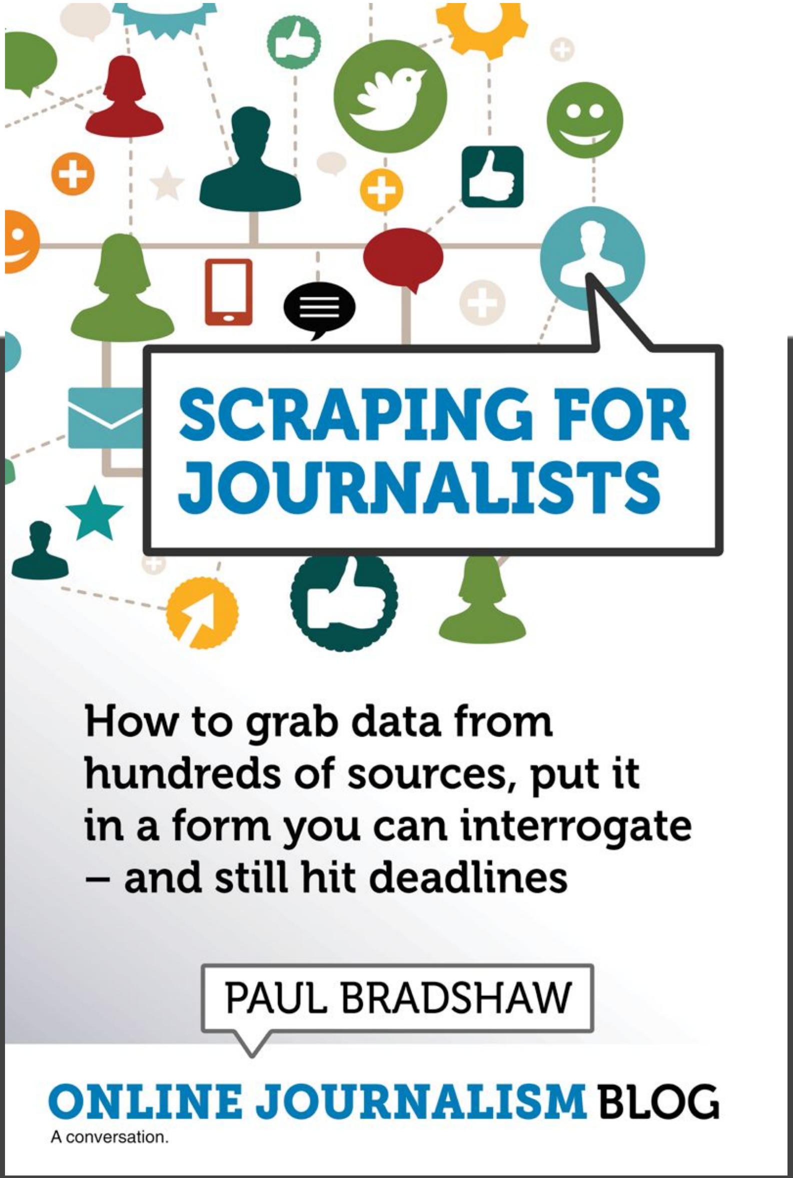
Scraping 2: libraries


SCRAPING FOR JOURNALISTS
How to grab data from hundreds of sources, put it in a form you can interrogate – and still hit deadlines
PAUL BRADSHAW
ONLINE JOURNALISM BLOG
A conversation.

