```
!pip install requests beautiful-soup4
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (2.27.1)
ERROR: Could not find a version that satisfies the requirement beautiful-soup4 (from versions: none)
ERROR: No matching distribution found for beautiful-soup4
```

```
import requests
from bs4 import BeautifulSoup
import os

class Crawler:
    def __init__(self):
        self.session = requests.Session()

    def get_soup(self, url):
        try:
            response = self.session.get(url, timeout=10)
            if response.status_code == 200:
                html = response.content
                # Save to a file
                filename = url.replace('https://', '').replace('http://', '').replace('/', '_') + '.html'
                if not os.path.exists('downloaded_pages'):
                    os.makedirs('downloaded_pages')
                with open(os.path.join('downloaded_pages', filename), 'wb') as f:
                    f.write(html)
                return BeautifulSoup(html, 'html.parser')
            elif response.status_code == 301:
                print(f'Redirected: {response.url}')
                return None
            elif response.status_code == 403:
                print('Forbidden')
                return None
            elif response.status_code == 404:
                print('Not Found')
                return None
            elif response.status_code == 500:
                print('Internal Server Error')
                return None
            else:
                print(f'Unexpected HTTP response: {response.status_code}')
                return None
        except requests.exceptions.RequestException as e:
            print(f'RequestException: {str(e)}')
            return None

    def crawl(self, url, max_depth=3):
        if max_depth < 0:
            return
        print(f'Depth: {max_depth} - Visiting: {url}')
        soup = self.get_soup(url)
        if soup is None:
            print("Failed to get page")
            return
```

```
            for link in soup.find_all('a'):
                href = link.get('href')
                if href and href.startswith('http'):
                    self.crawl(href, max_depth-1)

    if __name__ == "__main__":
        crawler = Crawler()
        crawler.crawl("https://example.com", 2)
```

```
         Depth: 2 - Visiting: https://example.com
         Depth: 1 - Visiting: https://www.iana.org/domains/example
         Depth: 0 - Visiting: http://www.icann.org/topics/idn/
         Depth: 0 - Visiting: http://www.icann.org/
         Depth: 0 - Visiting: http://www.icann.org/en/registries/agreements.htm
         Depth: 0 - Visiting: http://pti.icann.org
         Depth: 0 - Visiting: http://www.icann.org/
         Depth: 0 - Visiting: https://www.icann.org/privacy/policy
         Depth: 0 - Visiting: https://www.icann.org/privacy/tos
```

```
    #Crawl the Wikipedia
    crawler = Crawler()
    crawler.crawl("https://en.wikipedia.org/wiki/Main_Page", 3)
```

```
Depth: 0 - Visiting: https://wikimediafoundation.org/our-work/wikimedia-projects/
Depth: 0 - Visiting: https://shop.wikimedia.org
Depth: 0 - Visiting: https://www.mediawiki.org/wiki/Special:MyLanguage/Wikimedia_Cloud_Services
Depth: 0 - Visiting: https://en.wikipedia.org/wiki/Main_Page
Depth: 0 - Visiting: https://commons.wikimedia.org/wiki/Main_Page
Depth: 0 - Visiting: https://en.wiktionary.org/wiki/Wiktionary:Main_Page
Depth: 0 - Visiting: https://wikimediafoundation.org/
Depth: 0 - Visiting: https://wikimediafoundation.org/our-work/wikimedia-projects/
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Wikimedia_chapters
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Wikimedia_thematic_organizations
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Wikimedia_user_groups
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Grants:Start/About
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Projects
Depth: 0 - Visiting: https://shop.wikimedia.org
Depth: 0 - Visiting: https://labs.wikimedia.org
Depth: 0 - Visiting: https://store.wikimedia.org/pages/copy-of-privacy-policy
Depth: 0 - Visiting: https://www.mediawiki.org/wiki/Special:MyLanguage/Wikimedia_Apps/Synced_Reading_Lists
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/CheckUser
