WEB MINING LAB

Faculty: Dr.Sridhar.R

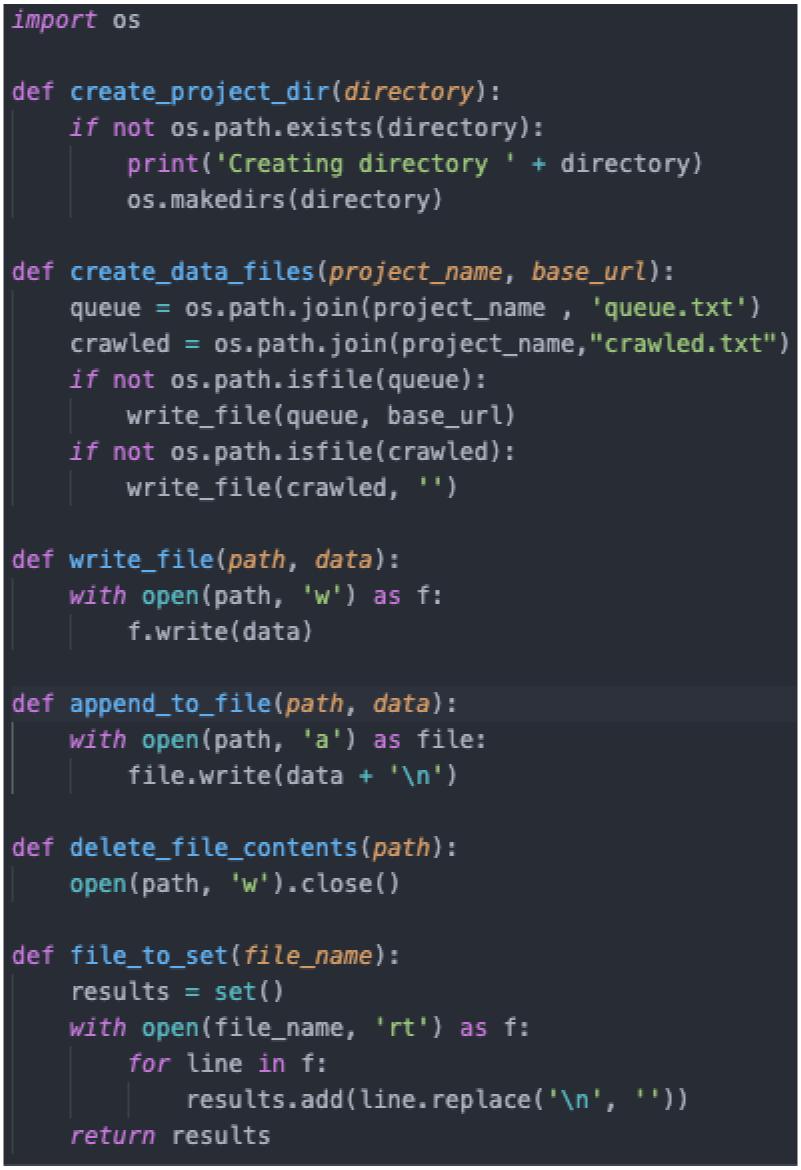
LAB4
DATE: 26TH AUG 2021

VIT CHENNAI

Aim: to design a basic web crawler

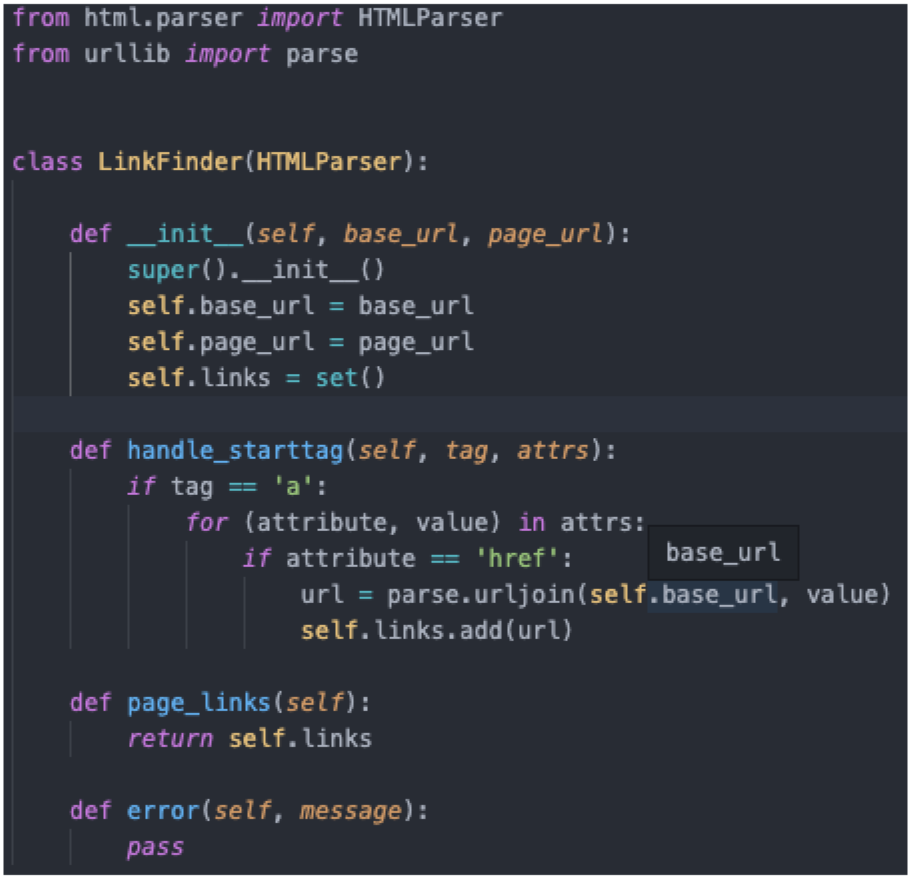
Code:

Helper functions (helper.py)



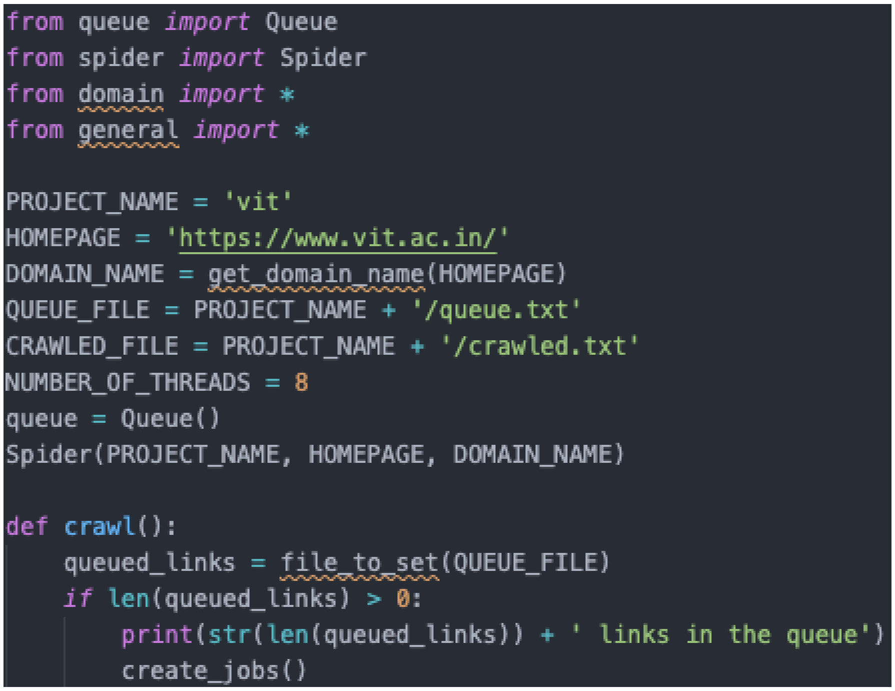
```python
import os

def create_project_dir(directory):
    if not os.path.exists(directory):
        print('Creating directory ' + directory)
        os.makedirs(directory)

def create_data_files(project_name, base_url):
    queue = os.path.join(project_name , 'queue.txt')
    crawled = os.path.join(project_name,"crawled.txt")
    if not os.path.isfile(queue):
        write_file(queue, base_url)
    if not os.path.isfile(crawled):
        write_file(crawled, '')

def write_file(path, data):
    with open(path, 'w') as f:
        f.write(data)

def append_to_file(path, data):
    with open(path, 'a') as file:
        file.write(data + '\n')

def delete_file_contents(path):
    open(path, 'w').close()

def file_to_set(file_name):
    results = set()
    with open(file_name, 'rt') as f:
        for line in f:
            results.add(line.replace('\n', ''))
    return results
```

