Web mining LAB Assignment -1 [Prof.Ranganthan Sridhar]

With the help of ‘Requests’ module to download and analyse page contents of ‘https://www.vit.ac.in/academics/home ‘or ‘https://www.vit.ac.in/admissions/overview’ or any other page in these menu tabs.

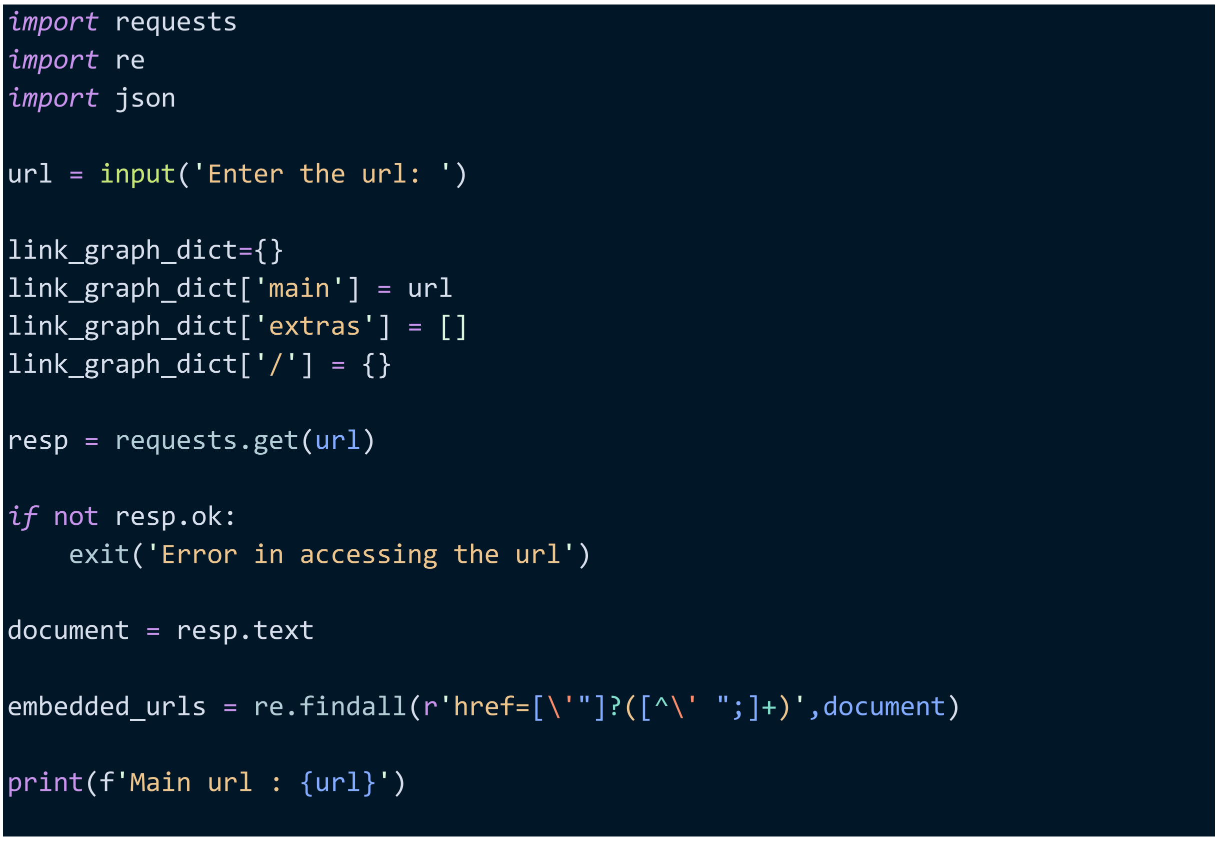
Find out all the URLs which are specified as links in these pages and draw the Link connection tree or Graph depicting the connections between pages. This should be done only with ‘requests module’ and pure python programming. For tree drawing you can use other drawing packages if needed. Instead of tree drawing you can also output a table indicating the connections.