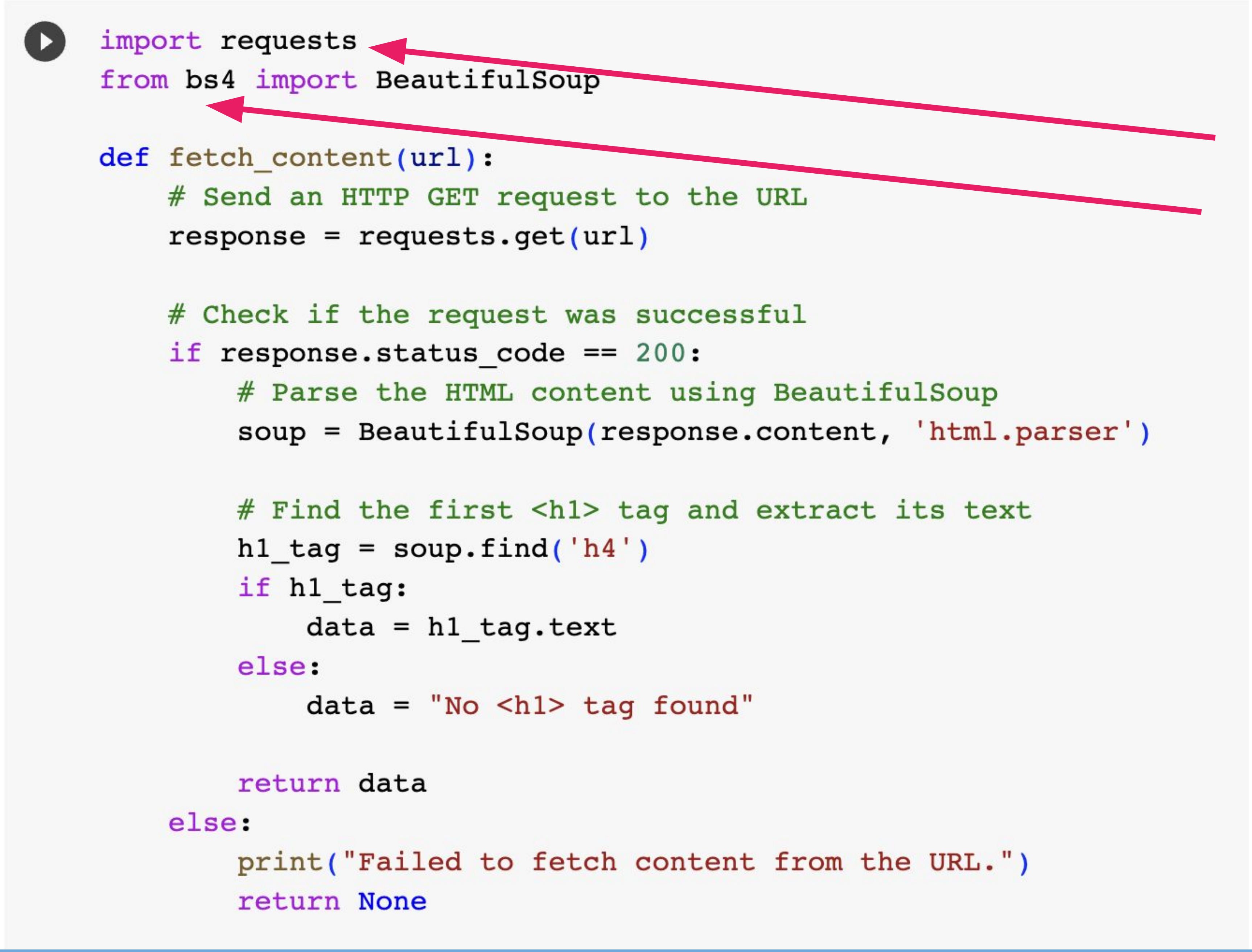
Paul Bradshaw
Leanpub.com/scrapingforjournalists

What we’ll cover

- What are **libraries** in Python - and why you need to know
- How to **import** libraries in a Python notebook in Google Colab

Libraries

- A library is a **collection of recipes (functions)** and other stuff that someone has created for a particular type of problem
- Make it possible to ‘stand on the shoulders of giants’ & use code created by others
- E.g. the **Beautiful Soup (bs4)** library is a collection of tools for solving scraping problems
- And **requests** is a library for fetching URLs
- **Pandas** is a library for data analysis
- **Matplotlib** is a library for visualisation



```python
import requests
from bs4 import BeautifulSoup

def fetch_content(url):
    # Send an HTTP GET request to the URL
    response = requests.get(url)

    # Check if the request was successful
    if response.status_code == 200:
        # Parse the HTML content using BeautifulSoup
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find the first <h1> tag and extract its text
        h1_tag = soup.find('h4')
        if h1_tag:
            data = h1_tag.text
        else:
            data = "No <h1> tag found"

        return data
    else:
        print("Failed to fetch content from the URL.")
        return None
```

