Name: Abhinaw Gupta
Submitted for: - 2credit course project - gdgvit
Reg No: - 15BCE0677

Abstract

This document represents my work for the 2-credit course at GDGVIT – 2017-18. In this project I have used python to scrap the website: "https://summerofcode.withgoogle.com/ organizations/" to fetch the details of all the organization affiliated to gsoc. This report includes the detailed discussion of the methodology and techniques I used for scrapping the organization website. I used selenium, request and beautifulsoup to fetch, parse and execute project requirements.

Introduction

The **Google Summer of Code**, often abbreviated to **GSoC**, is an international annual program, first held from May to August 2005, in which Google awards stipends, which depends on the purchasing power parity of the country the student's university belongs to, to all students who successfully complete a requested free and open-source software coding project during the summer. The program is open to university students aged 18 or over.

The program invites students who meet their eligibility criteria to post at most 5 applications that detail the software-coding project they wish to work on. These applications are then evaluated by the corresponding mentoring organization. Every participating organization must provide mentors for each of the project ideas received, if the organization believes the project would benefit from them. The mentors then rank the applications and decide among themselves which proposals to accept. Google then decides how many projects each organization gets considering the number of applications the organization has received, and asks the organizations to mark at most that many projects accordingly.

In the event of a single student being marked in more than one organization, Google mediates between all the involved organizations and decides who "gets" that student. The other mentoring organization then unmarks the student and marks a new proposal for acceptance, or gives their slot back to the pool, after which it is redistributed.

Thus, my task is to scrap the organization website to fetch the details of all organization related to google summer of code(GSOC).

Methodology

Prerequisites: -

- **i) Python 3.6**
- **ii) Pycharm**
- **iii) Libraries: -**
  - **a) Requests**
  - **b) Pillow**
  - **c) Selenium**
- **iv) Geckodriver**

v) Mozilla Firefox(or chrome)

Procedure followed:

i) Create a project in pycharm (I used pycharm)
ii) Setup gecko driver control for selenium
iii) Import webdriver from selenium
iv) Initialize url to be scrapped e.g.:
"https://summerofcode.withgoogle.com/organizations/"
v) Initialize binary for mozilla
vi) Initialize driver
vii) Get the diver link content
viii) Manual search the script to be scrapped (Here angularJS)
ix) Scarp the code and execute it in the binary
x) Scrap the executed result
xi) Search for the technology tag and other relevant data needed to be scrapped
xii) Store the data in dictionary
xiii) Display the result