Name: Gerosh George

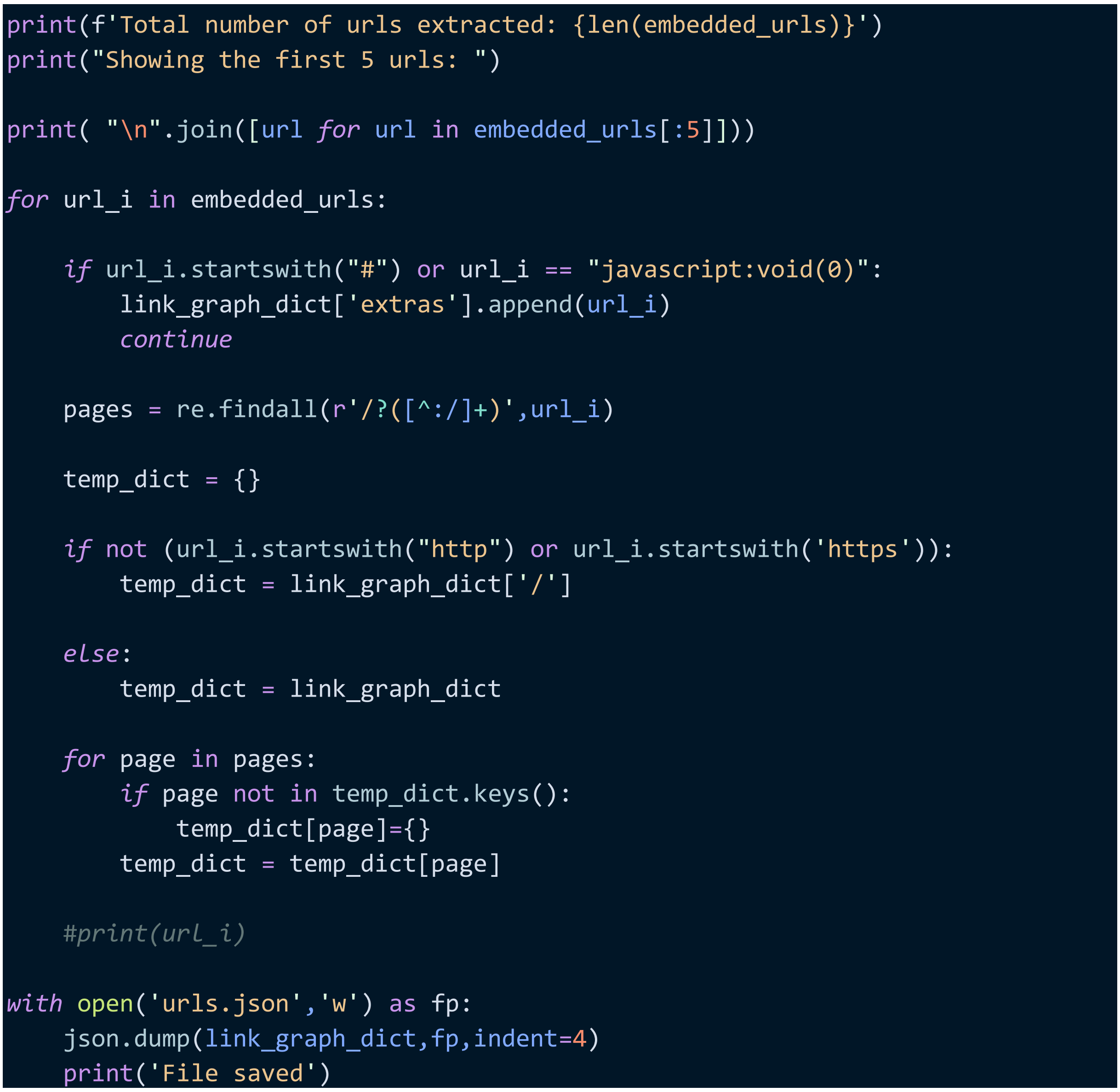
Reg No: 19BCE1403

1) Extracted all the embedded URLs from the web page and categorised them for easy understanding in a JSON file

