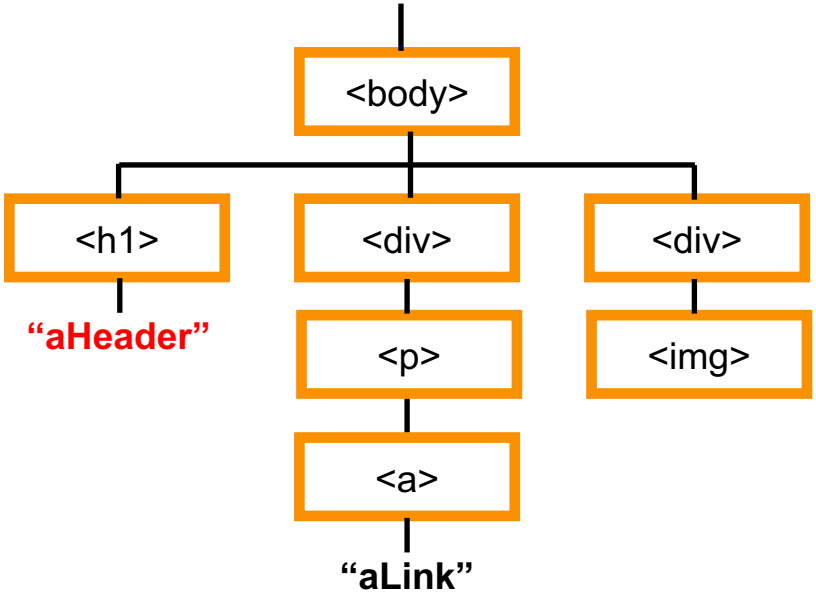


s.('div')[1]

or

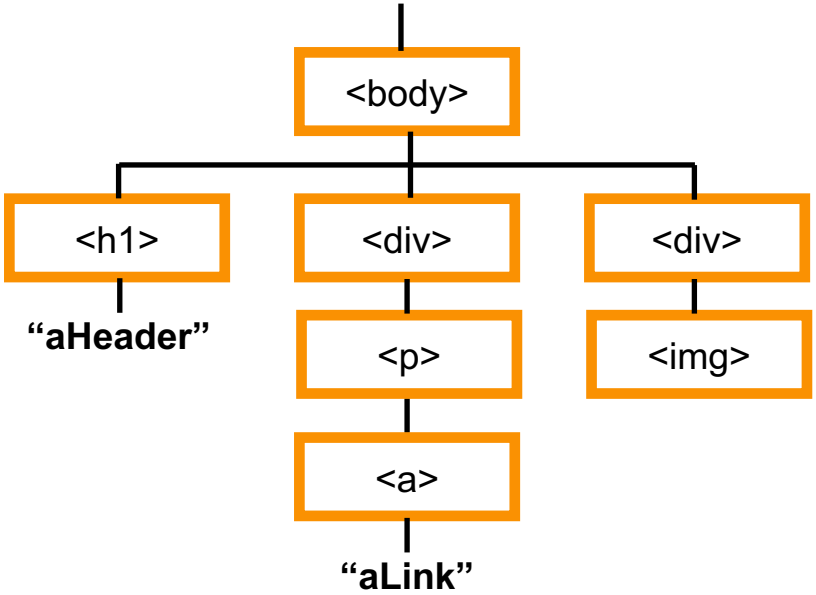
s.find_all('div')[1]

