

A PBL REPORT ON

Personalized analysis of colleges.

SUBMITTED BY

Ronak Karani

Tejas Choudhari

Shailesh Gupta

Rahul Chavan

Abhay Gupta



MENTORS

Prof. Dhiraj Amin

Prof. Gayatri Hegde

Prof. Suresh Babu

Prof. Madhura Vyawahare

Prof. Smita Joshi

Prof. Neetu P.

DEPARTMENT OF INFORMATION TECHNOLOGY

PILLAI COLLEGE OF ENGINEERING

NEW PANVEL – 410 206

UNIVERSITY OF MUMBAI

Academic Year 2017 – 18

AIM: To create a solution for Personalized analysis of colleges for selecting the best college for medical/engineering graduate course.

PROBLEM DEFINITION:

The need and importance of extracting data from the web is becoming increasingly loud and clear. Through web scraping data need to be scrapped out from the college websites. The extracted data can be of various features of the college likeno. faculty, students intake, placement activity, social and sports achievement and other. Through this a dataset needs to be prepared which will aid the students in deciding which college is best suited for them for taking admission as per their area of interest.

RELEVANCE OF SUBJECT:

- **PYTHON:**

Libraries or Packages are the collection of predefined Modules which can further used by the Programmer as per the Requirement. There is a huge set of Libraries which support Python. We used some Web Scraping Libraries and Web Framework Libraries out of this set to Scrap the data from college Websites and to build a GUI to present the data respectively. The different Libraries used are as following:

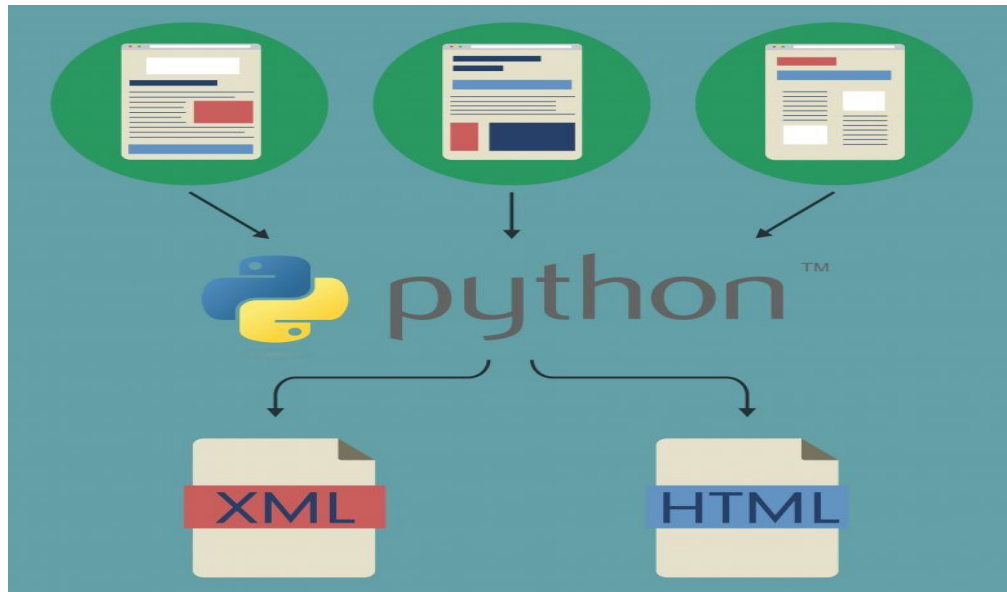
1. Requests

Requests allows you to send *organic, grass-fed* HTTP/1.1 requests, without the need for manual labor. There's no need to manually add query strings to your URLs, or to form-encode your POST data. Keep-alive and HTTP connection pooling are 100% automatic as it uses urllib3.

The main disadvantage of Requests is that it is useful only when one needs to scrap the data that is statically available when the website is loaded. So we can't scrap the Dynamically appearing data on the Website.

2. BeautifulSoup

Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. It can be used along with Requests. The code contents pulled by Requests is mostly Unstructured therefore BeautifulSoup is used to convert them into a Structured format. With the help of methods of BeautifulSoup the data can be extracted from the parse tree.