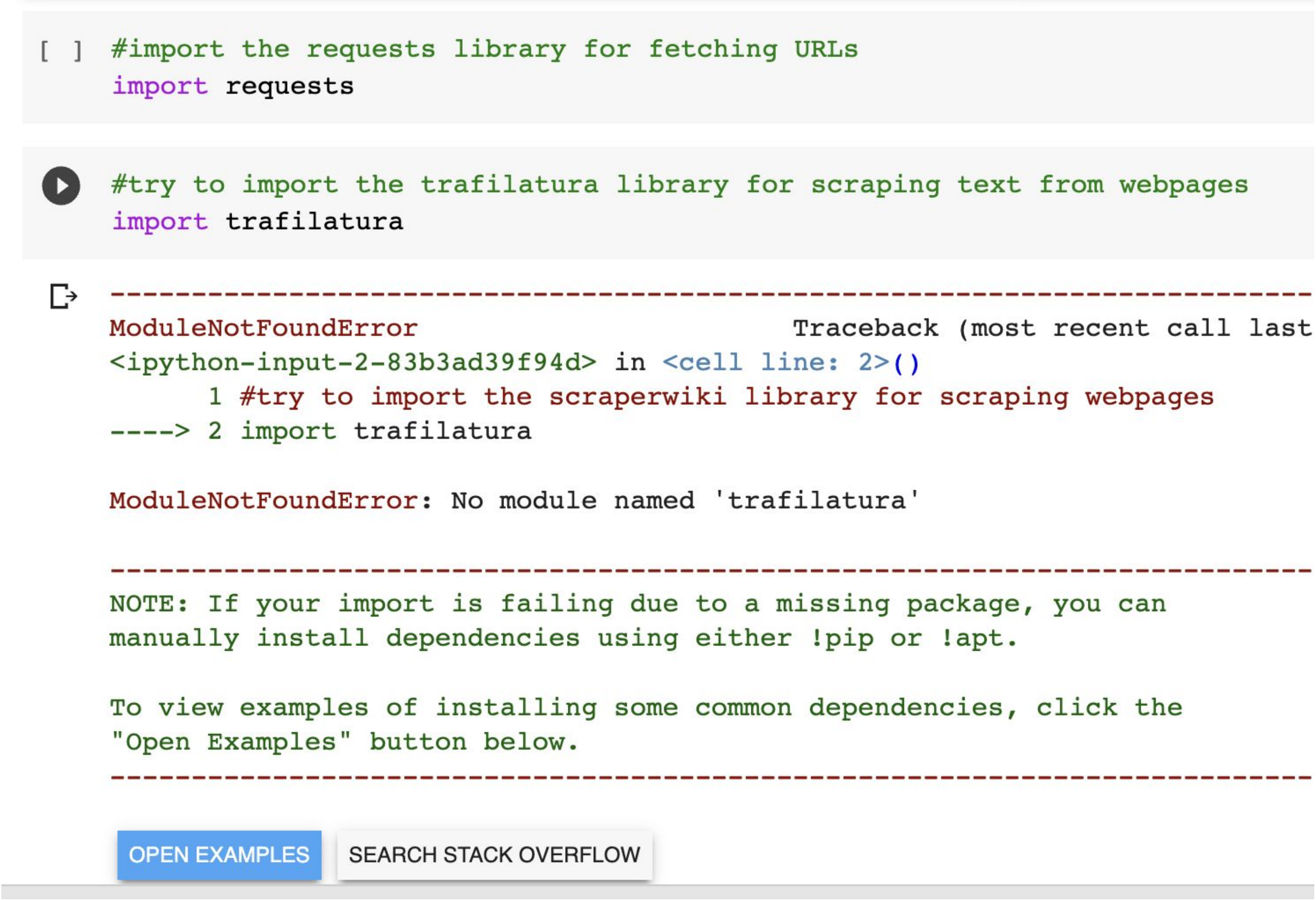
Spot the libraries

Libraries… in Colab

- (Some) libraries need **installing** first
- (All) libraries need **importing**

(How do you know?)

