web crawler

from urllib.request import urlopen  
  
url = "https://www.baidu.com"  
resp = urlopen(url)  
  
print(resp.read())

The output will be like:

Graphical user interface, text

Description automatically generated

charset=utf-8 shows how to decode:

print(resp.read().decode('utf-8'))

Text

Description automatically generated

We will find the numbers above will change into Chinese characters, this is because it was encoding with ‘utf-8’, if we want to see the exact character, we have to do decoding first.

Also, we can write the response into a file:

from urllib.request import urlopen  
  
url = "https://www.baidu.com"  
resp = urlopen(url)  
  
#print(resp.read().decode('utf-8'))  
  
with open("mybaidu.html", mode="w", encoding="utf-8") as f:  
 f.write(resp.read().decode("utf-8"))  
print("over!")

please remember to add

encoding="utf-8"

Otherwise, we will get:

Text

Description automatically generated

Graphical user interface, application, website

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

After running the code, we will get a file named “mybaidu.html”, if we open it, it will be just like the original website (i.e. the url)

Graphical user interface, text, application, email

Description automatically generated