Web Scraping

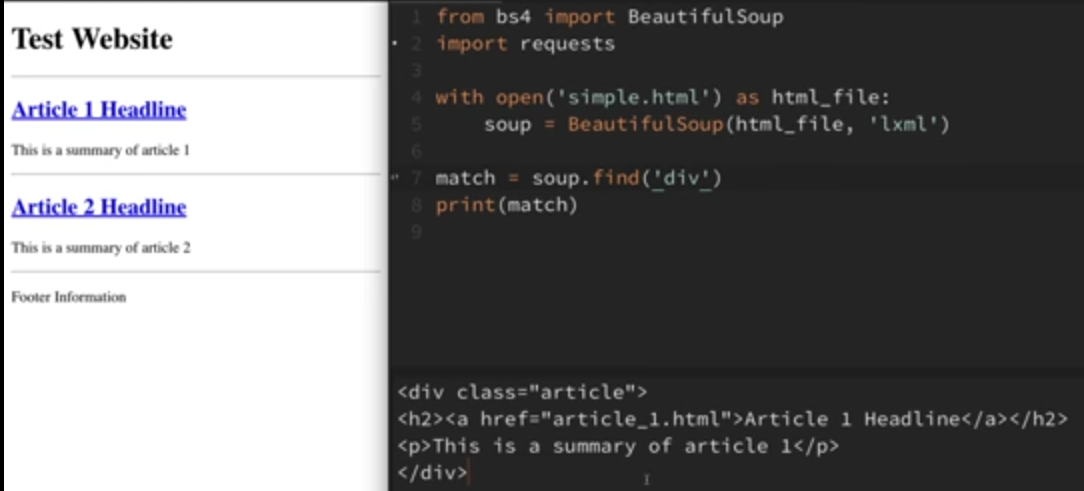
* A **parser** is a compiler or interpreter component that breaks data into smaller elements for easy translation into another language.
* A **parser** takes input in the form of a sequence of tokens or program instructions and usually builds a data structure in the form of a parse tree or an abstract syntax tree.
* Install beautifulsoup4 to have latest version of beautiful soup library. And then install particular parser eg: install lxml
* To open a website that is present in our local machine then steps are as follows:

1. Import beautifulsoup from bs4
2. Import requests used to access the websites.
3. Open the file.html from where u wanna return the data. That can be done as : ‘with open(‘filename.html’) as html\_file:’ here html\_file is the variable through which we’ll be able to access the data present in the file. Inside this create beautiful soup object with parameters html\_file and lxml parser declaration as shown below:

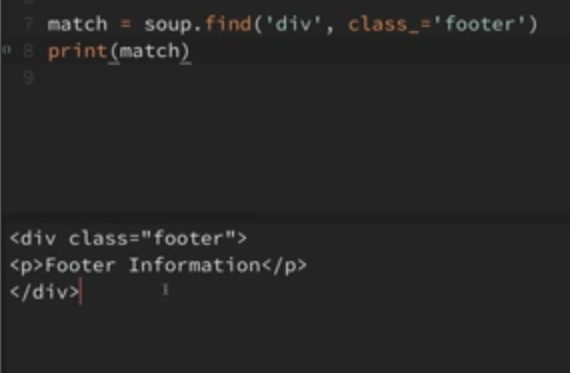
with open(‘file.html’) as html\_file:

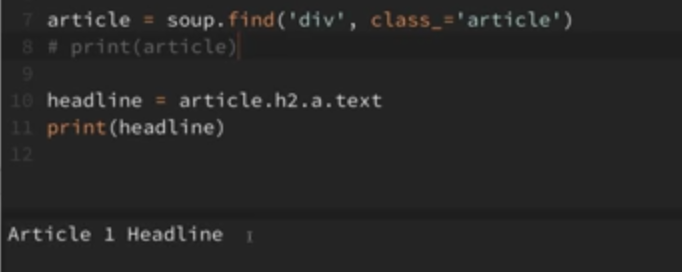
soup = BeautifulSoup(html\_file, ‘lxml’)