CODE:

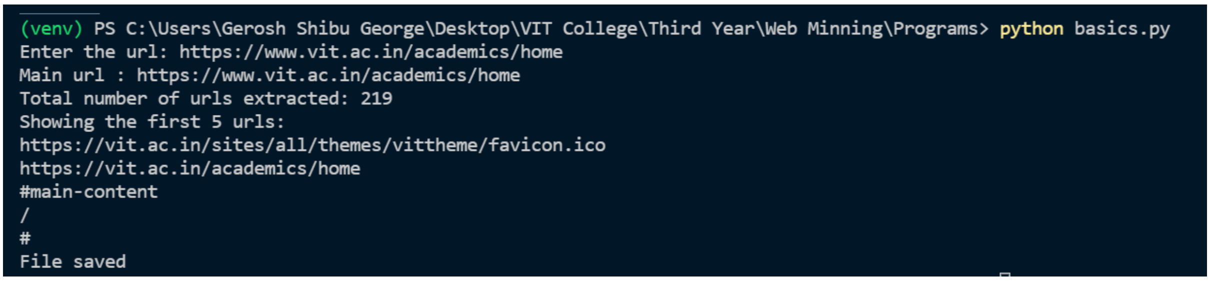


```
import requests
import re
import json

url = input('Enter the url: ')

link_graph_dict={}
link_graph_dict['main'] = url
link_graph_dict['extras'] = []
link_graph_dict['/'] = {}

resp = requests.get(url)

if not resp.ok:
    exit('Error in accessing the url')

document = resp.text

embedded_urls = re.findall(r'href=[\'"]?([^\' ";]+)',document)

print(f'Main url : {url}')
```



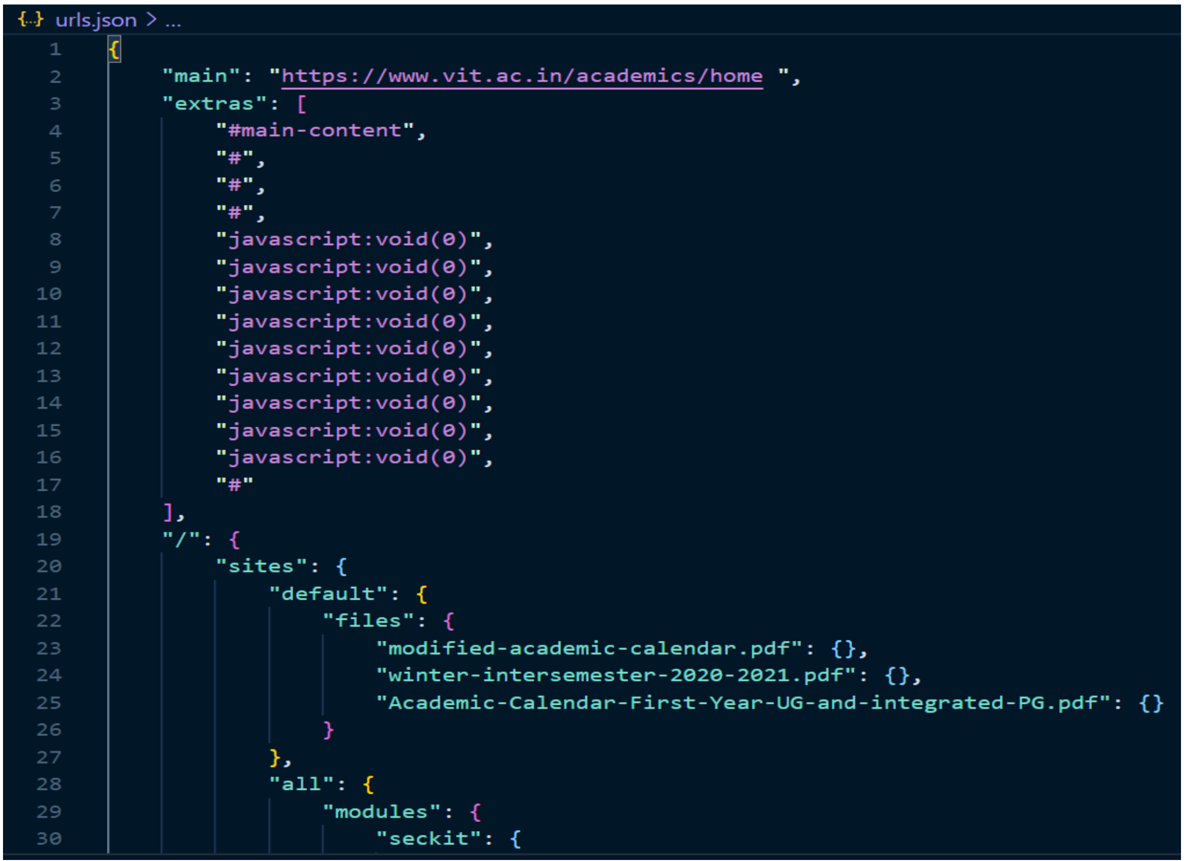
```python
print(f'Total number of urls extracted: {len(embedded_urls)}')
print("Showing the first 5 urls: ")

print( "\n".join([url for url in embedded_urls[:5]]))

for url_i in embedded_urls:

    if url_i.startswith("#") or url_i == "javascript:void(0)":
        link_graph_dict['extras'].append(url_i)
        continue

    pages = re.findall(r'/?([^:/]+)',url_i)

    temp_dict = {}

    if not (url_i.startswith("http") or url_i.startswith('https')):
        temp_dict = link_graph_dict['/']

    else:
        temp_dict = link_graph_dict

    for page in pages:
        if page not in temp_dict.keys():
            temp_dict[page]={}
        temp_dict = temp_dict[page]

    #print(url_i)

with open('urls.json','w') as fp:
    json.dump(link_graph_dict,fp,indent=4)
    print('File saved')
```