Trial and error...



```
#import the requests library for fetching URLs
import requests
```

```
#try to import the trafilatura library for scraping text from webpages
import trafilatura
```

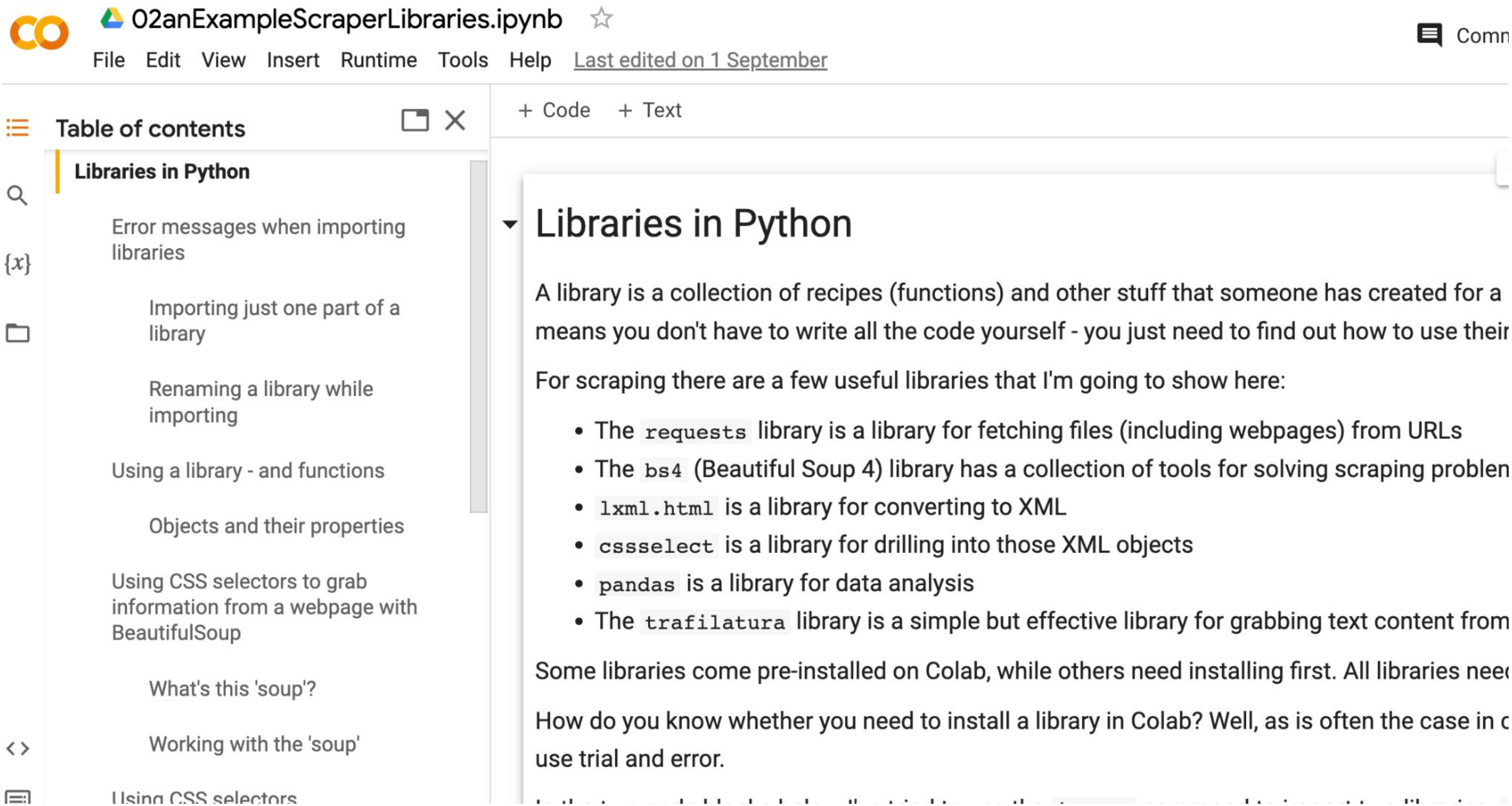
```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last
<ipython-input-2-83b3ad39f94d> in <cell line: 2>()
      1 #try to import the scraperwiki library for scraping webpages
----> 2 import trafilatura

ModuleNotFoundError: No module named 'trafilatura'

---------------------------------------------------------------------------
NOTE: If your import is failing due to a missing package, you can
manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the
"Open Examples" button below.
---------------------------------------------------------------------------
```

```
#install the library
!pip install trafilatura
#import the library
import trafilatura
```