BeautifulSoup Primer : Find the target element



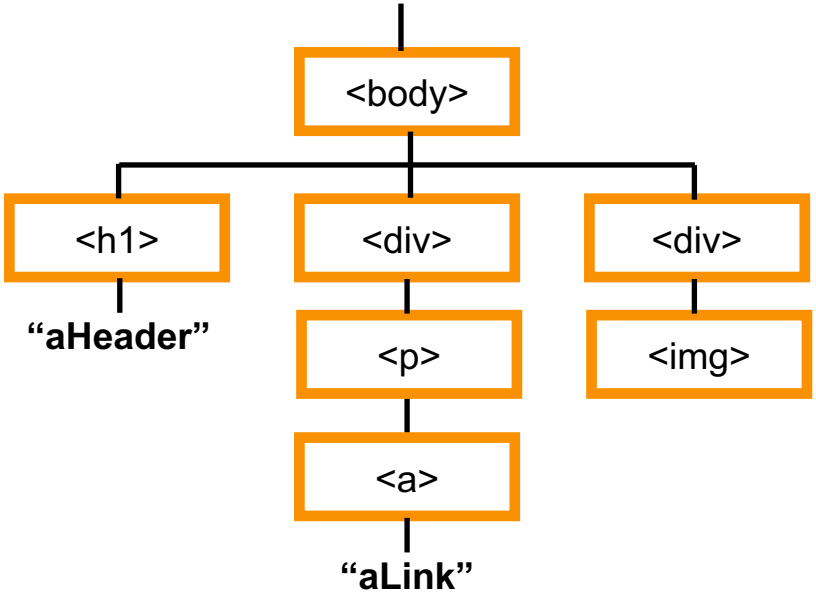

```
s.find(text='aHeader')
```

BeautifulSoup Primer : Find the target element




```
s.find(‘h1’ , string=’aHeader’)
```

BeautifulSoup Primer : Find the target element

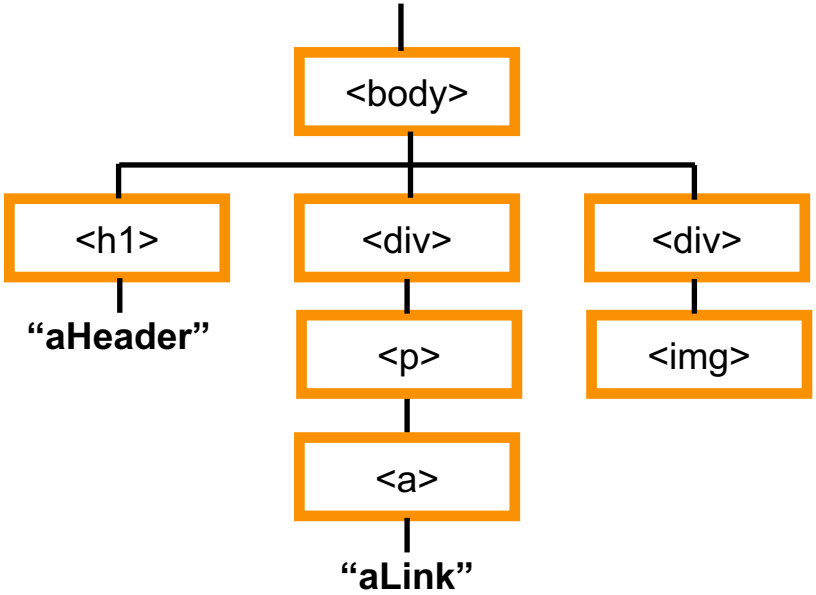



```
s.find( attrs={ 'class' : 'section1' } )
```