OUTPUT:

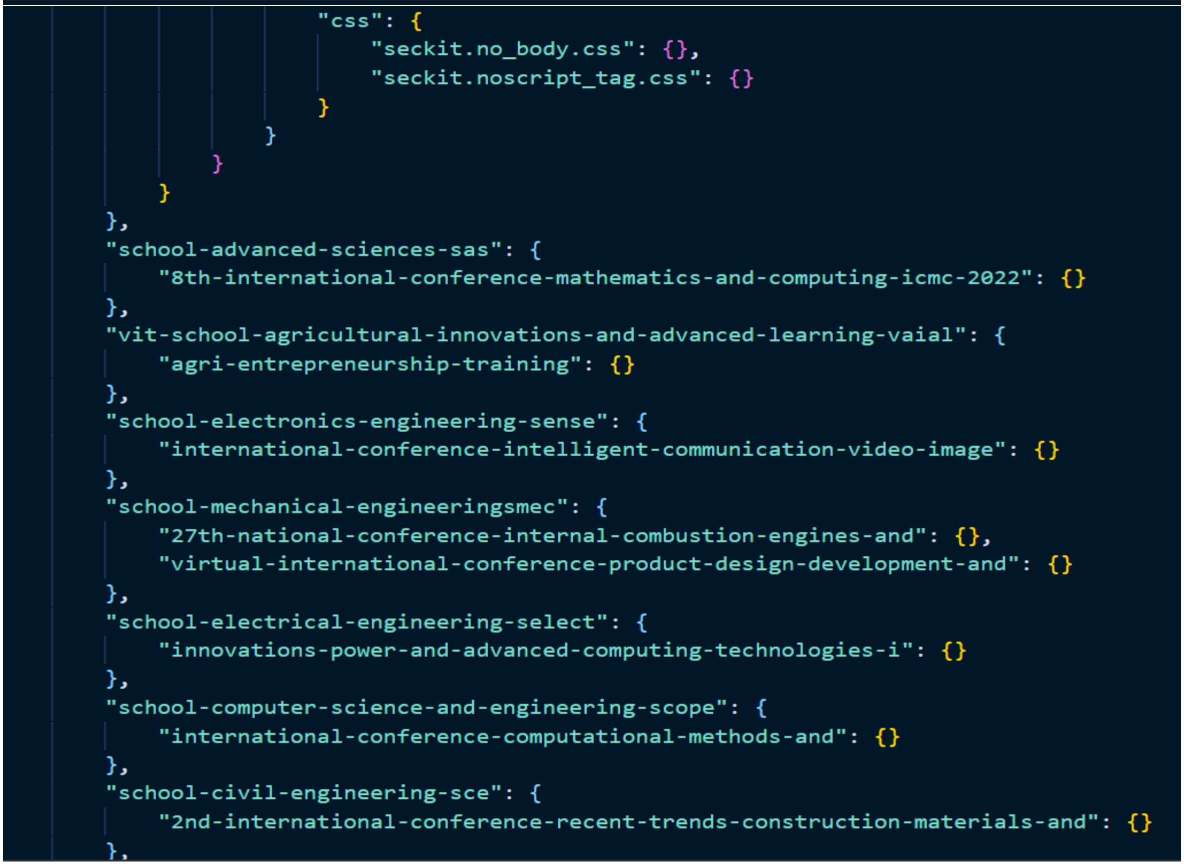


```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Minning\Programs> python basics.py
Enter the url: https://www.vit.ac.in/academics/home
Main url : https://www.vit.ac.in/academics/home
Total number of urls extracted: 219
Showing the first 5 urls:
https://vit.ac.in/sites/all/themes/vittheme/favicon.ico
https://vit.ac.in/academics/home
#main-content
/
#
File saved
```

