11/02/22

Expeximent-4

Implementation and analysis of DFS and BFS for application

Aim: Web crawling using DFS and BFS

Algorithm: 1) Define 2 empty sets to store internal and external links separately.

2) Define set url to temporarily store urls

3) Create beautifulsoup object and extract all anchor tags.

4) Get href tags from anchor tags.

5) Create absolute url using urljoin

Result: Web crawling using DFS and BFS has been successfully executed.

Experiment 4 - BFS and DFS for web crawling

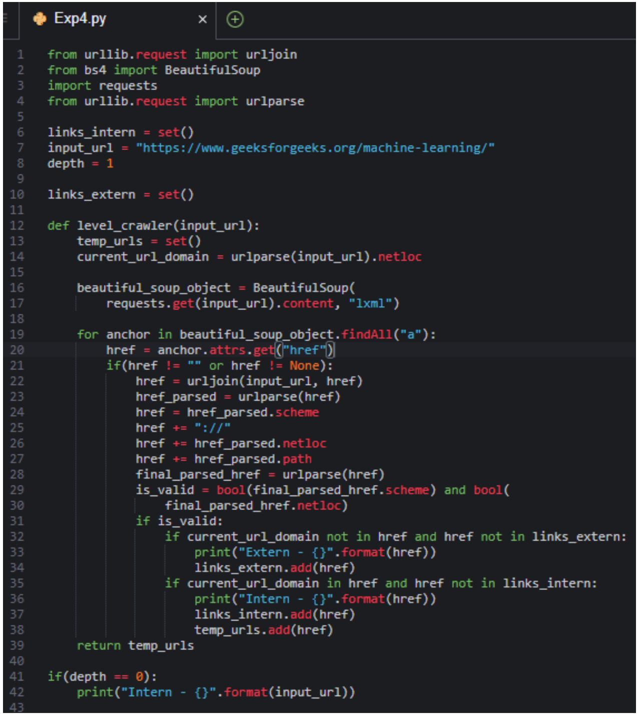
Dhawal Patil
RA1911003010575
CSE A2

Code (BFS)

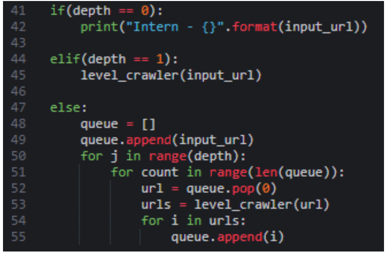
```
from urllib.request import urljoin
from bs4 import BeautifulSoup
import requests
from urllib.request import urlparse

links_intern = set()
input_url = "https://www.geeksforgeeks.org/machine-learning/"
depth = 1

links_extern = set()

def level_crawler(input_url):
        temp_urls = set()
        current_url_domain = urlparse(input_url).netloc

        beautiful_soup_object = BeautifulSoup(
                requests.get(input_url).content, "lxml")

        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_href.netloc)
                        if is_valid:
                                if current_url_domain not in href and href not in links_extern:
                                        print("Extern - {}".format(href))
                                        links_extern.add(href)
                                if current_url_domain in href and href not in links_intern:
                                        print("Intern - {}".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)

else:
        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)
```

Screenshot

Code (BFS)

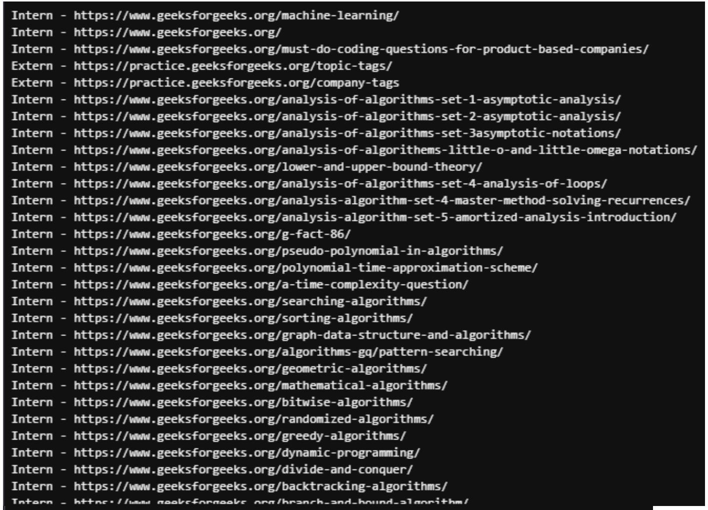


```python
from urllib.request import urljoin
from bs4 import BeautifulSoup
import requests
from urllib.request import urlparse

links_intern = set()
input_url = "https://www.geeksforgeeks.org/machine-learning/"
depth = 1

links_extern = set()

def level_crawler(input_url):
    temp_urls = set()
    current_url_domain = urlparse(input_url).netloc

    beautiful_soup_object = BeautifulSoup(
        requests.get(input_url).content, "lxml")

    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_href.netloc)
            if is_valid:
                if current_url_domain not in href and href not in links_extern:
                    print("Extern - {}".format(href))
                    links_extern.add(href)
                if current_url_domain in href and href not in links_intern:
                    print("Intern - {}".format(href))
                    links_intern.add(href)
                    temp_urls.add(href)
    return temp_urls

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

