In [1]:
```
# pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /Users/dhyan/opt/anaconda3
/lib/python3.9/site-packages (4.10.0)
Requirement already satisfied: soupsieve>1.2 in /Users/dhyan/opt/anaconda3/
lib/python3.9/site-packages (from beautifulsoup4) (2.2.1)

[notice] A new release of pip is available: 23.0 -> 23.0.1
[notice] To update, run: pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
# pip install requests
# !pip install bs4
```

```
Collecting bs4
  Downloading bs4-0.0.1.tar.gz (1.1 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: beautifulsoup4 in /Users/dhyan/opt/anaconda3
/lib/python3.9/site-packages (from bs4) (4.10.0)
Requirement already satisfied: soupsieve>1.2 in /Users/dhyan/opt/anaconda3/
lib/python3.9/site-packages (from beautifulsoup4->bs4) (2.2.1)
Building wheels for collected packages: bs4
  Building wheel for bs4 (setup.py) ... done
  Created wheel for bs4: filename=bs4-0.0.1-py3-none-any.whl size=1256 sha2
56=9ebb5b366ad4259bfa28e2d4dfa72c4ff29a15a5639466ea01efd3474bc9ab3a
  Stored in directory: /Users/dhyan/Library/Caches/pip/wheels/73/2b/cb/0999
80278a0c9a3e57ff1a89875ec07bfa0b6fcbebb9a8cad3
Successfully built bs4
Installing collected packages: bs4
Successfully installed bs4-0.0.1

[notice] A new release of pip is available: 23.0 -> 23.0.1
[notice] To update, run: pip install --upgrade pip
```

In [4]:
```
# pip install --upgrade pip
```

```
Requirement already satisfied: pip in /Users/dhyan/opt/anaconda3/lib/python
3.9/site-packages (23.0)
Collecting pip
  Downloading pip-23.0.1-py3-none-any.whl (2.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 11.9 MB/s eta 0:00
:0000:0100:01
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 23.0
    Uninstalling pip-23.0:
      Successfully uninstalled pip-23.0
Successfully installed pip-23.0.1
Note: you may need to restart the kernel to use updated packages.
```

In [5]:
```
# !pip install requests
```

```
Requirement already satisfied: requests in /Users/dhyan/opt/anaconda3/lib/p
ython3.9/site-packages (2.26.0)
Requirement already satisfied: charset-normalizer~=2.0.0 in /Users/dhyan/op
t/anaconda3/lib/python3.9/site-packages (from requests) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /Users/dhyan/opt/anaconda3/l
ib/python3.9/site-packages (from requests) (3.2)
Requirement already satisfied: certifi>=2017.4.17 in /Users/dhyan/opt/anaco
nda3/lib/python3.9/site-packages (from requests) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/dhyan/opt/an
aconda3/lib/python3.9/site-packages (from requests) (1.26.7)
```