JSON File which was created:



```
{
    "main": "https://www.vit.ac.in/academics/home ",
    "extras": [
        "#main-content",
        "#",
        "#",
        "#",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "javascript:void(0)",
        "#"
    ],
    "/": {
        "sites": {
            "default": {
                "files": {
                    "modified-academic-calendar.pdf": {},
                    "winter-intersemester-2020-2021.pdf": {},
                    "Academic-Calendar-First-Year-UG-and-integrated-PG.pdf": {}
                }
            },
            "all": {
                "modules": {
                    "seckit": {
```

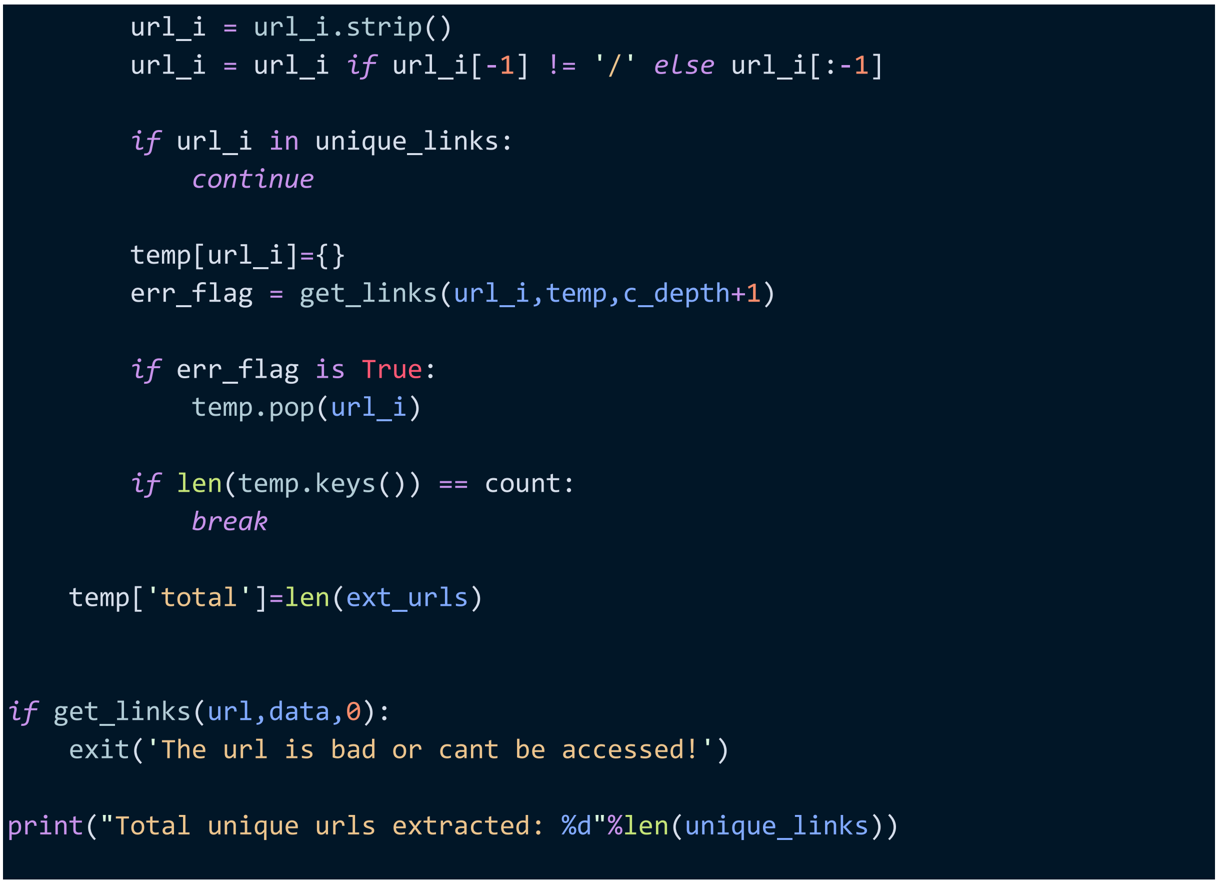


```
                        "css": {
                            "seckit.no_body.css": {},
                            "seckit.noscript_tag.css": {}
                        }
                    }
                }
            }
        },
        "school-advanced-sciences-sas": {
            "8th-international-conference-mathematics-and-computing-icmc-2022": {}
        },
        "vit-school-agricultural-innovations-and-advanced-learning-vaial": {
            "agri-entrepreneurship-training": {}
        },
        "school-electronics-engineering-sense": {
            "international-conference-intelligent-communication-video-image": {}
        },
        "school-mechanical-engineeringsmec": {