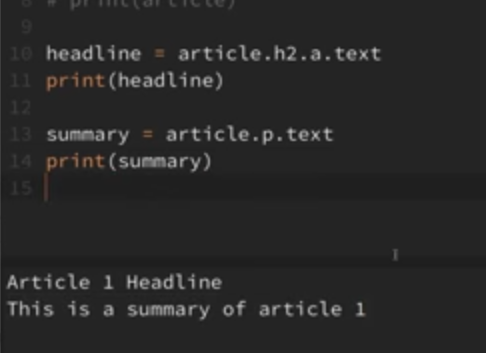
1. Then print(soup) it will print the whole html code of that page.
2. We can also make use of prettify() with soup that is print(soup.prettify()) to have proper indentation of the code.
3. To have specific tags data then use soup.tagname and if u only want text within the title tag then use soup.title.text
4. Now the above use will only return the first one matching the code. Eg soup.title.text the the ouput will be the text of very first tag within the code.
5. To have all or other tags within the code having same name we make use of find method or find­\_all method.



1. Czcs



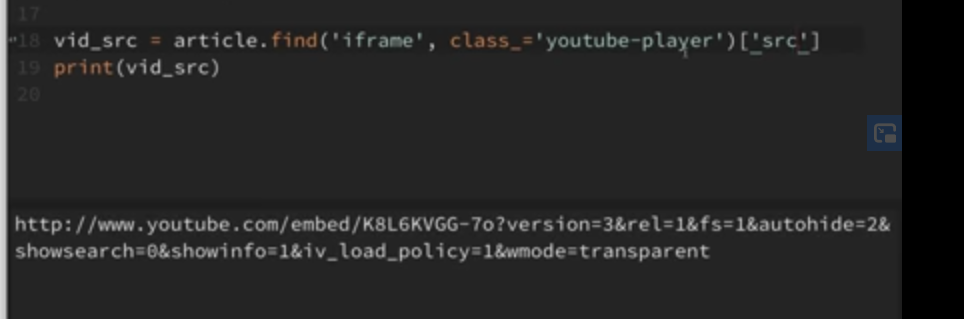




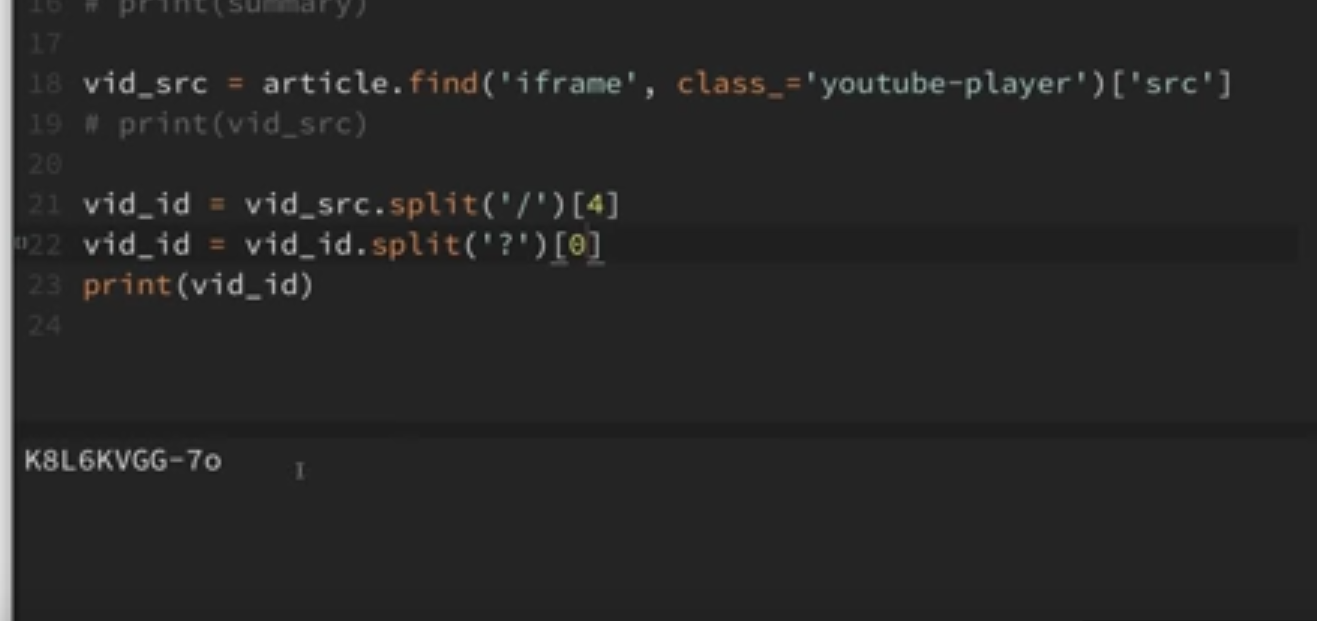


1. dcxcx

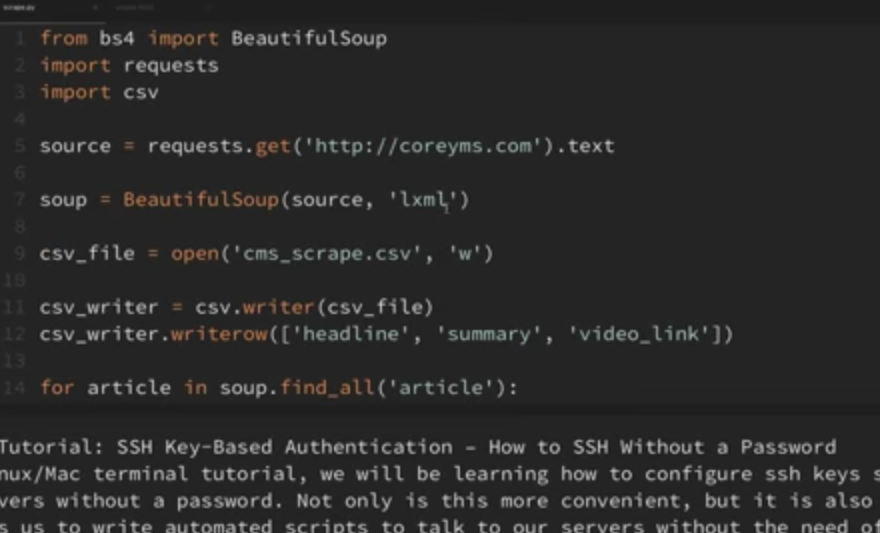
* Ddsds

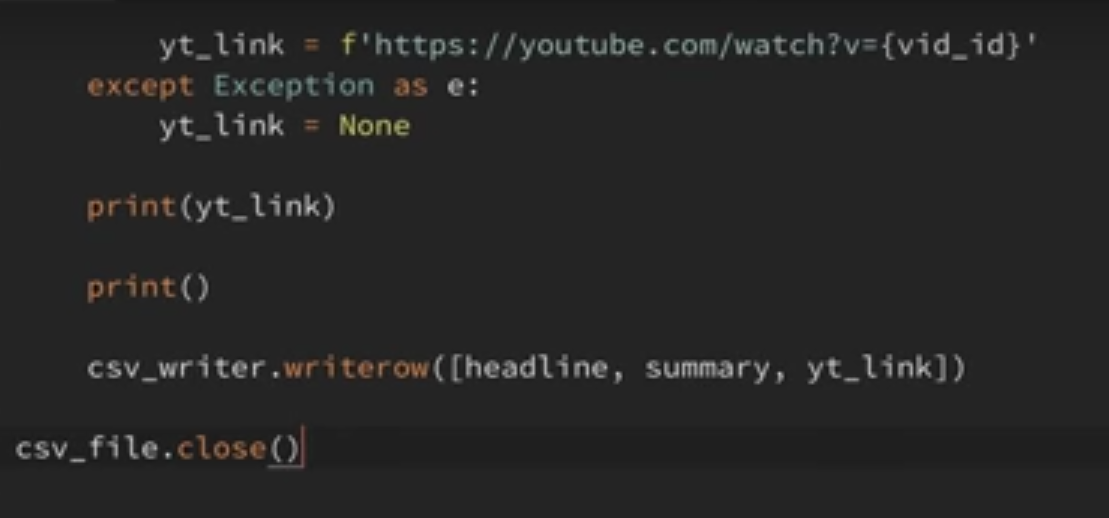


* Sdsd



* Csv file





* Dygsjfs
* sdfsf