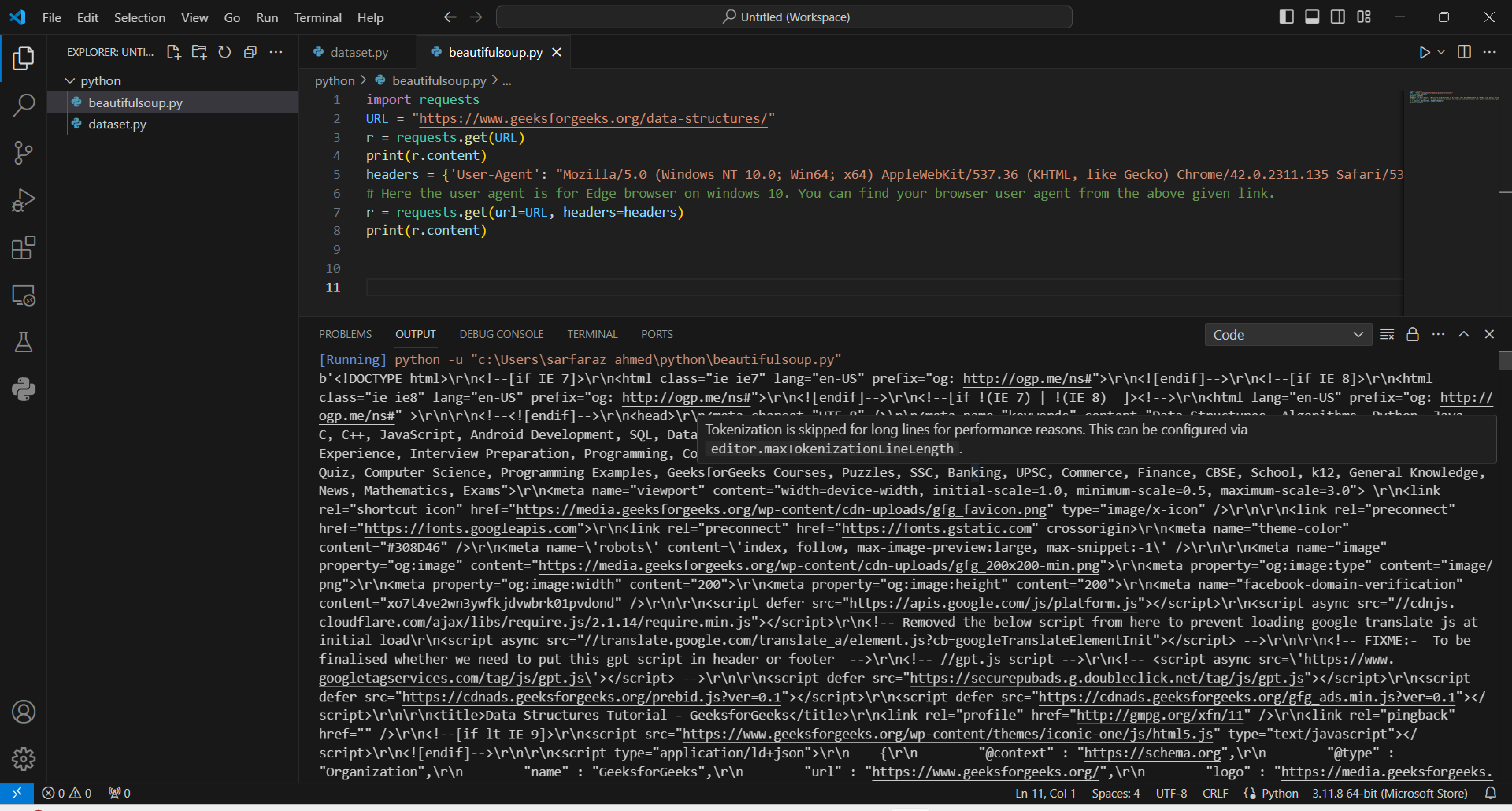
Assignment 4

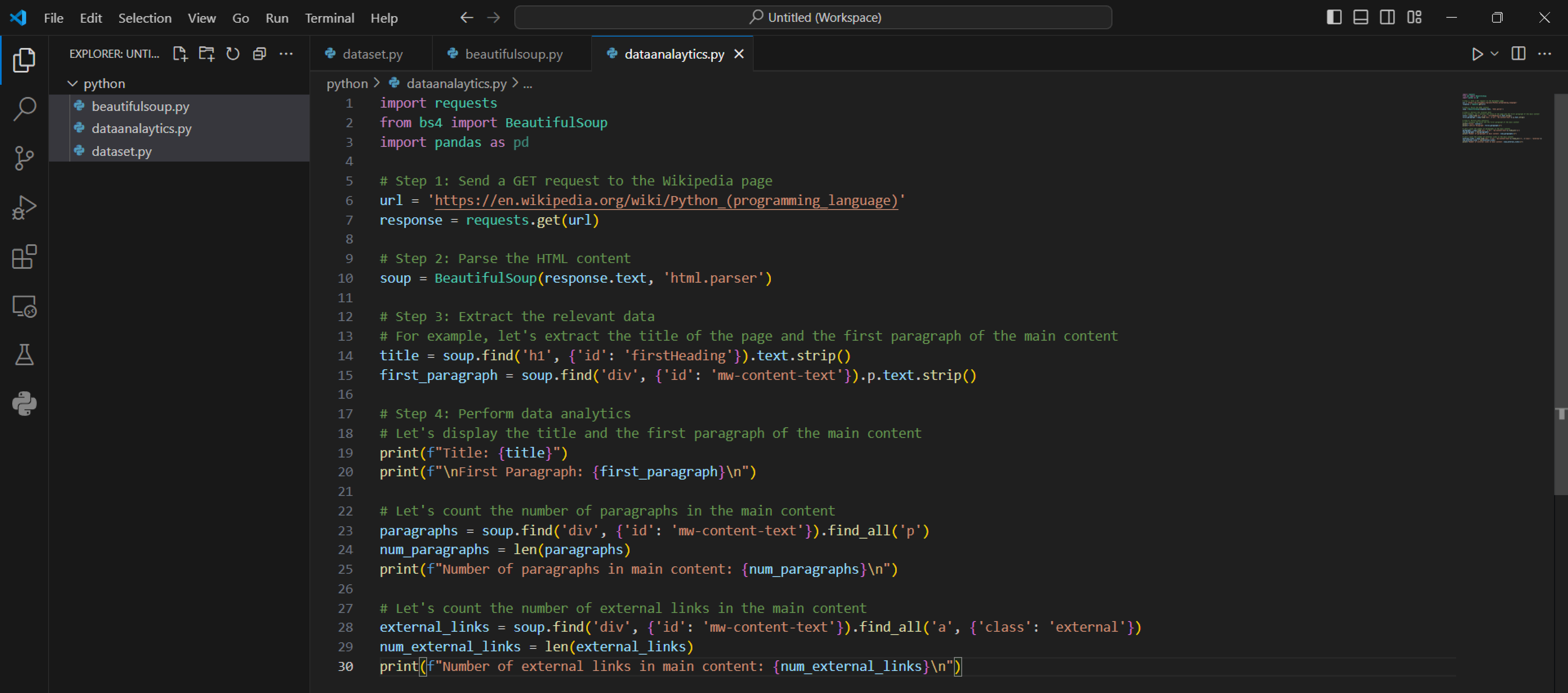
1. Take any of your own URL , do WEB Scraping using requests/beautifulsop modules and complete data analytics.



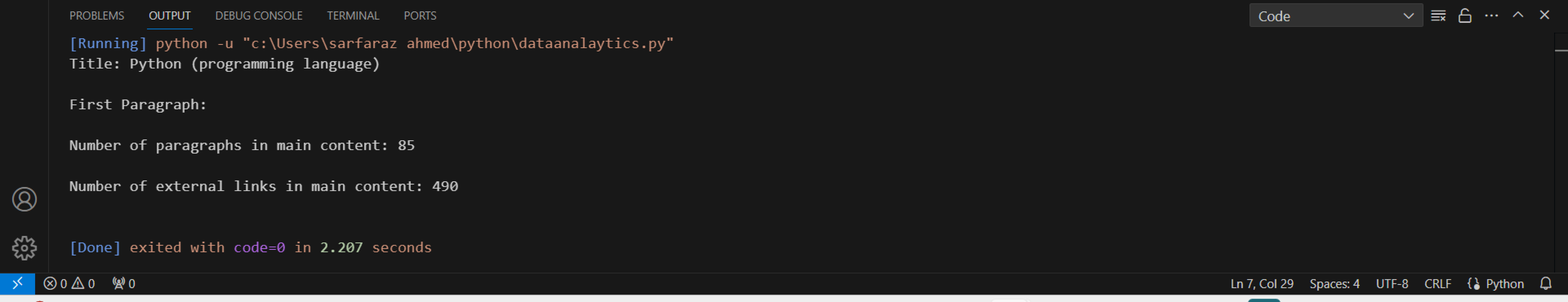
```python
import requests
URL = "https://www.geeksforgeeks.org/data-structures/"
r = requests.get(URL)
print(r.content)
headers = {'User-Agent': "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.135 Safari/53