Result

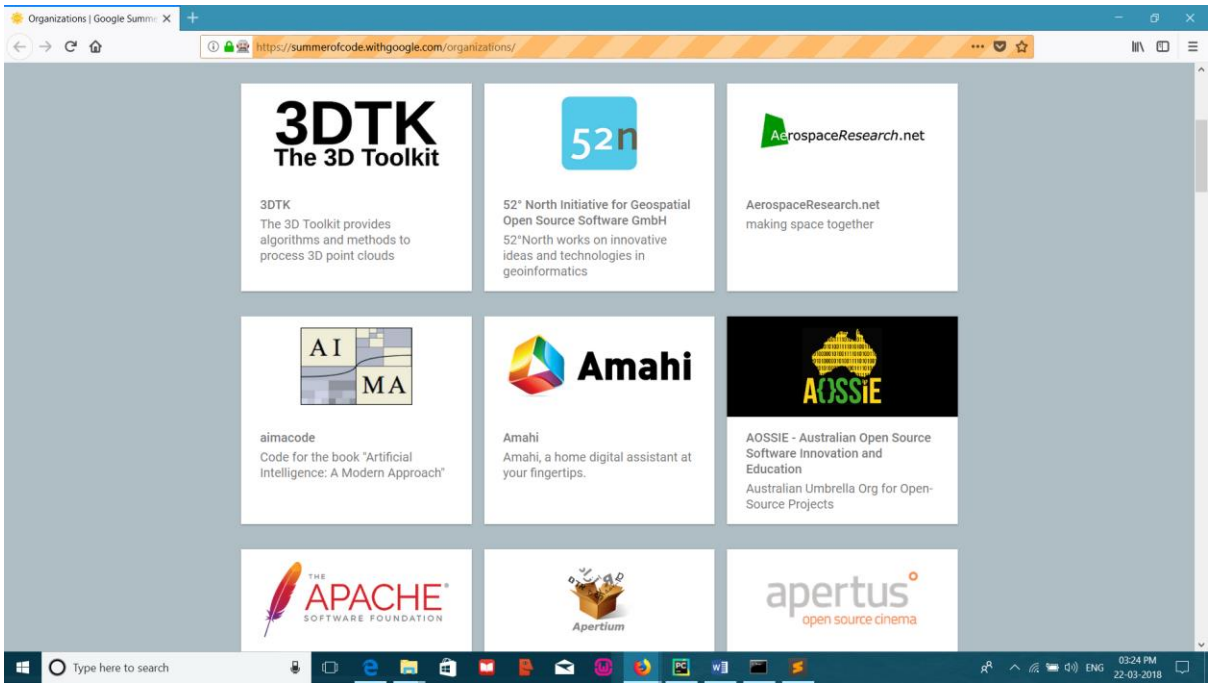

Figure: - 1 (Web browser controlled by gecko)

Scrapped organizations

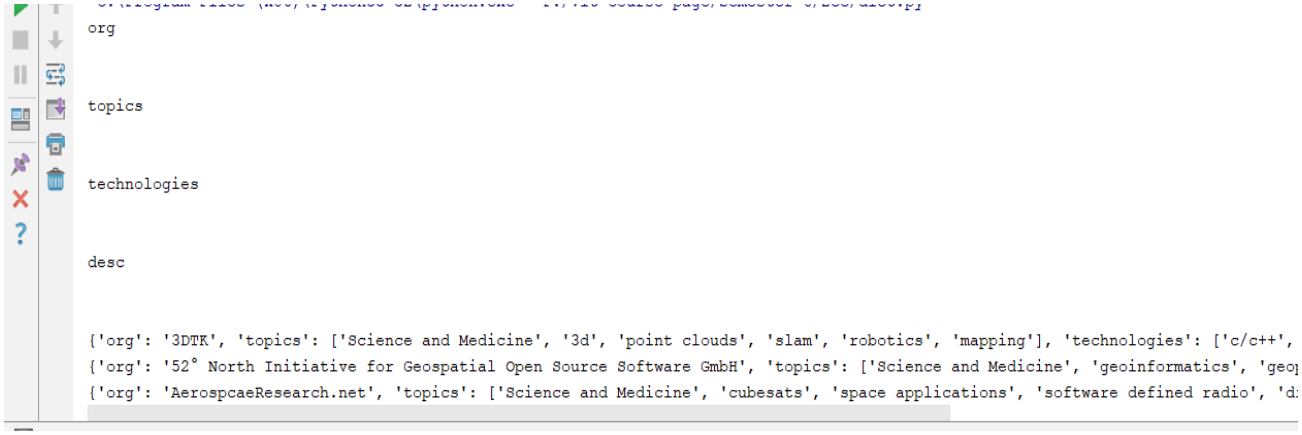

Figure – 2 – scrapped detail of organizaition

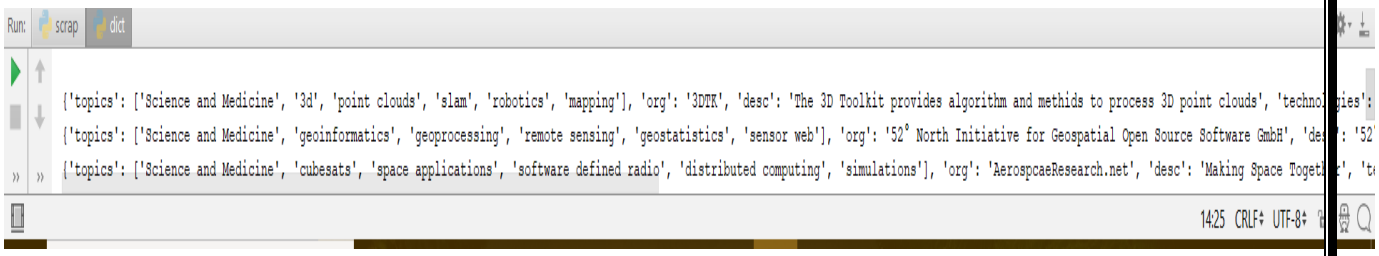

Figure – 3 scrapped dictionaries

Conclusion

The project gives a brief idea on how to scrap google summer of code website using selenium and python. The details of the organizations are stored in the dictionary and then printed as required.

References

i) http://flask.pocoo.org/docs/0.12/
ii) https://www.crummy.com/software/BeautifulSoup/bs4/d0c/