VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ▪ CHENNAI
www.vit.ac.in

Btech

21st March 2018

Scrap the GSOC Website

A Project Report

By

Harshit Bhatia (16BIT0109)

Under the guidance of

GDG 2 Credit Course Mentors

Acknowledgment

I sincerely thank Dr.G.Viswanathan - Chancellor, VIT University, for creating an opportunity to use the facilities available at VIT. I also thank GDG Team for giving us the opportunity to do this project. I also thank the Dean and the entire department of Information Technology, School of Information Technology and Engineering, for giving us this opportunity.

Introduction

Task

Create an API using Flask or any framework you are comfortable with (Flask is easy to learn) and scrape the GSOC website to fetch details of all the organizations in Google summer of codes and in the api send their name, link to their website, description (which comes on clicking learn more), The technologies they use and their contact email.

Example-

The api link is say http://localhost:8080/orgs . The user should get a list of organizations and each member of the list should look like this or should have same kind of schema.

{

organization: 3DTK,

link: http://threedtk.de/,

description: The 3D Toolkit ....,

technologies: [c/c++, cmake, opencv, ros, boost], c contact: johannes.schauer@uni-wuerzburg.de

}

Website

https://summerofcode.withgoogle.com/archive/2017/organizations/

FrameWorks used- Flask(Python webframework)

Library used- Beautiful Soup(Python library)