3.Selenium

Selenium Python bindings provides a simple API to write functional/acceptance tests using Selenium WebDriver. Through Selenium Python API one can access all functionalities of Selenium WebDriver in an intuitive way. Selenium Python bindings provide a convenient API to access Selenium WebDrivers like Firefox, Ie, Chrome, Remote etc. The current supported Python versions are 2.7, 3.5 and above.

Since the WebDrivers are like a Browser it helps to scrap the Dynamic data from a website by Automating the various tasks like filling form ,entering values in the Text fields,clicking Buttons,etc.

With Selenium after performing the required operations we can get that Dynamically generated source code of the webpage which can again be parsed with BeautifulSoup.

4.Flask

Flask is a micro web framework written in Python and based on the Werkzeug toolkit and Jinja2 template engine. It is BSD licensed.

Features:

- Contains development server and debugger
- Integrated support for unit testing
- RESTful request dispatching
- Uses Jinja2 templating
- Support for secure cookies (client side sessions)
- 100% WSGI 1.0 compliant
- Unicode-based

Thus Flask is powerful enough to host a website.

5.Pandas

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It is used to present the data scraped in a Tabular manner.

● **COMPUTER NETWORKS.**

With the knowledge of Computer Networks it gets easier to get the intuition of how the data flows along the internet. This helped us in building the foundation of gathering the data.

INTRODUCTION:

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. We have used Web Scraping to scrap/gather the data from the college websites'. With the help of Requests and BeautifulSoup we scraped the statically available data and used Selenium to scrap dynamically available data. For a College first we collected list all the Courses available in the college. Then for each course we gathered Fee details and Eligible Examination for admission. Also for each college we got the University to which they are affiliated. The locality of college and the various facilities provided by the college are too recorded. The data is stored in a Database which is further connected to the website.

On the Website students can create a account on registering with Email ID and choosing a Username for themselves . Once done with registration they can Log in with their Username and Password. Once logged in they are taken to the main Page where they can choose the Stream of their choice Engineering or Medical. After selecting the Stream students have option to check details of a specific College from a drop down menu or select a course of their choice along with the locality from where he/she wants to see the colleges.

METHODOLOGY:

- Send a HTTP request to the URL of the webpage you want to access. The server responds to the request by returning the HTML content of the webpage. For this task, we will use a third-party HTTP library for python requests.
- Once we have accessed the HTML content, we are left with the task of parsing the data. Since most of the HTML data is nested, we cannot extract data simply through string processing.
- One needs a parser which can create a nested/tree structure of the HTML data. There are many HTML parser libraries available but the most advanced one is html5lib.
- Now, all we need to do is navigating and searching the parse tree that we created, i.e. tree traversal. For this task, we will be using another third-party python library, BeautifulSoup. It is a Python library for pulling data out of HTML and XML files.

SOFTWARES USED:

JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

SPYDER

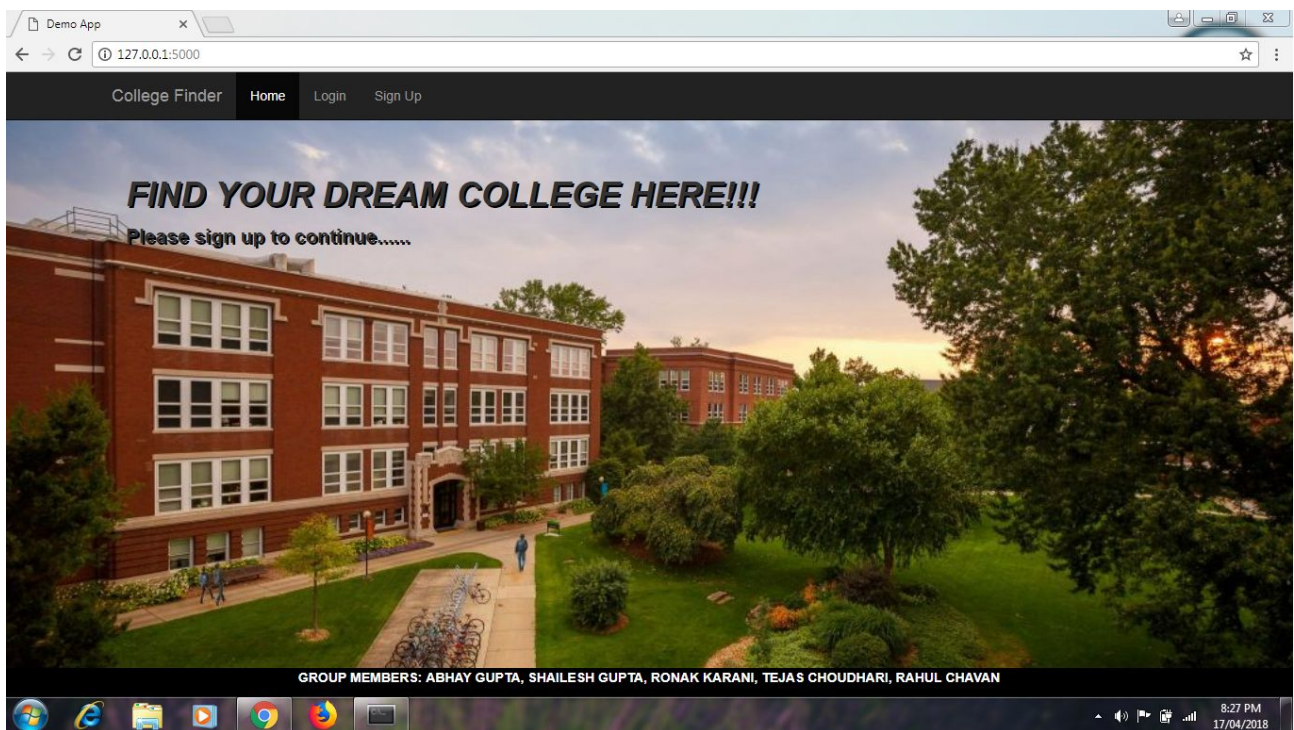
Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates NumPy, SciPy, Matplotlib and IPython, as well as other open source software.

SUBLIME TEXT

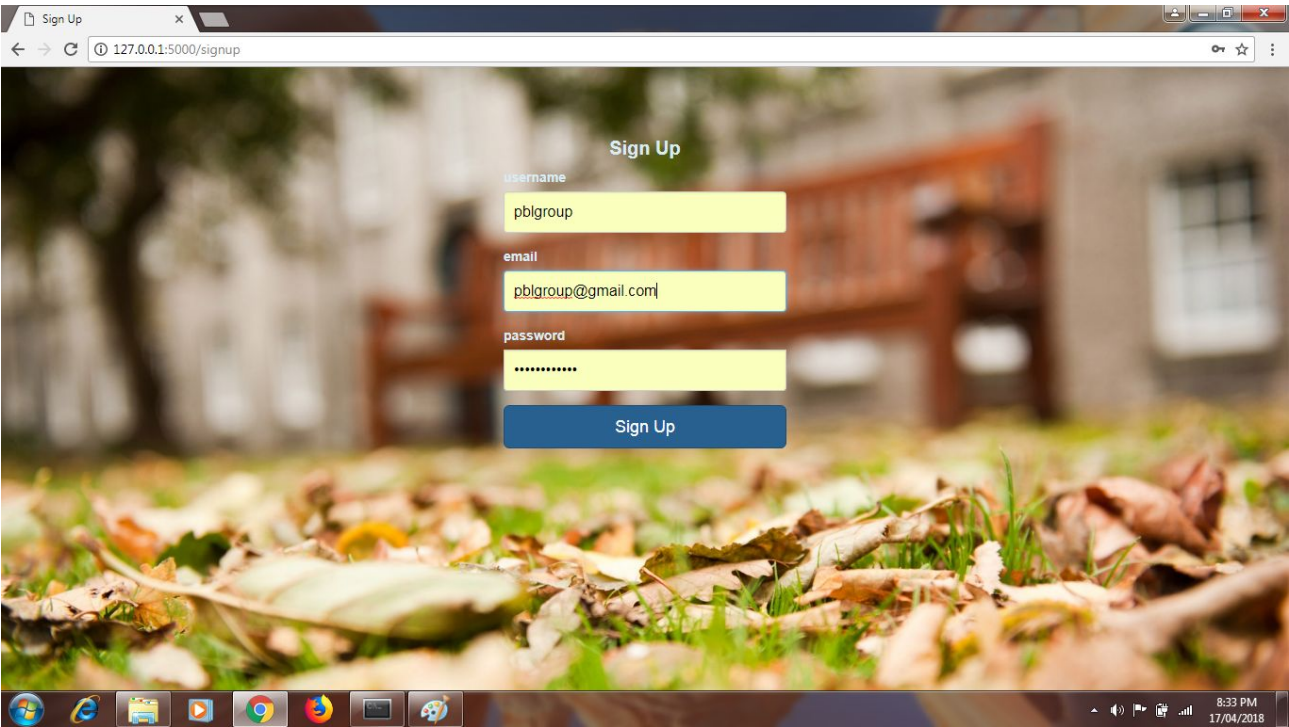
Sublime Text is a proprietary cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and functions can be added by users with plugins.