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Stewards
Depth: 0 - Visiting: https://www.mediawiki.org/wiki/Special:MyLanguage/Wikimedia_Cloud_Services
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Volunteer_Response_Team
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/IRC
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/Wikimedia_movement_affiliates
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Special:MyLanguage/CheckUser_policy
Depth: 0 - Visiting: https://lists.wikimedia.org/mailman/listinfo/wikimediaannounce-l
Depth: 0 - Visiting: https://lists.wikimedia.org/mailman/listinfo/wikimediaannounce-l
Depth: 0 - Visiting: https://lists.wikimedia.org/mailman/listinfo/wikimediaannounce-l
Depth: 0 - Visiting: https://wikimediafoundation.org/about/contact/
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/User:AKeton_(WMF)
Depth: 0 - Visiting: https://foundation.wikimedia.org/w/index.php?title=Template:Privacy_policy_navigation_2&action=edit
Depth: 0 - Visiting: https://foundation.wikimedia.org/w/index.php?title=Special:Translate/page-Template:Privacy_policy_navigation_2&language=en&action=page
Depth: 0 - Visiting: https://meta.wikimedia.org/wiki/Access_to_nonpublic_personal_data_policy/Noticeboard
Depth: 0 - Visiting: https://foundation.wikimedia.org/w/index.php?title=Policy:Privacy_policy&oldid=238682
Depth: 0 - Visiting: https://creativecommons.org/licenses/by-sa/3.0/
Depth: 0 - Visiting: https://foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Terms_of_Use
Depth: 0 - Visiting: https://developer.wikimedia.org
Depth: 0 - Visiting: https://stats.wikimedia.org/#/foundation.wikimedia.org
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-9-6f2571fd4462> in <cell line: 3>()
      1 #Crawl the Wikipedia
      2 crawler = Crawler()
----> 3 crawler.crawl("https://en.wikipedia.org/wiki/Main_Page", 3)

                                 14 frames
/usr/local/lib/python3.10/dist-packages/urllib3/util/connection.py in create_connection(address, timeout, source_address, socket_options)
     83             if source_address:
     84                 sock.bind(source_address)
---> 85             sock.connect(sa)
     86             return sock
     87

KeyboardInterrupt:

SEARCH STACK OVERFLOW
```

```
#Crawl Python Wevbsite
crawler = Crawler()
crawler.crawl("https://www.python.org", 1)
```

```
    Depth: 0 - Visiting: https://devguide.python.org/
    Depth: 0 - Visiting: https://docs.python.org/faq/
    Depth: 0 - Visiting: http://wiki.python.org/moin/Languages
    Depth: 0 - Visiting: http://python.org/dev/peps/
    Depth: 0 - Visiting: https://wiki.python.org/moin/PythonBooks
    Depth: 0 - Visiting: https://wiki.python.org/moin/
    Depth: 0 - Visiting: http://pyfound.blogspot.com/
    Depth: 0 - Visiting: http://pycon.blogspot.com/
    Depth: 0 - Visiting: http://planetpython.org/
    Depth: 0 - Visiting: https://wiki.python.org/moin/PythonEventsCalendar#Submitting_an_Event
    Depth: 0 - Visiting: http://docs.python.org/3/tutorial/introduction.html#using-python-as-a-calculator
    Depth: 0 - Visiting: https://docs.python.org
    Depth: 0 - Visiting: https://blog.python.org
    Depth: 0 - Visiting: https://pyfound.blogspot.com/2023/05/the-python-language-summit-2023-towards.html
    Depth: 0 - Visiting: https://pyfound.blogspot.com/2023/05/the-python-language-summit-2023-burnout.html
    Depth: 0 - Visiting: https://pyfound.blogspot.com/2023/05/the-python-language-summit-2023-making.html
    Depth: 0 - Visiting: https://pyfound.blogspot.com/2023/05/the-python-language-summit-2023-three.html
    Depth: 0 - Visiting: https://pyfound.blogspot.com/2023/05/the-python-language-summit-2023-python.html
    Depth: 0 - Visiting: http://www.djangoproject.com/
    Depth: 0 - Visiting: http://www.pylonsproject.org/
    Depth: 0 - Visiting: http://bottlepy.org
    Depth: 0 - Visiting: http://tornadoweb.org
    Depth: 0 - Visiting: http://flask.pocoo.org/
    Depth: 0 - Visiting: http://www.web2py.com/
    Depth: 0 - Visiting: http://wiki.python.org/moin/TkInter
    Depth: 0 - Visiting: https://wiki.gnome.org/Projects/PyGObject
    Depth: 0 - Visiting: http://www.riverbankcomputing.co.uk/software/pyqt/intro
    Depth: 0 - Visiting: https://wiki.qt.io/PySide