Libraries in Python

A library is a collection of recipes (functions) and other stuff that someone has created for a means you don't have to write all the code yourself - you just need to find out how to use their

For scraping there are a few useful libraries that I'm going to show here:

- The `requests` library is a library for fetching files (including webpages) from URLs
- The `bs4` (Beautiful Soup 4) library has a collection of tools for solving scraping problen
- `lxml.html` is a library for converting to XML
- `cssselect` is a library for drilling into those XML objects
- `pandas` is a library for data analysis
- The `trafilatura` library is a simple but effective library for grabbing text content from

Some libraries come pre-installed on Colab, while others need installing first. All libraries neec

How do you know whether you need to install a library in Colab? Well, as is often the case in c use trial and error.

https://colab.research.google.com/drive/13ULV_uHsQaTFW3oshohL99ZksNgLjEz8?usp=sharing

`import pandas as pd`

- A library can be **renamed** at the same time as it is imported (typically with shorter names for convenience)
- ...because when you use a function from a library you need to name the library

`from bs4 import BeautifulSoup`

- Sometimes you’ll find code where only part of a library is imported (just one function)
- In this case the name of the library is **bs4** but we only want to use **BeautifulSoup**
- You don’t need to know any of this for the code to work!

Using a library

- When you use a **function** from a library you name the library and the function, with a period joining them:
- **requests.**get(fullurl)
- **pandas.**DataFrame(columns=["title"])

  ...or if renamed when imported:
  pd. DataFrame(columns=["title"])

Hold on — functions?

Functions = recipes

- A **function** is a name for a recipe. Used in Excel, e.g. SUM, AVERAGE, VLOOKUP
- A function is always followed by parentheses to ‘pass’ any ingredients, e.g. =SUM(A1:A10)
- requests.**get(fullurl)**
- pd.**DataFrame(columns=["title"])**