Text Editor- Sublime Text and to run the codes I used cmd

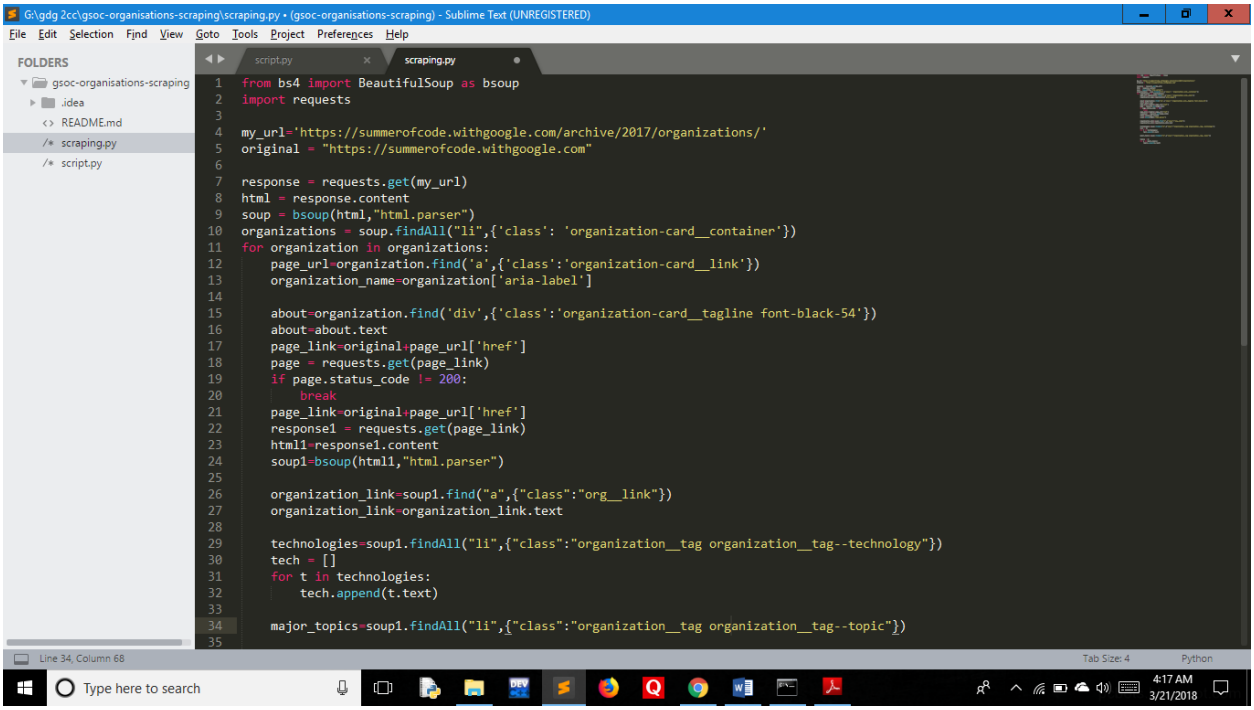
Code:-

Scraping.py(code for fetching the details from the website)

```
from bs4 import BeautifulSoup as bsoup
import requests

my_url='https://summerofcode.withgoogle.com/archive/2017/organizations/'
original = "https://summerofcode.withgoogle.com"

response = requests.get(my_url)
html = response.content
soup = bsoup(html,"html.parser")

organizations = soup.findAll("li",{'class': 'organization-card__container'})


for organization in organizations:
    page_url=organization.find('a',{'class':'organization-card__link'})
    organization_name=organization['aria-label']
```

```
about=organization.find('div',{'class':'organization-card__tagline font-black-54'})
about=about.text
page_link=original+page_url['href']
page = requests.get(page_link)
if page.status_code != 200:
    break
page_link=original+page_url['href']
response1 = requests.get(page_link)
html1=response1.content
soup1=bsoup(html1,"html.parser")

organization_link=soup1.find("a",{"class":"org__link"})
organization_link=organization_link.text

technologies=soup1.findAll("li",{"class":"organization__tag organization__tag--technology"})
tech = []
for t in technologies:
    tech.append(t.text)

major_topics=soup1.findAll("li",{"class":"organization__tag organization__tag--topic"})
```

```
topics = []
for q in major_topics:
        topics.append(q.text)
```



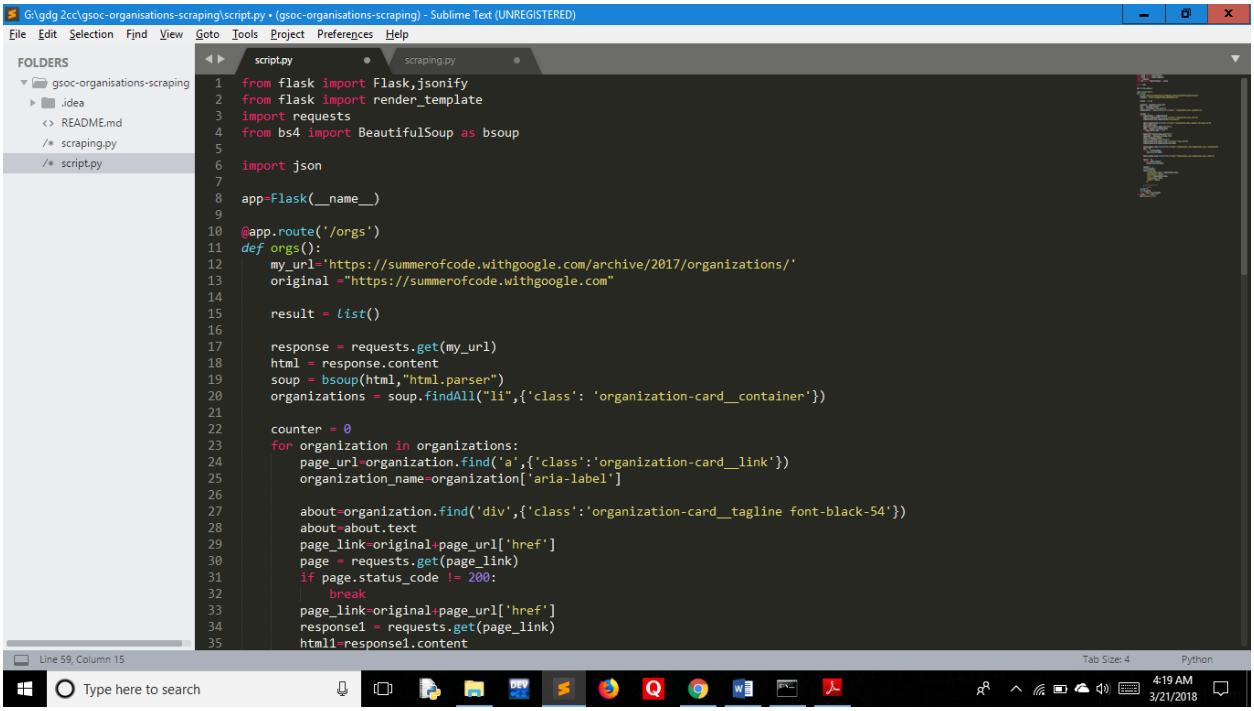
Script.py(Main File used Flask api to send in json format)

```
from flask import Flask,jsonify
from flask import render_template
import requests
from bs4 import BeautifulSoup as bsoup


import json


app=Flask(__name__)
```

```
@app.route('/orgs')
def orgs():
	my_url='https://summerofcode.withgoogle.com/archive/2017/organizatio
ns/'
	original ="https://summerofcode.withgoogle.com"

	result = list()

	response = requests.get(my_url)
	html = response.content
	soup = bsoup(html,"html.parser")
	organizations = soup.findAll("li",{'class': 'organization-card__container'})

	counter = 0
	for organization in organizations:
		page_url=organization.find('a',{'class':'organization-card__link'})
		organization_name=organization['aria-label']

		about=organization.find('div',{'class':'organization-card__tagline
font-black-54'})
		about=about.text
		page_link=original+page_url['href']
		page = requests.get(page_link)
		if page.status_code != 200:
```