or

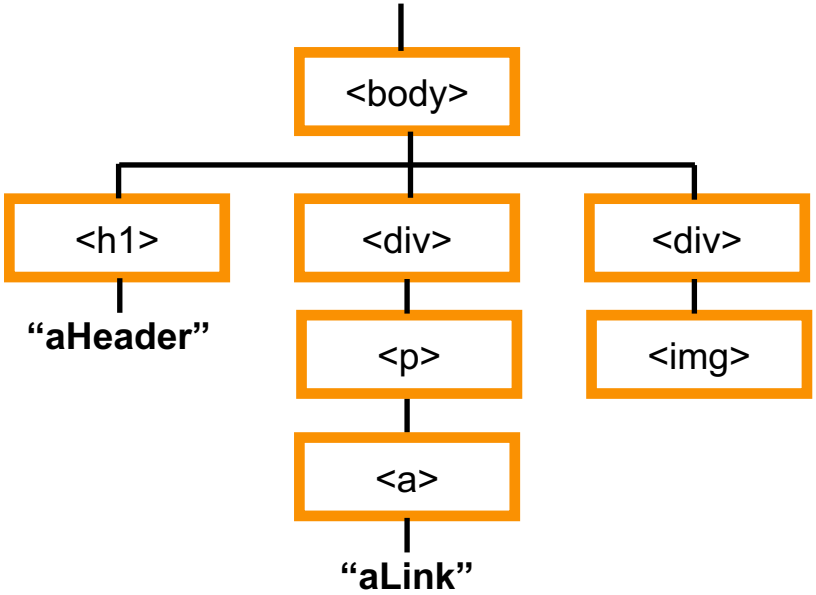
```
s.find(class='section1')
```

BeautifulSoup Primer : Traverse the DOM tree



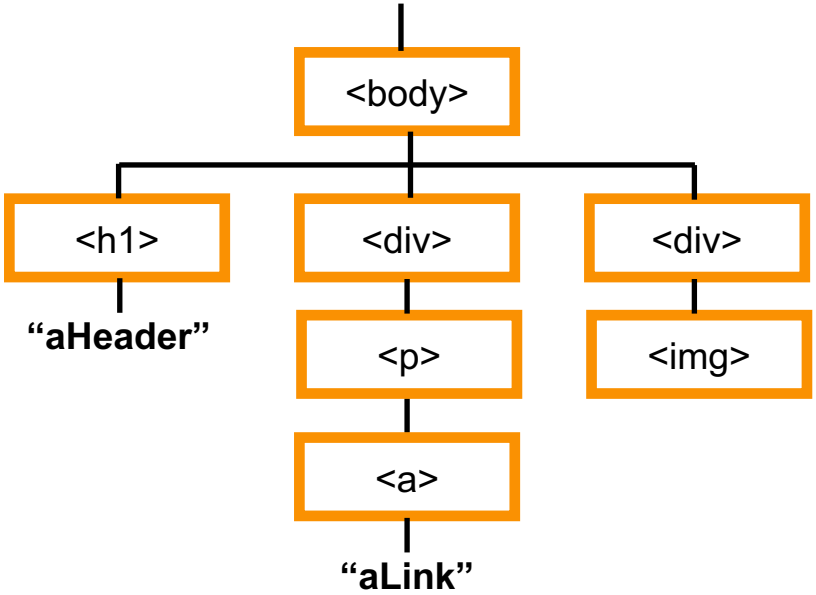
s.div.parent

BeautifulSoup Primer : Traverse the DOM tree



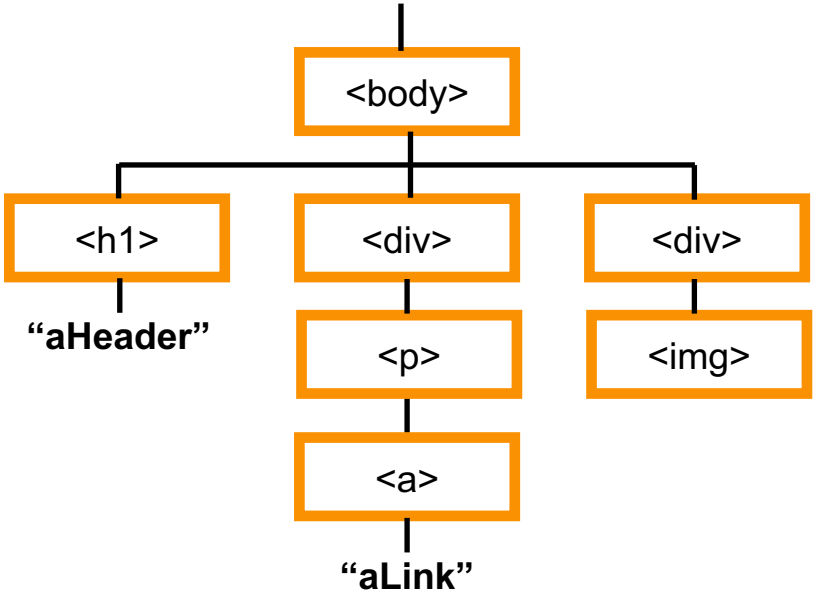
s.div.previous_sibling

BeautifulSoup Primer : Traverse the DOM tree