            "27th-national-conference-internal-combustion-engines-and": {},
            "virtual-international-conference-product-design-development-and": {}
        },
        "school-electrical-engineering-select": {
            "innovations-power-and-advanced-computing-technologies-i": {}
        },
        "school-computer-science-and-engineering-scope": {
            "international-conference-computational-methods-and": {}
        },
        "school-civil-engineering-sce": {
            "2nd-international-conference-recent-trends-construction-materials-and": {}
        },
```

2) Very naïve crawler that extracts the mentioned number of URLs from a webpage and go till nth depth which user can mention.

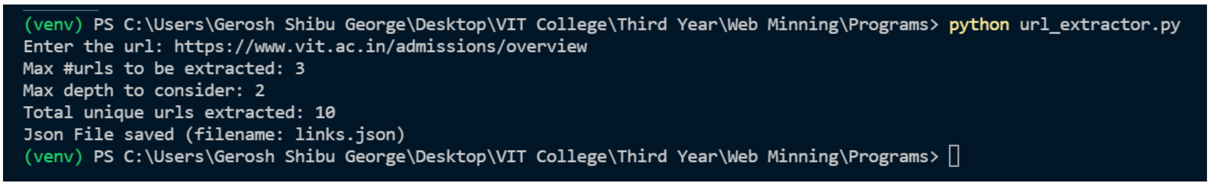
CODE:

```
import requests
import re
import json


url = input('Enter the url: ')
url = url if url[-1] != '/' else url[:-1]

count = int(input('Max #urls to be extracted: '))
depth = int(input('Max depth to consider: '))

unique_links=set()

data={
    url:{}
}

def get_links(url,url_dict,c_depth):

    if url in unique_links:
        return

    unique_links.add(url)

    if depth == c_depth:
        return

    temp=url_dict[url]

    resp=requests.get(url)

    if not resp.ok:
        print("Error URL: ",url)
        print("Status code: ",resp.status_code)
        return True

    document = resp.text

    ext_urls=re.findall(r'href=[\'"](https://[^\'"]+|http://[^\'"]+)',document)

    for url_i in ext_urls:
```