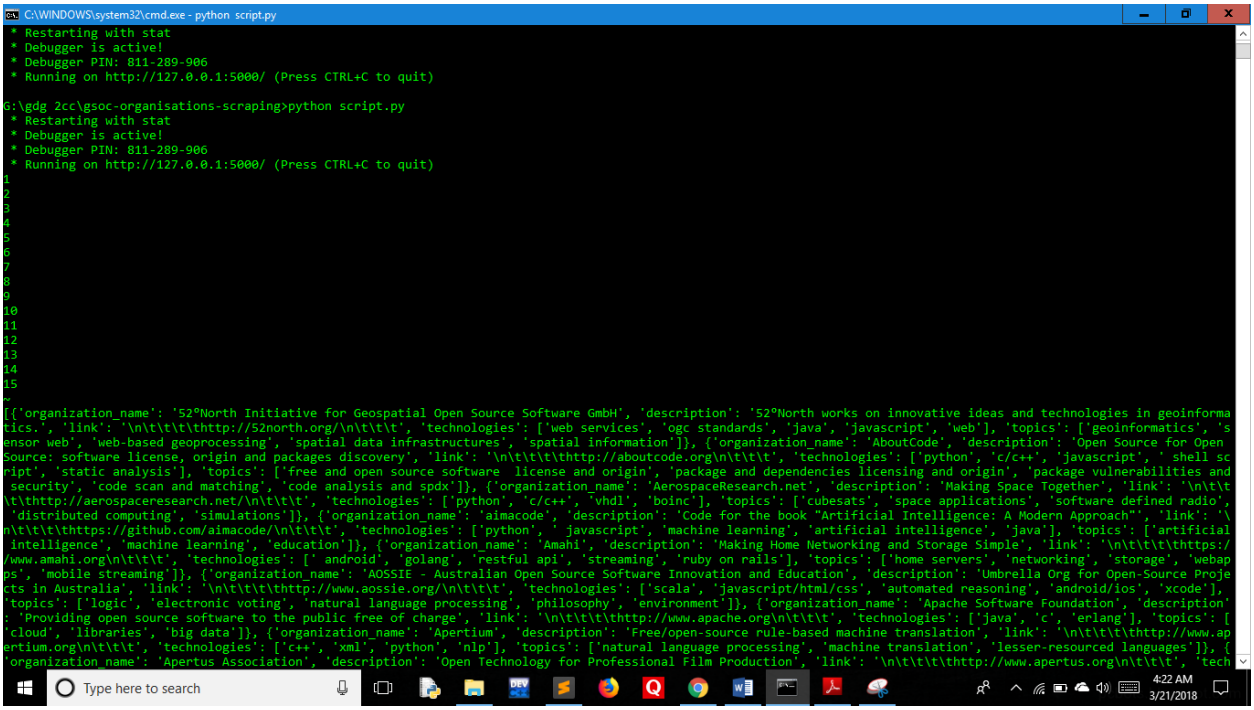
```
            break
        page_link=original+page_url['href']
        response1 = requests.get(page_link)
        html1=response1.content
        soup1=bsoup(html1,"html.parser")
        organization_link=soup1.find("a",{"class":"org__link"})
        organization_link=organization_link.text

        technologies=soup1.findAll("li",{"class":"organization__tag
organization__tag--technology"})
        tech = []
        for t in technologies:
            tech.append(t.text)

        major_topics=soup1.findAll("li",{"class":"organization__tag
organization__tag--topic"})

        topics = []
        for q in major_topics:
            topics.append(q.text)

        counter += 1
        print(counter)
        result.append({
            'organization_name': organization_name,
```