    Depth: 0 - Visiting: https://kivy.org/
    Depth: 0 - Visiting: http://www.wxpython.org/
    Depth: 0 - Visiting: http://www.scipy.org
    Depth: 0 - Visiting: http://pandas.pydata.org/
    Depth: 0 - Visiting: http://ipython.org
    Depth: 0 - Visiting: http://buildbot.net/
    Depth: 0 - Visiting: http://trac.edgewall.org/
    Depth: 0 - Visiting: http://roundup.sourceforge.net/
    Depth: 0 - Visiting: http://www.ansible.com
    Depth: 0 - Visiting: https://saltproject.io
    Depth: 0 - Visiting: https://www.openstack.org
    Depth: 0 - Visiting: https://xon.sh
    Depth: 0 - Visiting: http://brochure.getpython.info/
    Depth: 0 - Visiting: https://docs.python.org/3/license.html
    Depth: 0 - Visiting: https://wiki.python.org/moin/BeginnersGuide
    Depth: 0 - Visiting: https://devguide.python.org/
    Depth: 0 - Visiting: https://docs.python.org/faq/
```

```
Depth: 0 - Visiting: https://mail.python.org/mailman/listinfo/python-dev
Depth: 0 - Visiting: https://github.com/python/pythondotorg/issues
Depth: 0 - Visiting: https://status.python.org/
```

```
#Crawl the New York Times website
crawler = Crawler()
crawler.crawl("https://www.nytimes.com", 2)
```

```
Depth: 1 - Visiting: https://www.nytimes.com/section/todayspaper
Depth: 0 - Visiting: https://www.nytimes.com/section/todayspaper
Depth: 0 - Visiting: https://www.nytimes.com/section/todayspaper
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us/articles/115014792127-Copyright-notice
Forbidden
Failed to get page
Depth: 0 - Visiting: https://www.nytco.com/
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us/articles/115015385887-Contact-Us
Forbidden
Failed to get page
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us/articles/115015727108-Accessibility
Forbidden
Failed to get page
Depth: 0 - Visiting: https://www.nytco.com/careers/
Depth: 0 - Visiting: https://nytmediakit.com/
Depth: 0 - Visiting: https://www.tbrandstudio.com/
Depth: 0 - Visiting: https://www.nytimes.com/privacy/cookie-policy#how-do-i-manage-trackers
Depth: 0 - Visiting: https://www.nytimes.com/privacy/privacy-policy
Forbidden
Failed to get page
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-service
Forbidden
Failed to get page
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us/articles/115014893968-Terms-of-sale
Forbidden
Failed to get page
Depth: 0 - Visiting: https://www.nytimes.com/ca/?action=click&region=Footer&pgtype=Homepage
Depth: 0 - Visiting: https://www.nytimes.com/international/?action=click&region=Footer&pgtype=Homepage
Depth: 0 - Visiting: https://help.nytimes.com/hc/en-us
Forbidden
Failed to get page
Depth: 0 - Visiting: https://www.nytimes.com/subscription?campaignId=37WXW
Depth: 1 - Visiting: https://cooking.nytimes.com/
Depth: 0 - Visiting: http://www.facebook.com/sharer/sharer.php?u=https://cooking.nytimes.com/recipes/1023145-tajin-grilled-chicken%3Fsmid=fb-share
Depth: 0 - Visiting: http://www.pinterest.com/pin/create/button/?url=https%3A%2F%2Fcooking.nytimes.com%2Frecipes%2F1023145-tajin-grilled-chicken%3Fsmid=pin-share&descript
---------------------------------------------------------------------------
OSError                                   Traceback (most recent call last)
<ipython-input-11-f5b98cacb230> in <cell line: 3>()
      1 #Crawl the New York Times website
      2 crawler = Crawler()
----> 3 crawler.crawl("https://www.nytimes.com", 2)

                                 3 frames
<ipython-input-8-9f98ede16001> in get_soup(self, url)
     16                 if not os.path.exists('downloaded_pages'):
     17                     os.makedirs('downloaded_pages')
---> 18                 with open(os.path.join('downloaded_pages', filename), 'wb') as f:
     19                     f.write(html)
     20                 return BeautifulSoup(html, 'html.parser')

OSError: [Errno 36] File name too long: 'downloaded_pages/www.pinterest.com_pin_create_button_?url=https%3A%2F%2Fcooking.nytimes.com%2Frecipes%2F1023145-tajin-grilled-chicken%3Fsmid=pin-share&description=NYT%20Cooking:%20Tajín%20Grilled%20Chicken&media=https%3A%2F%2Fstatic01.nyt.com%2Fimages%2F2022%2F05%2F25%2Fdining%2F25Memorial-tajin-grilled-chicken%2Fmerlin_206068182_ffd1a522-033b-4519-969f-1990ecb8a8f0-threeByTwoMediumAt2X.jpg.html'

SEARCH STACK OVERFLOW
```