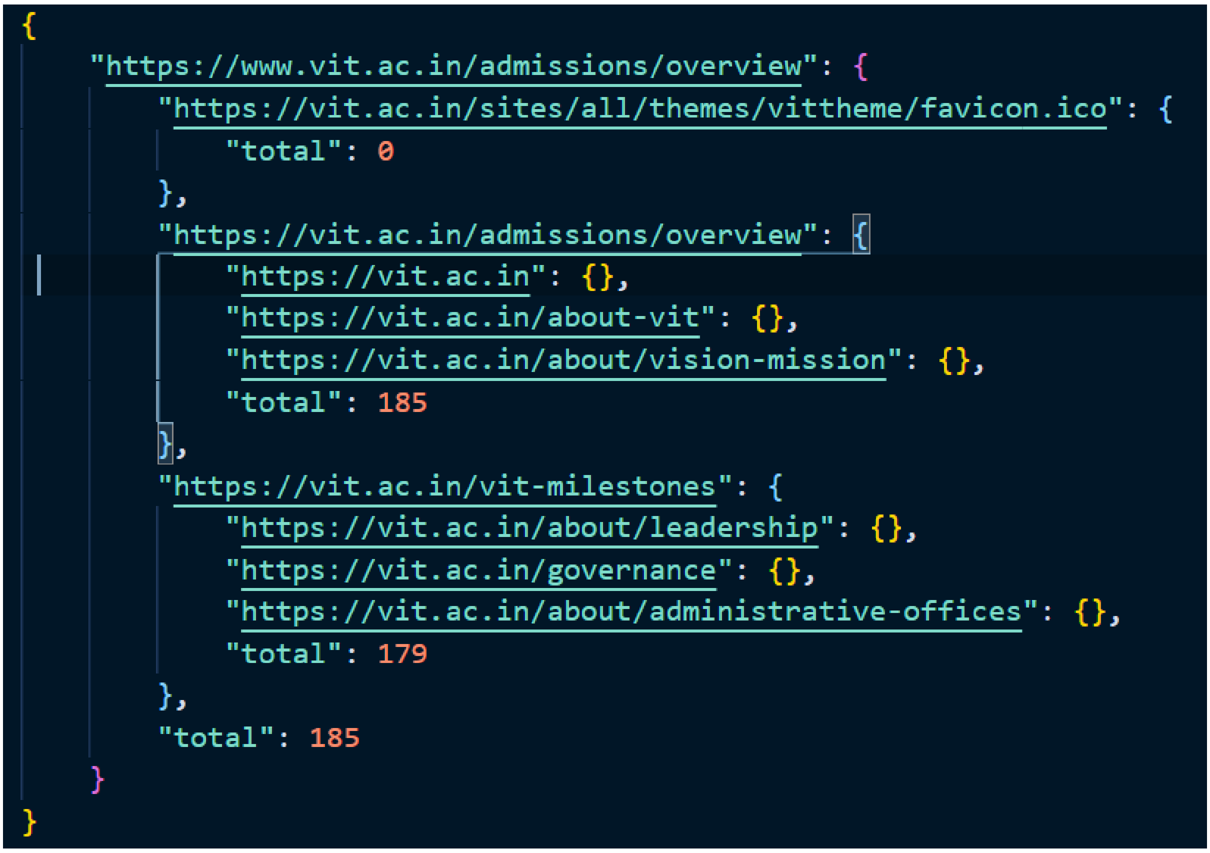
```
        url_i = url_i.strip()
        url_i = url_i if url_i[-1] != '/' else url_i[:-1]

        if url_i in unique_links:
            continue

        temp[url_i]={}
        err_flag = get_links(url_i,temp,c_depth+1)

        if err_flag is True:
            temp.pop(url_i)

        if len(temp.keys()) == count:
            break

    temp['total']=len(ext_urls)


if get_links(url,data,0):
    exit('The url is bad or cant be accessed!')

print("Total unique urls extracted: %d"%len(unique_links))
```