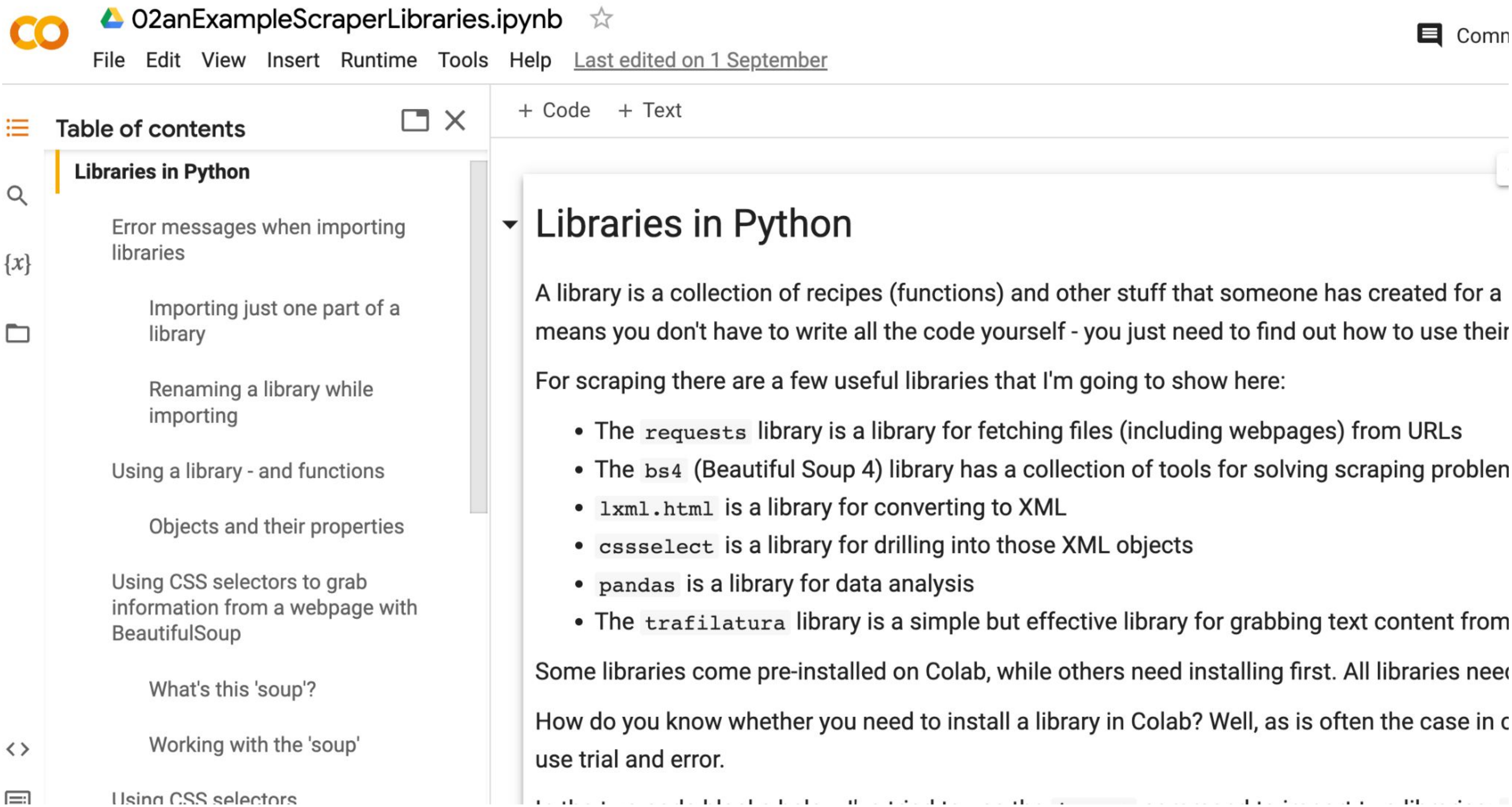
Recap

- A library is (pre-)installed, and imported:

```
!pip install trafilatura
import trafilatura
import requests
```

- Functions (recipes) from that library are joined by a period and followed by parentheses:

```
html = requests.get(“http://blah.com”)
```



Libraries in Python

Error messages when importing libraries

Importing just one part of a library

Renaming a library while importing

Using a library - and functions

Objects and their properties

Using CSS selectors to grab information from a webpage with BeautifulSoup

What's this 'soup'?

Working with the 'soup'

Using CSS selectors

Libraries in Python

A library is a collection of recipes (functions) and other stuff that someone has created for a means you don't have to write all the code yourself - you just need to find out how to use their

For scraping there are a few useful libraries that I'm going to show here:

- The `requests` library is a library for fetching files (including webpages) from URLs
- The `bs4` (Beautiful Soup 4) library has a collection of tools for solving scraping problen
- `lxml.html` is a library for converting to XML
- `cssselect` is a library for drilling into those XML objects
- `pandas` is a library for data analysis
- The `trafilatura` library is a simple but effective library for grabbing text content from

Some libraries come pre-installed on Colab, while others need installing first. All libraries neec

How do you know whether you need to install a library in Colab? Well, as is often the case in c use trial and error.

https://colab.research.google.com/drive/13ULV_uHsQaTFW3oshohL99ZksNgLjEz8?usp=sharing

Try it now:

- Create a notebook and import the libraries we will need:
  - **import requests**
  - **from bs4 import BeautifulSoup**
  - **import pandas as pd**

- Tip: If you get an error, ask Gemini/ChatGPT what you might have done wrong