```

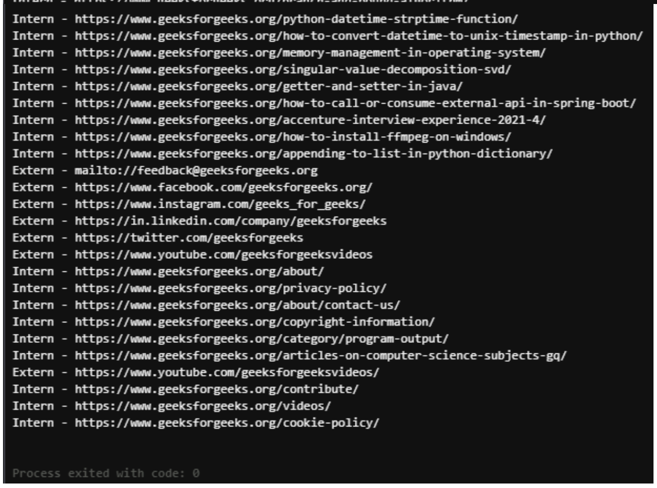


```python
if(depth == 0):
    print("Intern - {}".format(input_url))

elif(depth == 1):
    level_crawler(input_url)

else:
    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)
```

Output (BFS)



```
Intern - https://www.geeksforgeeks.org/machine-learning/
Intern - https://www.geeksforgeeks.org/
Intern - https://www.geeksforgeeks.org/must-do-coding-questions-for-product-based-companies/
Extern - https://practice.geeksforgeeks.org/topic-tags/
Extern - https://practice.geeksforgeeks.org/company-tags
Intern - https://www.geeksforgeeks.org/analysis-of-algorithms-set-1-asymptotic-analysis/
Intern - https://www.geeksforgeeks.org/analysis-of-algorithms-set-2-asymptotic-analysis/
Intern - https://www.geeksforgeeks.org/analysis-of-algorithms-set-3asymptotic-notations/
Intern - https://www.geeksforgeeks.org/analysis-of-algorithems-little-o-and-little-omega-notations/
Intern - https://www.geeksforgeeks.org/lower-and-upper-bound-theory/
Intern - https://www.geeksforgeeks.org/analysis-of-algorithms-set-4-analysis-of-loops/
Intern - https://www.geeksforgeeks.org/analysis-algorithm-set-4-master-method-solving-recurrences/
Intern - https://www.geeksforgeeks.org/analysis-algorithm-set-5-amortized-analysis-introduction/
Intern - https://www.geeksforgeeks.org/g-fact-86/
Intern - https://www.geeksforgeeks.org/pseudo-polynomial-in-algorithms/
Intern - https://www.geeksforgeeks.org/polynomial-time-approximation-scheme/
Intern - https://www.geeksforgeeks.org/a-time-complexity-question/
Intern - https://www.geeksforgeeks.org/searching-algorithms/
Intern - https://www.geeksforgeeks.org/sorting-algorithms/
Intern - https://www.geeksforgeeks.org/graph-data-structure-and-algorithms/
Intern - https://www.geeksforgeeks.org/algorithms-gq/pattern-searching/
Intern - https://www.geeksforgeeks.org/geometric-algorithms/
Intern - https://www.geeksforgeeks.org/mathematical-algorithms/
Intern - https://www.geeksforgeeks.org/bitwise-algorithms/
Intern - https://www.geeksforgeeks.org/randomized-algorithms/
Intern - https://www.geeksforgeeks.org/greedy-algorithms/
Intern - https://www.geeksforgeeks.org/dynamic-programming/
Intern - https://www.geeksforgeeks.org/divide-and-conquer/
Intern - https://www.geeksforgeeks.org/backtracking-algorithms/
Intern - https://www.geeksforgeeks.org/branch-and-bound-algorithm/
```