Finding all the links:(link_find.py)



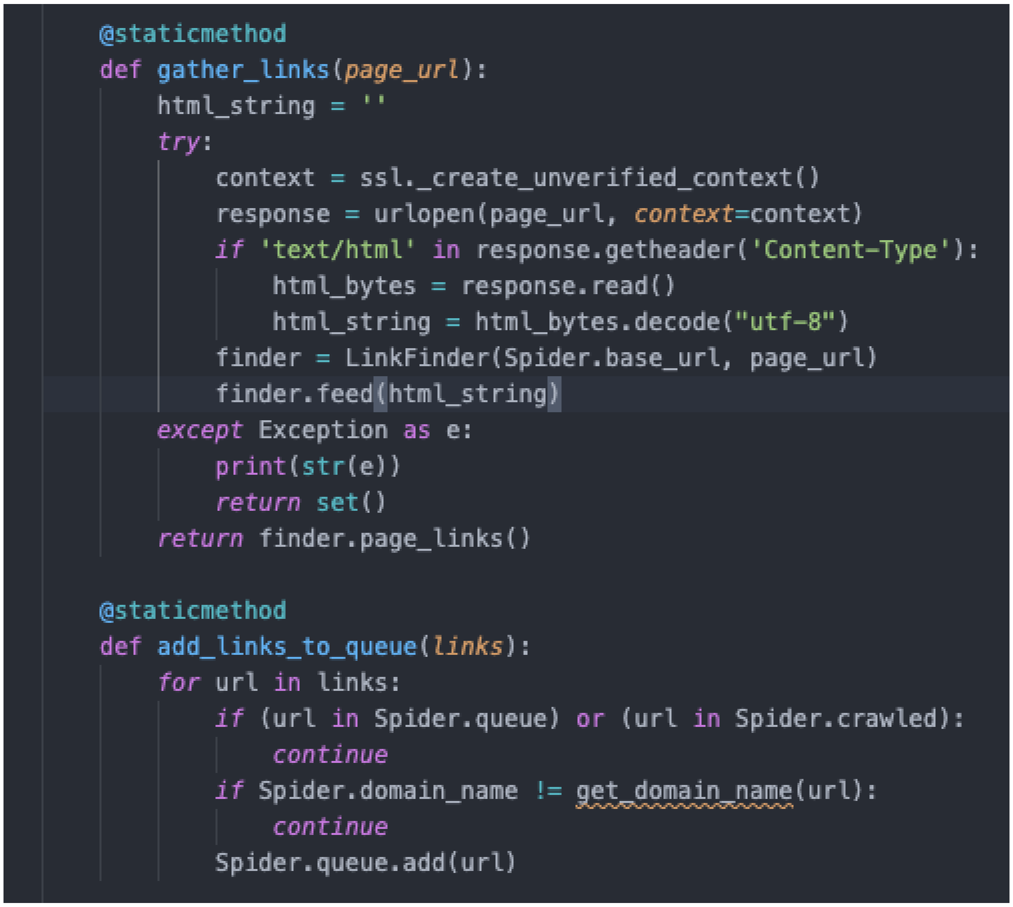
```python
from html.parser import HTMLParser
from urllib import parse


class LinkFinder(HTMLParser):

    def __init__(self, base_url, page_url):
        super().__init__()
        self.base_url = base_url
        self.page_url = page_url
        self.links = set()

    def handle_starttag(self, tag, attrs):
        if tag == 'a':
            for (attribute, value) in attrs:
                if attribute == 'href':
                    url = parse.urljoin(self.base_url, value)
                    self.links.add(url)

    def page_links(self):
        return self.links

    def error(self, message):
        pass
```

