4.4 Web Scraping in Python

Web Scraping in Python

Introduction

Extracting specific content from a webpage is called 'Web Scraping". Web pages could be quite complicated to read and understand in its HTML format.
To understand a structure of the web page, we can use a tool such as web browser's inspect tool.

In Firefox or Chrome Browsers, you can right click on any item of interest, and choose Inspect (Right Click + Q). The item will be highlighted in the page, and the source code related to the selected item will be highlighted in a source-code view, (DOM and Style Inspector: Ctrl+Shift+C). You can identify the related tags, class or ID related to the interested item by observing the corresponding source code.

BeautifulSoup4

To navigate and search the HTML document using python, a library known as BeautifulSoup can be used. You can read about the library and its functions here.

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Beautiful library takes in an HTML web bage, and provide us python methods to navigate and access the individual tags within it.

Let us try some example uses:

There are many more techniques which you can use to locate any specific content easily.

Exercise : Extract the topic and the link of the Google search results

As an exercise , let us try to extract only the search results topic and the link from a google search results web page.

First, you have to open the specific web page in the web browser and use the Inspect tool to study the structure of the page, specially around the item of interest.

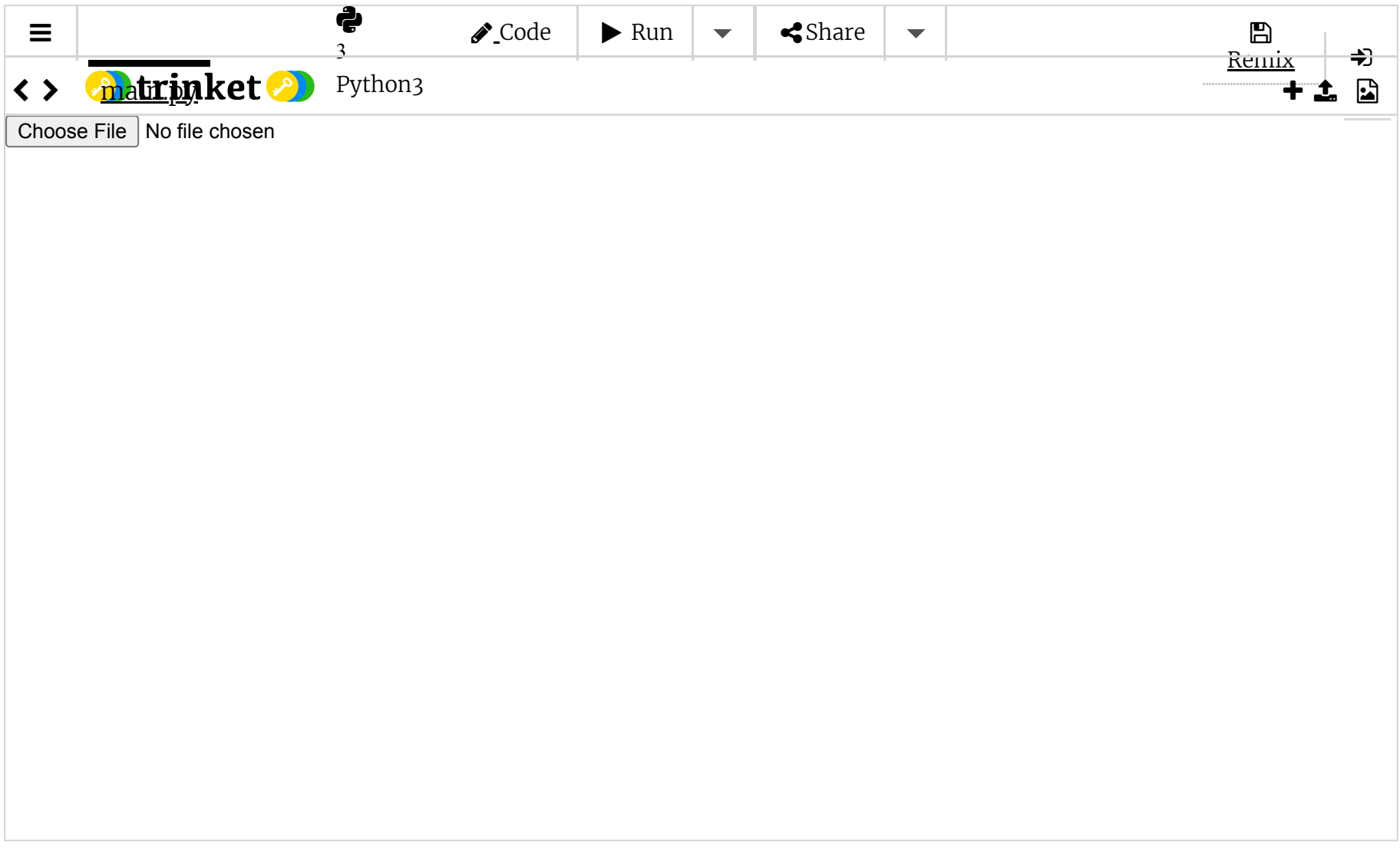
We are using the following link, which will produce the Google search results for the term "Sri+Lanka"

https://www.google.com/search?q=Sri+Lanka

After inspecting the web page html source in the browser, we figured-out the titles of search results are in h3 tags, and only the titles use the h3 tags in entire page. Because of this, we can ask the beautiful soup for the h3 tags, if we need all the titles.

Since the related link must also be quite close to the title, we decide to check the other siblings of the title. You can iteratively check for a id, class, or a numbered order of the specific child to isolated the necessary data.

In the following trinket, you can try this for yourself.



GET IN TOUCH

University of Moratuwa
Centre for Open & Distance Learning
CODL

011 308 2787/8

011 265 0301 ext. 3850,3851

open@uom.lk

University Website

CODL Website

Data retention summary