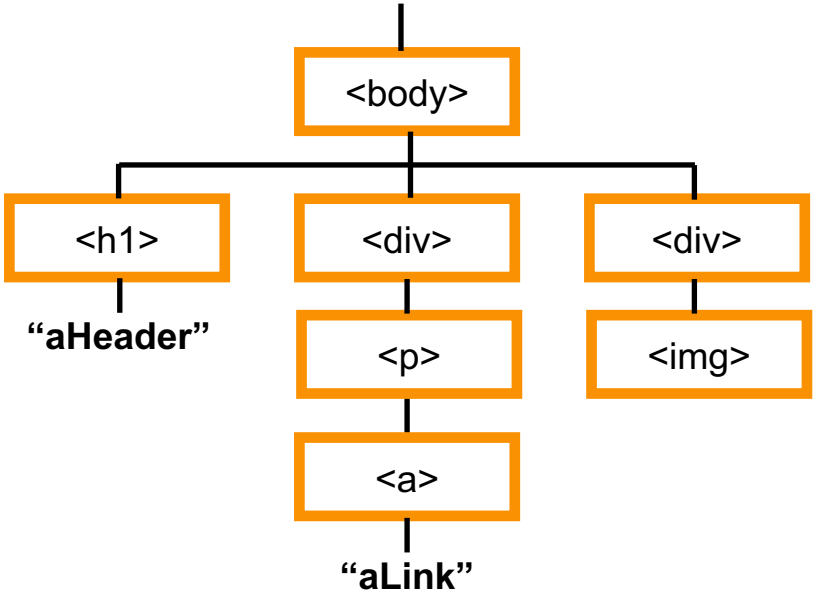
s.div.next_sibling

BeautifulSoup Primer : Traverse the DOM tree



s.div.next_element

BeautifulSoup Primer : Traverse the DOM tree

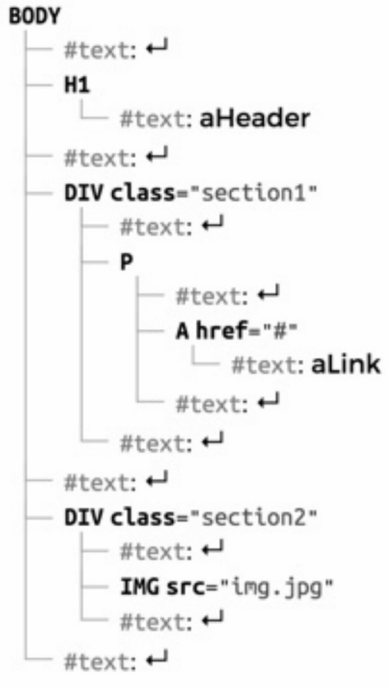


s.div.parent

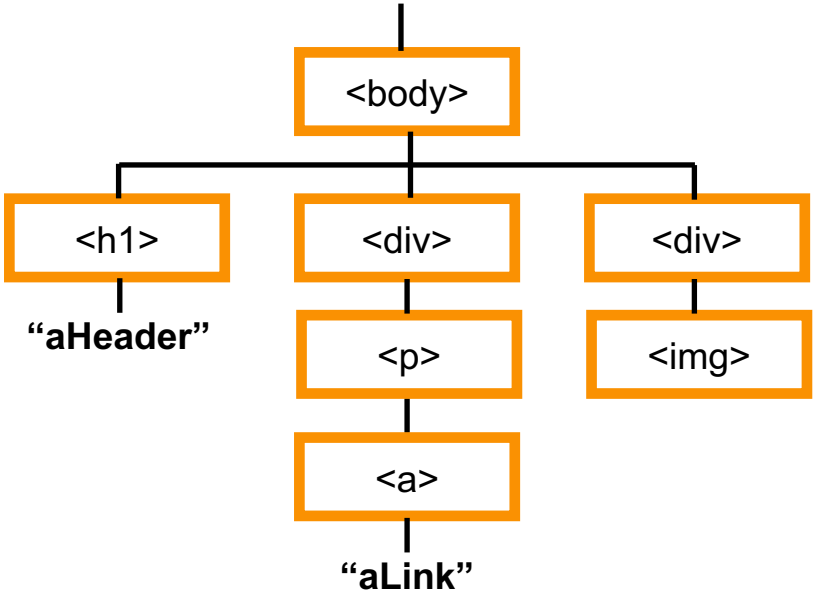
Warning : for white spaces !!!