Crawler (main.py)



```python
from queue import Queue
from spider import Spider
from domain import *
from general import *

PROJECT_NAME = 'vit'
HOMEPAGE = 'https://www.vit.ac.in/'
DOMAIN_NAME = get_domain_name(HOMEPAGE)
QUEUE_FILE = PROJECT_NAME + '/queue.txt'
CRAWLED_FILE = PROJECT_NAME + '/crawled.txt'
NUMBER_OF_THREADS = 8
queue = Queue()
Spider(PROJECT_NAME, HOMEPAGE, DOMAIN_NAME)

def crawl():
    queued_links = file_to_set(QUEUE_FILE)
    if len(queued_links) > 0:
        print(str(len(queued_links)) + ' links in the queue')
        create_jobs()
```

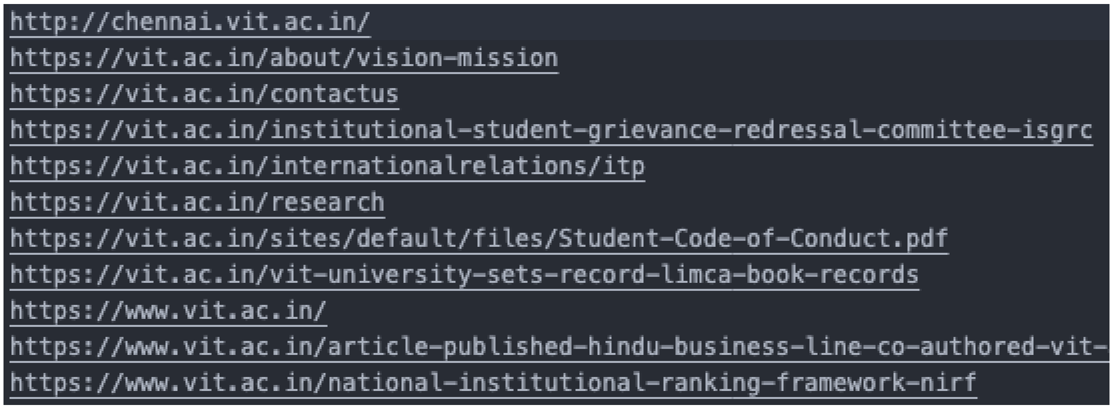
Setting links in the queue (spider.py)



```
@staticmethod
def gather_links(page_url):
    html_string = ''
    try:
        context = ssl._create_unverified_context()
        response = urlopen(page_url, context=context)
        if 'text/html' in response.getheader('Content-Type'):
            html_bytes = response.read()
            html_string = html_bytes.decode("utf-8")
        finder = LinkFinder(Spider.base_url, page_url)
        finder.feed(html_string)
    except Exception as e:
        print(str(e))
        return set()
    return finder.page_links()

@staticmethod
def add_links_to_queue(links):
    for url in links:
        if (url in Spider.queue) or (url in Spider.crawled):
            continue
        if Spider.domain_name != get_domain_name(url):
            continue
        Spider.queue.add(url)
```