OUTPUT:

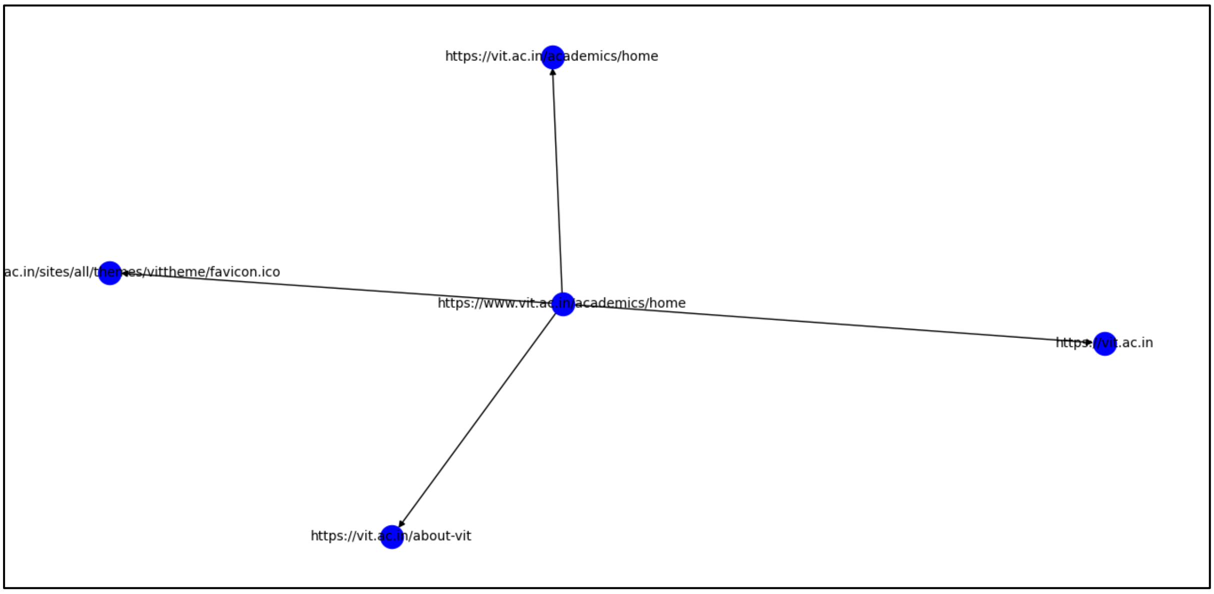


```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Minning\Programs> python url_extractor.py
Enter the url: https://www.vit.ac.in/admissions/overview
Max #urls to be extracted: 3
Max depth to consider: 2
Total unique urls extracted: 10
Json File saved (filename: links.json)
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Minning\Programs> 
```

JSON FILE:



```
{
    "https://www.vit.ac.in/admissions/overview": {
        "https://vit.ac.in/sites/all/themes/vittheme/favicon.ico": {
            "total": 0
        },
        "https://vit.ac.in/admissions/overview": {
            "https://vit.ac.in": {},
            "https://vit.ac.in/about-vit": {},
            "https://vit.ac.in/about/vision-mission": {},
            "total": 185
        },
        "https://vit.ac.in/vit-milestones": {
            "https://vit.ac.in/about/leadership": {},
            "https://vit.ac.in/governance": {},
            "https://vit.ac.in/about/administrative-offices": {},
            "total": 179
        },
        "total": 185
    }
}
```


https://vit.ac.in/academics/home
ac.in/sites/all/themes/vittheme/favicon.ico
https://www.vit.ac.in/academics/home
https://vit.ac.in
https://vit.ac.in/about-vit

Link Connection Graph of main URL (https://www.vit.ac.in/academics/home) and only showing the first level to avoid complexity (i.e., showing the first four embedded links that were extracted).