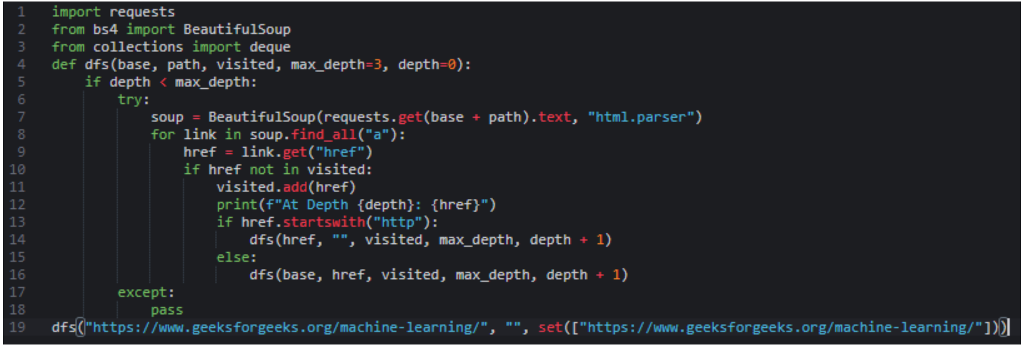
```
Intern - https://www.geeksforgeeks.org/python-datetime-strptime-function/
Intern - https://www.geeksforgeeks.org/how-to-convert-datetime-to-unix-timestamp-in-python/
Intern - https://www.geeksforgeeks.org/memory-management-in-operating-system/
Intern - https://www.geeksforgeeks.org/singular-value-decomposition-svd/
Intern - https://www.geeksforgeeks.org/getter-and-setter-in-java/
Intern - https://www.geeksforgeeks.org/how-to-call-or-consume-external-api-in-spring-boot/
Intern - https://www.geeksforgeeks.org/accenture-interview-experience-2021-4/
Intern - https://www.geeksforgeeks.org/how-to-install-ffmpeg-on-windows/
Intern - https://www.geeksforgeeks.org/appending-to-list-in-python-dictionary/
Extern - mailto://feedback@geeksforgeeks.org
Extern - https://www.facebook.com/geeksforgeeks.org/
Extern - https://www.instagram.com/geeks_for_geeks/
Extern - https://in.linkedin.com/company/geeksforgeeks
Extern - https://twitter.com/geeksforgeeks
Extern - https://www.youtube.com/geeksforgeeksvideos
Intern - https://www.geeksforgeeks.org/about/
Intern - https://www.geeksforgeeks.org/privacy-policy/
Intern - https://www.geeksforgeeks.org/about/contact-us/
Intern - https://www.geeksforgeeks.org/copyright-information/
Intern - https://www.geeksforgeeks.org/category/program-output/
Intern - https://www.geeksforgeeks.org/articles-on-computer-science-subjects-gq/
Extern - https://www.youtube.com/geeksforgeeksvideos/
Intern - https://www.geeksforgeeks.org/contribute/
Intern - https://www.geeksforgeeks.org/videos/
Intern - https://www.geeksforgeeks.org/cookie-policy/


Process exited with code: 0
```

Code (DFS)

```
import requests
from bs4 import BeautifulSoup
from collections import deque
def dfs(base, path, visited, max_depth=3, depth=0):
    if depth < max_depth:
        try:
            soup = BeautifulSoup(requests.get(base + path).text, "html.parser")
            for link in soup.find_all("a"):
                href = link.get("href")
                if href not in visited:
                    visited.add(href)
                    print(f"At Depth {depth}: {href}")
                    if href.startswith("http"):
                        dfs(href, "", visited, max_depth, depth + 1)
                    else:
                        dfs(base, href, visited, max_depth, depth + 1)
        except:
            pass
dfs("https://www.geeksforgeeks.org/machine-learning/", "", set(["https://www.geeksforgeeks.org/machine-learning/"]))
```