In [16]:

```
# Import libraries
from urllib.request import urljoin
from bs4 import BeautifulSoup
import requests
from urllib.request import urlparse


# Set for storing urls with same domain
links_intern = set()
input_url = "https://thomasdorfer.medium.com/"
depth = 2

# Set for storing urls with different domain
links_extern = set()


# Method for crawling a url at next level
def level_crawler(input_url):
    temp_urls = set()
    current_url_domain = urlparse(input_url).netloc

    # Creates beautiful soup object to extract html tags
    beautiful_soup_object = BeautifulSoup(
        requests.get(input_url).content, "lxml")

    # Access all anchor tags from input
    # url page and divide them into internal
    # and external categories
    for anchor in beautiful_soup_object.findAll("a"):
        href = anchor.attrs.get("href")
        if(href != "" or href != None):
            href = urljoin(input_url, href)
            href_parsed = urlparse(href)
            href = href_parsed.scheme
            href += "://"
            href += href_parsed.netloc
            href += href_parsed.path
            final_parsed_href = urlparse(href)
            is_valid = bool(final_parsed_href.scheme) and bool(final_parsed
            if is_valid:
                if current_url_domain not in href and href not in links_ext
                    print("External - {}".format(href))
                    links_extern.add(href)
                if current_url_domain in href and href not in links_intern
                    print("Internal - {}".format(href))
                    links_intern.add(href)
                    temp_urls.add(href)
    return temp_urls
```

```
# if(depth == 0):
#       print("Intern - {}".format(input_url))

# elif(depth == 1):
#       level_crawler(input_url)
queue = []
queue.append(input_url)
for j in range(depth):
    for count in range(len(queue)):
        url = queue.pop(0)
        urls = level_crawler(url)
        for i in urls:
            queue.append(i)
```

```
External - https://rsci.app.link/
External - https://medium.com/m/signin
External - https://medium.com/
External - https://medium.com/search
Internal - https://thomasdorfer.medium.com/
Internal - https://thomasdorfer.medium.com/followers
Internal - https://thomasdorfer.medium.com/about
External - https://towardsdatascience.com/
Internal - https://thomasdorfer.medium.com/how-to-stay-on-top-of-the-latest
-ai-research-e8993523ef3e
External - https://medium.com/tag/artificial-intelligence
Internal - https://thomasdorfer.medium.com/enhanced-object-detection-how-to
-effectively-implement-yolov8-afd1bf6132ae
Internal - https://thomasdorfer.medium.com/comparing-list-comprehensions-vs
-built-in-functions-in-python-which-is-better-1e2c9646fafe
External - https://medium.com/tag/data-science
Internal - https://thomasdorfer.medium.com/enhanced-debugging-in-python-tra
cebacks-just-got-a-major-upgrade-bd77fb32db38
Internal - https://thomasdorfer.medium.com/effective-data-visualization-9-v
aluable-tips-to-increase-the-quality-of-your-charts-5fec31144a6d
Internal - https://thomasdorfer.medium.com/how-to-effectively-use-lambda-fu
nctions-in-python-as-a-data-scientist-fd6171554053
Internal - https://thomasdorfer.medium.com/data-science-in-small-and-big-co
mpanies-5cb32be1491a
Internal - https://thomasdorfer.medium.com/why-simple-models-are-often-bett
er-e2428964811a
Internal - https://thomasdorfer.medium.com/can-chatgpt-explain-code-813ff4c
a1ab0
External - https://betterprogramming.pub/
Internal - https://thomasdorfer.medium.com/accelerate-your-learning-with-ch
atgpt-d409f1e986f2
External - https://mlwhiz.medium.com/
External - https://odsc.medium.com/
External - https://towardsdatascience.medium.com/
External - https://anangsha.medium.com/
External - https://medium.com/@mary.newhauser
External - https://help.medium.com/hc/en-us
External - https://medium.statuspage.io/
External - https://about.medium.com/creators/
External - https://blog.medium.com/
External - https://medium.com/jobs-at-medium/work-at-medium-959d1a85284e
External - https://policy.medium.com/medium-privacy-policy-f03bf92035c9
External - https://policy.medium.com/medium-terms-of-service-9db0094a1e0f
External - https://medium.com/about
External - https://speechify.com/medium
```

External - https://unsplash.com/@pablogamedev
External - https://unsplash.com/photos/_SEbdtH4ZLM
External - https://unsplash.com/@hairspies
External - https://unsplash.com/photos/mXw0CfTPUrM
External - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6360409/
External - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9307347/
External - https://unsplash.com/@nci
External - https://unsplash.com/photos/BDKid0yJcAk
External - https://unsplash.com/@etiennemartin
External - https://unsplash.com/photos/2_K82gx9Uk8
External - https://medium.com/tag/machine-learning
External - https://medium.com/tag/modeling
External - https://medium.com/tag/occams-razor
External - https://itunes.apple.com/app/medium-everyones-stories/id828256236
External - https://play.google.com/store/apps/details
External - https://en.wikipedia.org/wiki/Immediately_invoked_function_expression
External - https://docs.python.org/3/library/functions.html
External - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html
External - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.agg.html
External - https://en.wikipedia.org/wiki/Closure_(computer_programming)
External - https://docs.python.org/3/tutorial/controlflow.html
External - https://realpython.com/python-lambda/
External - https://twitter.com/ThomasADorfer
External - https://www.linkedin.com/in/thomasdorfer/
Internal - https://thomasdorfer.medium.com/membership
External - https://medium.com/tag/python
External - https://medium.com/tag/programming
External - https://medium.com/tag/coding
External - https://medium.com/tag/software-development
External - https://www.pexels.com/@camilo-calderon-3343529/
External - https://www.pexels.com/video/a-video-footage-of-busy-street-4997787/
External - https://arxiv.org/abs/1506.02640
External - https://github.com/ultralytics/ultralytics
External - http://cocodataset.org/
External - https://github.com/ultralytics/ultralytics/blob/main/LICENSE
External - https://pypi.org/project/ultralytics/
External - https://unsplash.com/@jhc
External - https://unsplash.com/photos/jViepQKI01Q
External - https://docs.ultralytics.com/modes/predict/
External - https://pypi.org/project/opencv-python/
External - https://pypi.org/project/supervision/
External - https://docs.ultralytics.com/
External - https://www.youtube.com/watch
External - https://medium.com/@thomasdorfer/membership
External - https://medium.com/tag/computer-vision
External - https://medium.com/@bjarne.meyn
External - https://medium.com/@george.hajal
External - https://medium.com/@xxayani
External - https://medium.com/@milosalaveni
External - https://fabridigua.medium.com
External - https://medium.com/@karencampa
External - https://medium.com/@ganesh.gadsing
External - https://medium.com/@cbmtvsdn
External - https://unsplash.com/@ratushny
External - https://unsplash.com/photos/O33IVNPb0RI
External - https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-