# Here the user agent is for Edge browser on windows 10. You can find your browser user agent from the above given link.
r = requests.get(url=URL, headers=headers)
print(r.content)
```

```
[Running] python -u "c:\Users\sarfaraz ahmed\python\beautifulsoup.py"
b'<!DOCTYPE html>\r\n<!--[if IE 7]>\r\n<html class="ie ie7" lang="en-US" prefix="og: http://ogp.me/ns#">\r\n<![endif]-->\r\n<!--[if IE 8]>\r\n<html
class="ie ie8" lang="en-US" prefix="og: http://ogp.me/ns#">\r\n<![endif]-->\r\n<!--[if !(IE 7) | !(IE 8)  ]><!-->\r\n<html lang="en-US" prefix="og: http://
ogp.me/ns#" >\r\n\r\n<!--<![endif]-->\r\n<head>\r\n
C, C++, JavaScript, Android Development, SQL, Data
Experience, Interview Preparation, Programming, Co
Quiz, Computer Science, Programming Examples, GeeksforGeeks Courses, Puzzles, SSC, Banking, UPSC, Commerce, Finance, CBSE, School, k12, General Knowledge,
News, Mathematics, Exams">\r\n<meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=0.5, maximum-scale=3.0"> \r\n<link
rel="shortcut icon" href="https://media.geeksforgeeks.org/wp-content/cdn-uploads/gfg_favicon.png" type="image/x-icon" />\r\n\r\n<link rel="preconnect"
href="https://fonts.googleapis.com">\r\n<link rel="preconnect" href="https://fonts.gstatic.com" crossorigin>\r\n<meta name="theme-color"
content="#308D46" />\r\n<meta name=\'robots\' content=\'index, follow, max-image-preview:large, max-snippet:-1\' />\r\n\r\n<meta name="image"
property="og:image" content="https://media.geeksforgeeks.org/wp-content/cdn-uploads/gfg_200x200-min.png">\r\n<meta property="og:image:type" content="image/
png">\r\n<meta property="og:image:width" content="200">\r\n<meta property="og:image:height" content="200">\r\n<meta name="facebook-domain-verification"
content="xo7t4ve2wn3ywfkjdvwbrk01pvdond" />\r\n\r\n<script defer src="https://apis.google.com/js/platform.js"></script>\r\n<script async src="//cdnjs.
cloudflare.com/ajax/libs/require.js/2.1.14/require.min.js"></script>\r\n<!-- Removed the below script from here to prevent loading google translate js at
initial load\r\n<script async src="//translate.google.com/translate_a/element.js?cb=googleTranslateElementInit"></script> -->\r\n\r\n<!-- FIXME:-  To be
finalised whether we need to put this gpt script in header or footer  -->\r\n<!-- //gpt.js script -->\r\n<!-- <script async src=\'https://www.
googletagservices.com/tag/js/gpt.js\'></script> -->\r\n\r\n<script defer src="https://securepubads.g.doubleclick.net/tag/js/gpt.js"></script>\r\n<script
defer src="https://cdnads.geeksforgeeks.org/prebid.js?ver=0.1"></script>\r\n<script defer src="https://cdnads.geeksforgeeks.org/gfg_ads.min.js?ver=0.1"></
script>\r\n\r\n<title>Data Structures Tutorial - GeeksforGeeks</title>\r\n<link rel="profile" href="http://gmpg.org/xfn/11" />\r\n<link rel="pingback"
href="" />\r\n<!--[if lt IE 9]>\r\n<script src="https://www.geeksforgeeks.org/wp-content/themes/iconic-one/js/html5.js" type="text/javascript"></
script>\r\n<![endif]-->\r\n\r\n<script type="application/ld+json">\r\n    {\r\n        "@context" : "https://schema.org",\r\n        "@type" :
"Organization",\r\n        "name" : "GeeksforGeeks",\r\n        "url" : "https://www.geeksforgeeks.org/",\r\n        "logo" : "https://media.geeksforgeeks.
```

Tokenization is skipped for long lines for performance reasons. This can be configured via `editor.maxTokenizationLineLength`.



```python
import requests
from bs4 import BeautifulSoup
import pandas as pd

# Step 1: Send a GET request to the Wikipedia page
url = 'https://en.wikipedia.org/wiki/Python_(programming_language)'
response = requests.get(url)

# Step 2: Parse the HTML content
soup = BeautifulSoup(response.text, 'html.parser')

# Step 3: Extract the relevant data
# For example, let's extract the title of the page and the first paragraph of the main content
title = soup.find('h1', {'id': 'firstHeading'}).text.strip()
first_paragraph = soup.find('div', {'id': 'mw-content-text'}).p.text.strip()

# Step 4: Perform data analytics
# Let's display the title and the first paragraph of the main content
print(f"Title: {title}")
print(f"\nFirst Paragraph: {first_paragraph}\n")

# Let's count the number of paragraphs in the main content
paragraphs = soup.find('div', {'id': 'mw-content-text'}).find_all('p')
num_paragraphs = len(paragraphs)
print(f"Number of paragraphs in main content: {num_paragraphs}\n")

# Let's count the number of external links in the main content
external_links = soup.find('div', {'id': 'mw-content-text'}).find_all('a', {'class': 'external'})
num_external_links = len(external_links)
print(f"Number of external links in main content: {num_external_links}\n")
```



```
[Running] python -u "c:\Users\sarfaraz ahmed\python\dataanalaytics.py"
Title: Python (programming language)

First Paragraph:

Number of paragraphs in main content: 85

Number of external links in main content: 490


[Done] exited with code=0 in 2.207 seconds
```