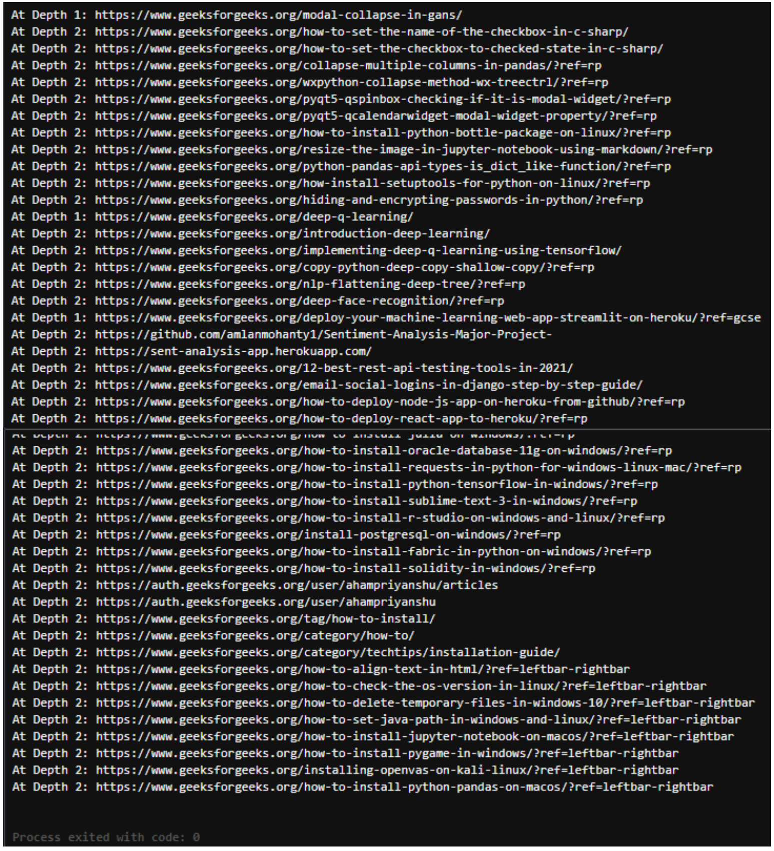
Screenshot

Code (DFS)



```
import requests
from bs4 import BeautifulSoup
from collections import deque
def dfs(base, path, visited, max_depth=3, depth=0):
    if depth < max_depth:
        try:
            soup = BeautifulSoup(requests.get(base + path).text, "html.parser")
            for link in soup.find_all("a"):
                href = link.get("href")
                if href not in visited:
                    visited.add(href)
                    print(f"At Depth {depth}: {href}")
                    if href.startswith("http"):
                        dfs(href, "", visited, max_depth, depth + 1)
                    else:
                        dfs(base, href, visited, max_depth, depth + 1)
        except:
            pass
dfs("https://www.geeksforgeeks.org/machine-learning/", "", set(["https://www.geeksforgeeks.org/machine-learning/"]))
```

Output (DFS)



```
At Depth 1: https://www.geeksforgeeks.org/modal-collapse-in-gans/
At Depth 2: https://www.geeksforgeeks.org/how-to-set-the-name-of-the-checkbox-in-c-sharp/
At Depth 2: https://www.geeksforgeeks.org/how-to-set-the-checkbox-to-checked-state-in-c-sharp/
At Depth 2: https://www.geeksforgeeks.org/collapse-multiple-columns-in-pandas/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/wxpython-collapse-method-wx-treectrl/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/pyqt5-qspinbox-checking-if-it-is-modal-widget/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/pyqt5-qcalendarwidget-modal-widget-property/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-python-bottle-package-on-linux/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/resize-the-image-in-jupyter-notebook-using-markdown/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/python-pandas-api-types-is_dict_like-function/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-install-setuptools-for-python-on-linux/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/hiding-and-encrypting-passwords-in-python/?ref=rp
At Depth 1: https://www.geeksforgeeks.org/deep-q-learning/
At Depth 2: https://www.geeksforgeeks.org/introduction-deep-learning/
At Depth 2: https://www.geeksforgeeks.org/implementing-deep-q-learning-using-tensorflow/
At Depth 2: https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/nlp-flattening-deep-tree/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/deep-face-recognition/?ref=rp
At Depth 1: https://www.geeksforgeeks.org/deploy-your-machine-learning-web-app-streamlit-on-heroku/?ref=gcse
At Depth 2: https://github.com/amlanmohanty1/Sentiment-Analysis-Major-Project-
At Depth 2: https://sent-analysis-app.herokuapp.com/
At Depth 2: https://www.geeksforgeeks.org/12-best-rest-api-testing-tools-in-2021/
At Depth 2: https://www.geeksforgeeks.org/email-social-logins-in-django-step-by-step-guide/
At Depth 2: https://www.geeksforgeeks.org/how-to-deploy-node-js-app-on-heroku-from-github/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-deploy-react-app-to-heroku/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-julia-on-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-oracle-database-11g-on-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-requests-in-python-for-windows-linux-mac/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-python-tensorflow-in-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-sublime-text-3-in-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-r-studio-on-windows-and-linux/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/install-postgresql-on-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-fabric-in-python-on-windows/?ref=rp
At Depth 2: https://www.geeksforgeeks.org/how-to-install-solidity-in-windows/?ref=rp
At Depth 2: https://auth.geeksforgeeks.org/user/ahampriyanshu/articles
At Depth 2: https://auth.geeksforgeeks.org/user/ahampriyanshu
At Depth 2: https://www.geeksforgeeks.org/tag/how-to-install/
At Depth 2: https://www.geeksforgeeks.org/category/how-to/
At Depth 2: https://www.geeksforgeeks.org/category/techtips/installation-guide/
At Depth 2: https://www.geeksforgeeks.org/how-to-align-text-in-html/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-check-the-os-version-in-linux/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-delete-temporary-files-in-windows-10/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-set-java-path-in-windows-and-linux/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-install-jupyter-notebook-on-macos/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-install-pygame-in-windows/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/installing-openvas-on-kali-linux/?ref=leftbar-rightbar
At Depth 2: https://www.geeksforgeeks.org/how-to-install-python-pandas-on-macos/?ref=leftbar-rightbar


Process exited with code: 0
```