Index-Report_Master.pdf
External - https://openai.com/dall-e-2/
External - https://en.wikipedia.org/wiki/GPT-3
External - https://openai.com/blog/chatgpt/
External - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8371605/
External - https://www.technologyreview.com/collection/the-download/
External - https://tldr.tech/
External - https://arxiv.org/help/subscribe
External - https://arxiv.org/category_taxonomy
External - https://scholar.google.com/scholar_alerts
External - https://twitter.com/ylecun
External - https://twitter.com/geoffreyhinton
External - https://twitter.com/RichardSSutton
External - https://twitter.com/AndrewYNg
External - https://twitter.com/chrmanning
External - https://www.ieee.org/publications/index.html
External - https://www.nature.com/natmachintell/
External - https://www.sciencedirect.com/journal/pattern-recognition
External - https://arxiv.org/
External - https://podcasts.apple.com/us/podcast/the-twiml-ai-podcast-forme
rly-this-week-in-machine/id1116303051
External - https://podcasts.apple.com/us/podcast/data-skeptic/id890348705
External - https://www.thetalkingmachines.com/home
External - https://podcasts.apple.com/us/podcast/practical-ai-machine-learn
ing-data-science/id1406537385
External - https://unsplash.com/@austindistel
External - https://unsplash.com/photos/Hg3BHX6U5jg
External - https://www.deeplearning.ai/the-batch/
External - https://alphasignal.ai/
External - https://thesequence.substack.com/
External - https://unsplash.com/@wocintechchat
External - https://unsplash.com/photos/faEfWCdOKIg
External - https://nips.cc/
External - https://icml.cc/
External - https://iclr.cc/
External - https://cvpr2022.thecvf.com/
External - https://www.2022.aclweb.org/
External - https://unsplash.com/@productschool
External - https://unsplash.com/photos/4jtHJX4SNk8
External - https://medium.com/tag/learning
External - https://medium.com/tag/research
External - https://medium.com/tag/growth
External - https://medium.com/tag/tips-and-tricks
External - https://pixabay.com/users/mohamed_hassan-5229782/
External - https://pixabay.com/illustrations/planning-finance-business-4077
086/
External - https://en.wikipedia.org/wiki/Chartjunk
Internal - https://thomasdorfer.medium.com/the-expressive-power-of-the-scat
ter-plot-c2f3354d3d97
External - https://archive.ics.uci.edu/ml/datasets/auto+mpg
External - https://www.oreilly.com/library/view/colorwise/9781492097839/
External - https://medium.com/u/fc065d3295b8
Internal - https://thomasdorfer.medium.com/the-case-against-the-pie-chart-4
3f4c3fccc6
External - https://www.theinspiration.com/2023/01/population-density-maps-b
y-terence-fosstodon/
External - https://www.color-blindness.com/coblis-color-blindness-simulator
/
External - https://www.highcharts.com/blog/tutorials/10-guidelines-for-data
viz-accessibility/
External - https://www.bbc.com/news/av/business-33464903