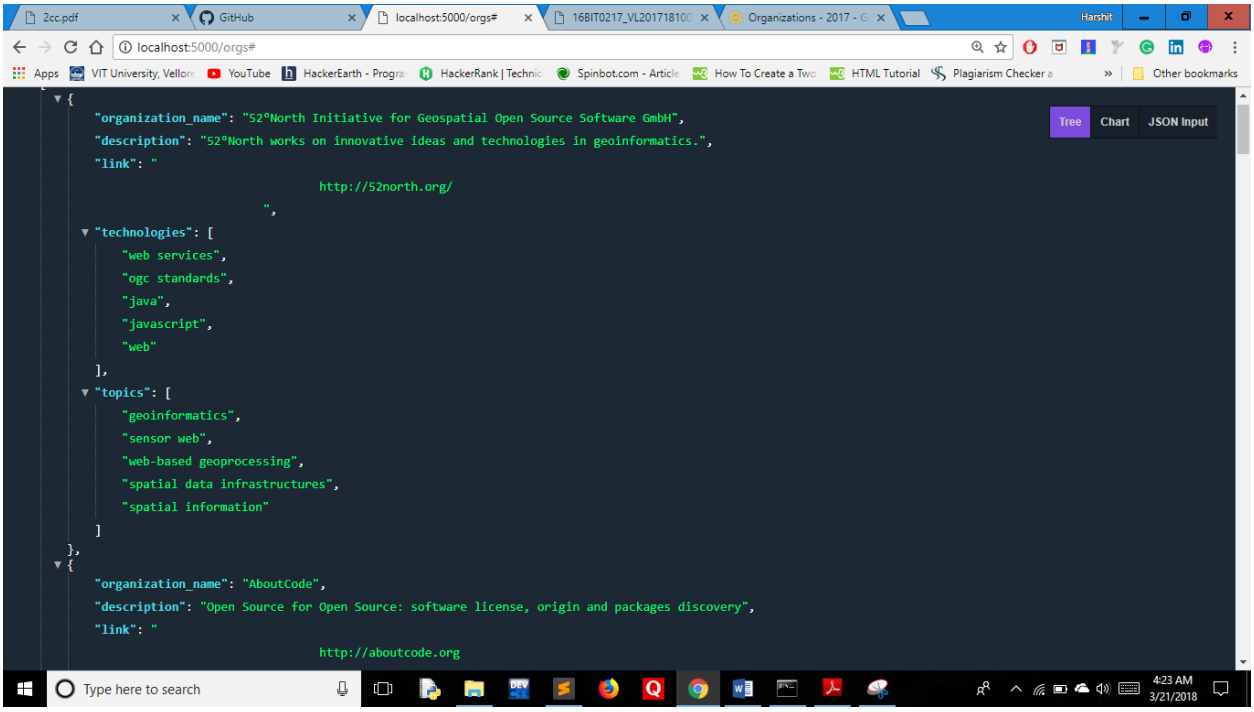
```
                'description': about,
                'link': organization_link,
                'technologies':tech,
                'topics': topics
                })

            # if (counter==5):
            #     break
    print("~")
    print(result)
    return json.dumps(result)
if __name__ == "__main__":
    app.run(debug=True)
```



Output:

I am showing you data upto 15 organisations.





Refrences:-

https://summerofcode.withgoogle.com/archive/2017/organizations/

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

http://flask.pocoo.org/docs/0.12/