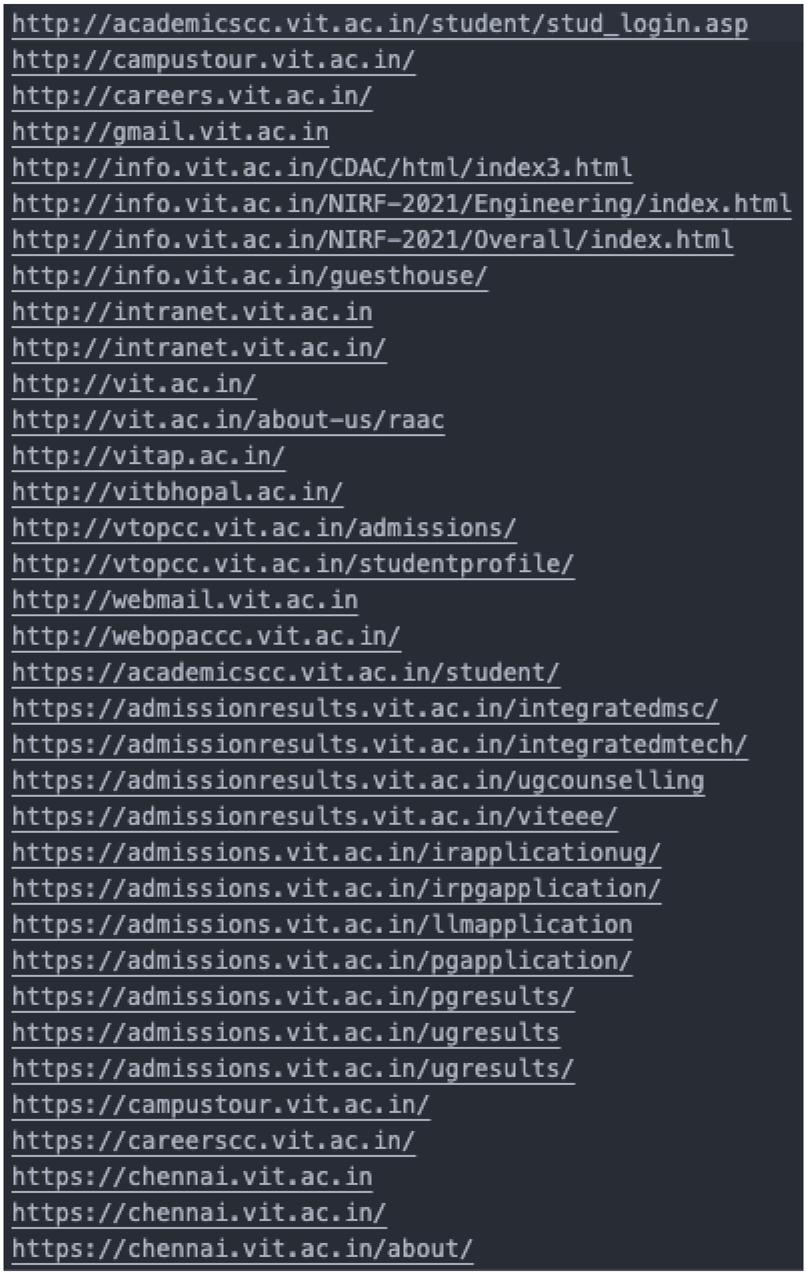
OutPut:

Links crawled:



```
http://chennai.vit.ac.in/
https://vit.ac.in/about/vision-mission
https://vit.ac.in/contactus
https://vit.ac.in/institutional-student-grievance-redressal-committee-isgrc
https://vit.ac.in/internationalrelations/itp
https://vit.ac.in/research
https://vit.ac.in/sites/default/files/Student-Code-of-Conduct.pdf
https://vit.ac.in/vit-university-sets-record-limca-book-records
https://www.vit.ac.in/
https://www.vit.ac.in/article-published-hindu-business-line-co-authored-vit-
https://www.vit.ac.in/national-institutional-ranking-framework-nirf
```

Links in Queue:



```
http://academicscc.vit.ac.in/student/stud_login.asp
http://campustour.vit.ac.in/
http://careers.vit.ac.in/
http://gmail.vit.ac.in
http://info.vit.ac.in/CDAC/html/index3.html
http://info.vit.ac.in/NIRF-2021/Engineering/index.html
http://info.vit.ac.in/NIRF-2021/Overall/index.html
http://info.vit.ac.in/guesthouse/
http://intranet.vit.ac.in
http://intranet.vit.ac.in/
http://vit.ac.in/
http://vit.ac.in/about-us/raac
http://vitap.ac.in/
http://vitbhopal.ac.in/
http://vtopcc.vit.ac.in/admissions/
http://vtopcc.vit.ac.in/studentprofile/
http://webmail.vit.ac.in
http://webopaccc.vit.ac.in/
https://academicscc.vit.ac.in/student/
https://admissionresults.vit.ac.in/integratedmsc/
https://admissionresults.vit.ac.in/integratedmtech/
https://admissionresults.vit.ac.in/ugcounselling
https://admissionresults.vit.ac.in/viteee/
https://admissions.vit.ac.in/irapplicationug/
https://admissions.vit.ac.in/irpgapplication/
https://admissions.vit.ac.in/llmapplication
https://admissions.vit.ac.in/pgapplication/
https://admissions.vit.ac.in/pgresults/
https://admissions.vit.ac.in/ugresults
https://admissions.vit.ac.in/ugresults/
https://campustour.vit.ac.in/
https://careerscc.vit.ac.in/
https://chennai.vit.ac.in
https://chennai.vit.ac.in/
https://chennai.vit.ac.in/about/
```