External - https://www.sciencedirect.com/science/article/pii/B978155860307350037X
External - https://medium.com/tag/data-visualization
External - https://medium.com/tag/business
External - https://medium.com/tag/data-analysis
External - https://medium.com/tag/charts
External - https://unsplash.com/@anniespratt
External - https://unsplash.com/@chuttersnap
External - https://unsplash.com/@disruptxn
External - https://unsplash.com/photos/IgUR1iX0mqM
External - https://unsplash.com/@dylan_nolte
External - https://unsplash.com/photos/NIrgENd0sAY
External - https://unsplash.com/@neonbrand
External - https://unsplash.com/photos/1-aA2Fadydc
External - https://unsplash.com/@monicomelty
External - https://unsplash.com/photos/oc_XTqWezp4
External - https://unsplash.com/@huntersrace
External - https://unsplash.com/photos/MYbhN8KaaEc
External - https://medium.com/tag/data-scientist
External - https://medium.com/tag/careers
External - https://medium.com/tag/tech
External - https://medium.com/tag/office-hours
External - https://docs.python.org/3/whatsnew/3.8.html
External - https://unsplash.com/@aindraus
External - https://unsplash.com/photos/Bb9jWuTMPUk
External - https://help.openai.com/en/articles/6783457-chatgpt-faq
External - https://medium.com/tag/chatgpt
External - https://medium.com/tag/algorithms
External - https://unsplash.com/@cdr6934
External - https://unsplash.com/photos/ieic5Tq8YMk
External - https://medium.com/plans
External - https://pixabay.com/vectors/error-warning-computer-crash-6641731/
External - https://www.python.org/downloads/release/python-3110/
External - https://docs.python.org/3/whatsnew/3.11.html
External - https://medium.com/tag/debugging

In [ ]:

```
import numpy as np
import pandas as pd
from queue import Queue
import requests
from bs4 import BeautifulSoup
q=Queue()
url = input("Enter URL")
q.put(url)
visited= set()
# visited.add(url)
urls=[]
# print(q.get())
while(q.empty()==False and len(urls)<50):
    m = q.qsize()
    for i in range(m):
        link = q.get()
        if link in visited:
            continue

        visited.add(link)
        print(link)
        urls.append(link)
#         print(urls[0])
        try:
            reqs = requests.get(link)
        except:
            continue
        html = reqs.content
        soup = BeautifulSoup(html, 'html.parser')

        for new_url in soup.find_all('a'):
            q.put(new_url.get('href'))

print(urls)
```

```
Enter URLhttps://thomasdorfer.medium.com/
https://thomasdorfer.medium.com/
https://rsci.app.link/?%24canonical_url=https%3A%2F%2Fmedium.com%2F&%7Efeat
ure=LoOpenInAppButton&%7Echannel=ShowUser&source=---two_column_layout_nav--
--------------------------------
https://medium.com/m/signin?operation=login&redirect=https%3A%2F%2Fthomasdo
rfer.medium.com%2F&source=user_profile---two_column_layout_nav-------------
----------global_nav-----------
https://medium.com/?source=---two_column_layout_nav-----------------------
----------
https://medium.com/m/signin?operation=register&redirect=https%3A%2F%2Fmediu
m.com%2Fnew-story&source=---two_column_layout_nav----------------------new
_post_sidenav-----------
https://medium.com/search?source=---two_column_layout_nav-----------------
----------------
/?source=user_profile---------------------------------------
/followers?source=user_profile------------------------------------
/about?source=user_profile----------------------------------------
https://towardsdatascience.com/?source=user_profile---------0-------------
--------------
/how-to-stay-on-top-of-the-latest-ai-research-e8993523ef3e?source=user_prof
ile---------0----------------------------
https://medium.com/tag/artificial-intelligence?source=user_profile---------
----------artificial_intelligence-----------------
```