จะเกิด ช่องว่าง **(space)** และ ขึ้นบรรทัดใหม่ **(new lines)** ระหว่าง **tag**

```
<body>
        <h1>aHeader</h1>
        <div class=”section1”>
                <p>
                        <a href=”#”>aLink</a>
                </p>
        </div>
        <div class=”section2”>
                <img src=”img.jpg/>
        </div>
</body>
```

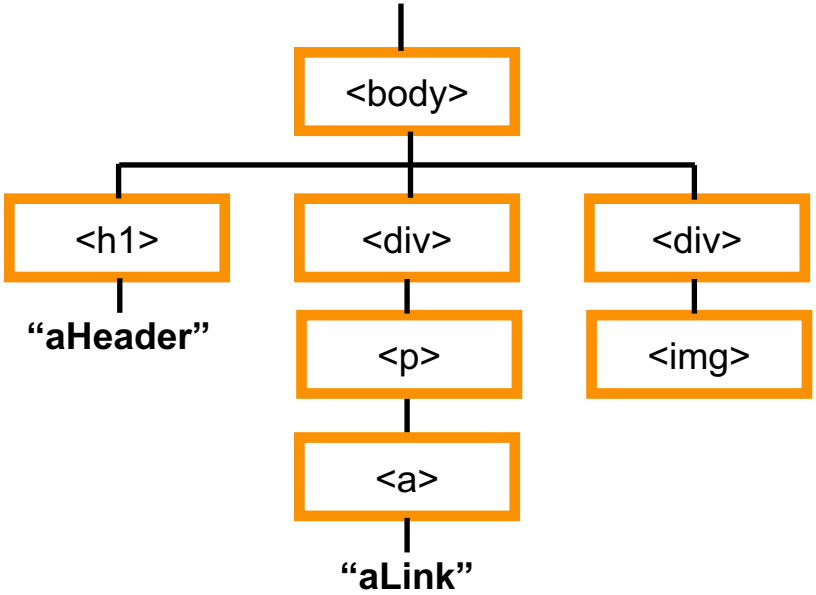


BeautifulSoup Primer : Traverse the DOM tree



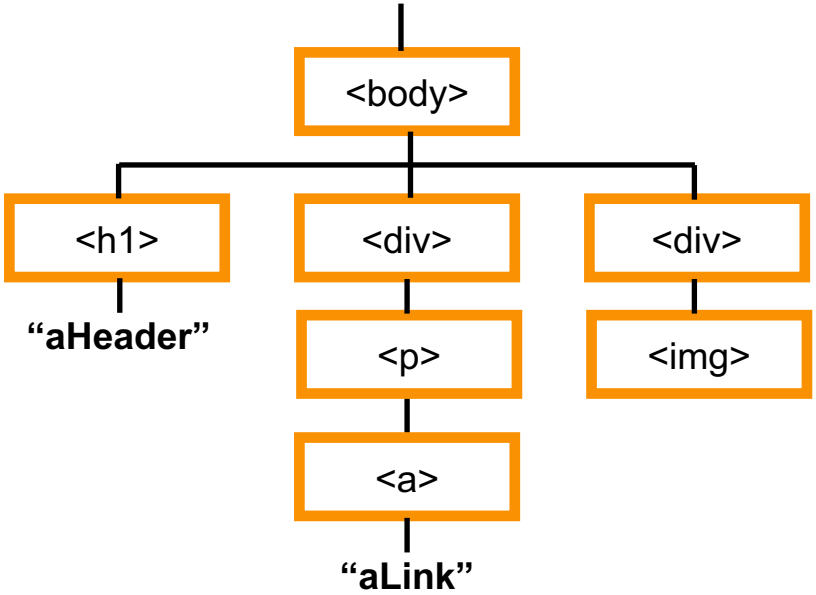

```
s.a.find_next()
```

BeautifulSoup Primer : Traverse the DOM tree



s.a.find_next(‘img’)

BeautifulSoup Primer : Traverse the DOM tree




```
s.find('div', class_='section2') \
    .find_previous_sibling('h1')
```

Workshop 2.2 : Data Scraping

02-web_scraping.ipynb

- ดึงข้อมูลรายชื่อบริษัทในเครือปัจจุบันในสังกัด GDH
- ดึงข้อมูลรายชื่อผู้กำกับภาพยนตร์ในสังกัด GDH
- ดึงข้อมูลรายชื่อนักแสดงในสังกัดนาดาวบางกอก
- ดึงข้อมูลรายชื่อภาพยนตร์ในเครือ GDH พร้อมทั้ง วันเปิดตัว , รายได้ และ ผู้กำกับ