**Web Scraping**

* Scrap the HTML document
  + static website
    - [requests](https://requests.readthedocs.io/en/latest/) (recommended)
    - [urllib.request](https://docs.python.org/3/library/urllib.request.html#module-urllib.request)
  + dynamic website
    - [webdr iver](https://selenium-python.readthedocs.io/getting-started.html)
* HTML document
  + HTML is a tree
  + tags 标签
  + attributes 属性
  + [HTML Examples](https://www.w3schools.com/html/html_examples.asp)
* Searching the tree
  + Intro to [Regular Expression (RegEx, 正则表达式)](https://www.w3schools.com/python/python_regex.asp)
    - anchors 锚点 (^, $)
    - alternations 交替 ((a|b), [ab])
    - metacharacters 元字符 (\w, \d)
    - repetitions and quantifiers 重复 ({}, \*)
    - looking around 环视 ((?<=), (?=))
  + [BeautifulSoup](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)
    - soup.tag
    - soup.find\_all("tag")
    - soup.find\_all("tag",{"attribute":"name"})
* Iteration 迭代
  + for loop 对…循环
  + list comprehension 列表推导式
    - [output\_function(item) for item in item\_collection]
    - [output\_function(item) for item in item\_collection if item\_judgement]
    - [output\_function(i) for item in item\_collection for i in item]
  + [review 1D object types](https://www.geeksforgeeks.org/differences-and-applications-of-list-tuple-set-and-dictionary-in-python/#:~:text=List%2C%20Tuple%2C%20Set%2C%20and,data%20in%20an%20efficient%20manner.&text=allow%20duplicate%20elements-,Set%20will%20not%20allow%20duplicate%20elements,doesn't%20allow%20duplicate%20keys.&text=Tuple%20can%20be%20created%20using%20tuple()%20function.&text=Dictionary%20can%20be%20created%20using%20dict()%20function.)
  + [range()](https://www.w3schools.com/python/python_regex.asp)
* Store results
  + create a [pandas DataFrame](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html)