https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2Fe8993523ef3e&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fhow-to-stay-on-top-of-the-latest-ai-research-e8993523ef3e&source=---------0-----------------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------1----------
/how-to-use-argument-parsing-for-greater-efficiency-in-machine-learning-workflows-2f637eaf5f6a?source=user_profile---------1----------
https://medium.com/tag/machine-learning?source=user_profile----------machine_learning----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2F2f637eaf5f6a&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fhow-to-use-argument-parsing-for-greater-efficiency-in-machine-learning-workflows-2f637eaf5f6a&source=---------1----------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------2----------
/enhanced-object-detection-how-to-effectively-implement-yolov8-afd1bf6132ae?source=user_profile---------2----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2Fafd1bf6132ae&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fenhanced-object-detection-how-to-effectively-implement-yolov8-afd1bf6132ae&source=---------2----------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------3----------
/comparing-list-comprehensions-vs-built-in-functions-in-python-which-is-better-1e2c9646fafe?source=user_profile---------3----------
https://medium.com/tag/data-science?source=user_profile----------data_science----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2F1e2c9646fafe&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fcomparing-list-comprehensions-vs-built-in-functions-in-python-which-is-better-1e2c9646fafe&source=---------3----------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------4----------
/enhanced-debugging-in-python-tracebacks-just-got-a-major-upgrade-bd77fb32db38?source=user_profile---------4----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2Fbd77fb32db38&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fenhanced-debugging-in-python-tracebacks-just-got-a-major-upgrade-bd77fb32db38&source=---------4----------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------5----------
/effective-data-visualization-9-valuable-tips-to-increase-the-quality-of-your-charts-5fec31144a6d?source=user_profile---------5----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2F5fec31144a6d&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Feffective-data-visualization-9-valuable-tips-to-increase-the-quality-of-your-charts-5fec31144a6d&source=---------5----------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------6----------
/how-to-effectively-use-lambda-functions-in-python-as-a-data-scientist-fd6171554053?source=user_profile---------6----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2Ffd6171554053&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fhow-to-effectively-use-lambda-functions-in-python-as-a-data-s

cientist-fd6171554053&source=---------6-----------------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------7----------
/data-science-in-small-and-big-companies-5cb32be1491a?source=user_profile---------7----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2F5cb32be1491a&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fdata-science-in-small-and-big-companies-5cb32be1491a&source=---------7-----------------bookmark_preview----------
https://towardsdatascience.com/?source=user_profile---------8----------
/why-simple-models-are-often-better-e2428964811a?source=user_profile---------8----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2Fe2428964811a&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fwhy-simple-models-are-often-better-e2428964811a&source=---------8-----------------bookmark_preview----------
/can-chatgpt-explain-code-813ff4ca1ab0?source=user_profile---------9----------
https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fbookmark%2Fp%2F813ff4ca1ab0&operation=register&redirect=https%3A%2F%2Fthomasdorfer.medium.com%2Fcan-chatgpt-explain-code-813ff4ca1ab0&source=---------9----------bookmark_preview----------
/?source=---two_column_layout_sidebar----------
/followers?source=---two_column_layout_sidebar----------
https://medium.com/m/signin?actionUrl=%2F_%2Fapi%2Fsubscriptions%2Fnewsletters%2Fdb1dd30387ee&operation=register&redirect=https%3A%2F%2Fthomasdorfer.medium.com%2F&newsletterV3=7c54f9b62b90&newsletterV3Id=db1dd30387ee&user=Thomas+A+Dorfer&userId=7c54f9b62b90&source=---two_column_layout_sidebar----------subscribe_user----------
https://kozyrkov.medium.com/?source=blogrolls_sidebar---two_column_layout_sidebar----------
https://medium.com/@radecicdario?source=blogrolls_sidebar---two_column_layout_sidebar----------
https://odsc.medium.com/?source=blogrolls_sidebar---two_column_layout_sidebar----------
https://medium.com/@kenneth.b.jee?source=blogrolls_sidebar---two_column_layout_sidebar----------
https://gmyrianthous.medium.com/?source=blogrolls_sidebar---two_column_layout_sidebar----------
/following?source=blogrolls_sidebar---two_column_layout_sidebar----------
https://help.medium.com/hc/en-us?source=---two_column_layout_sidebar----------
https://medium.statuspage.io/?source=---two_column_layout_sidebar----------
https://about.medium.com/creators/?source=---two_column_layout_sidebar----------
https://blog.medium.com/?source=---two_column_layout_sidebar----------
https://medium.com/jobs-at-medium/work-at-medium-959d1a85284e?source=---two_column_layout_sidebar----------
https://policy.medium.com/medium-privacy-policy-f03bf92035c9?source=---two_column_layout_sidebar----------
https://policy.medium.com/medium-terms-of-service-9db0094a1e0f?source=---two_column_layout_sidebar----------

In [ ]:

```
import requests
from bs4 import BeautifulSoup

# zenrows_api_base = "https://api.zenrows.com/v1/?apikey=YOUR_KEY"
seed_url = ""

def extract_links(soup):
    return [a.get("href")
            for a in soup.find_all("a")
            if a.get("href") and a.get("href").startswith("/")]

def call_url(url):
    response = requests.get(zenrows_api_base, params={"url": url})
    soup = BeautifulSoup(response.text, "html.parser")
    links = extract_links(url, soup)
    print(links)

call_url(seed_url)
```