PROJECT INPUT AND OUTPUT:

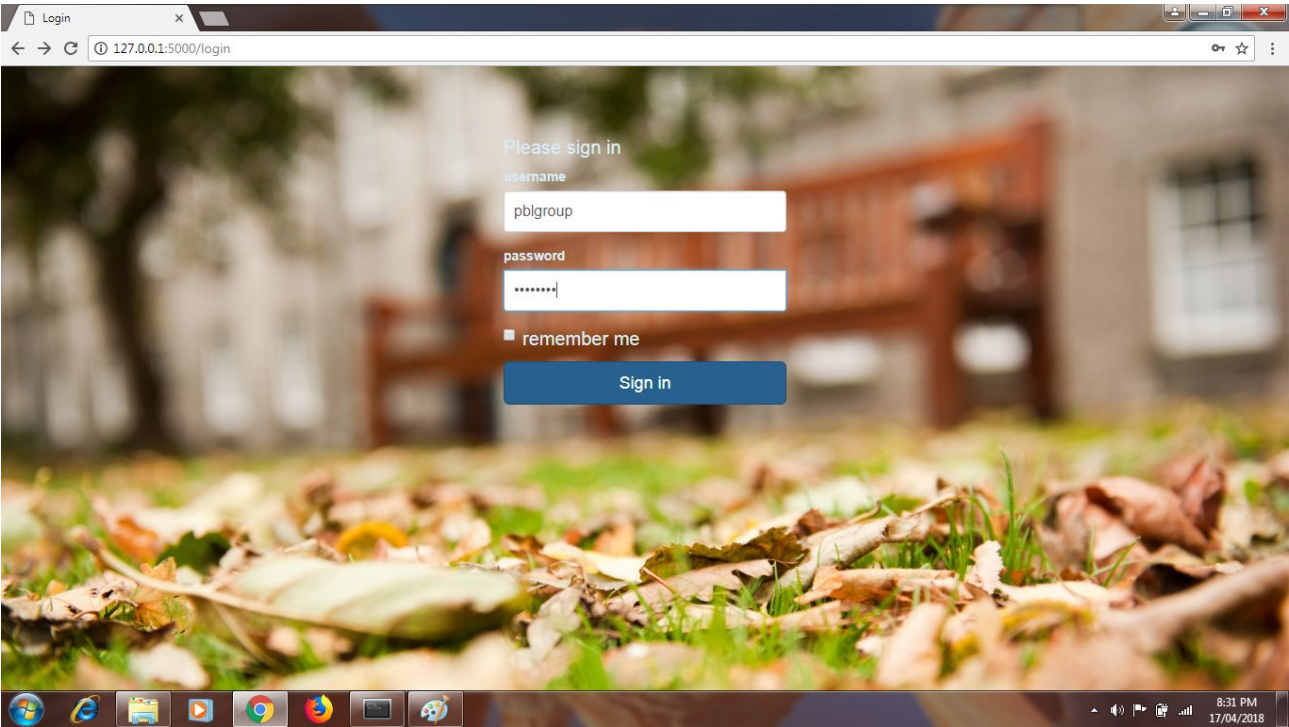
1.Home Page



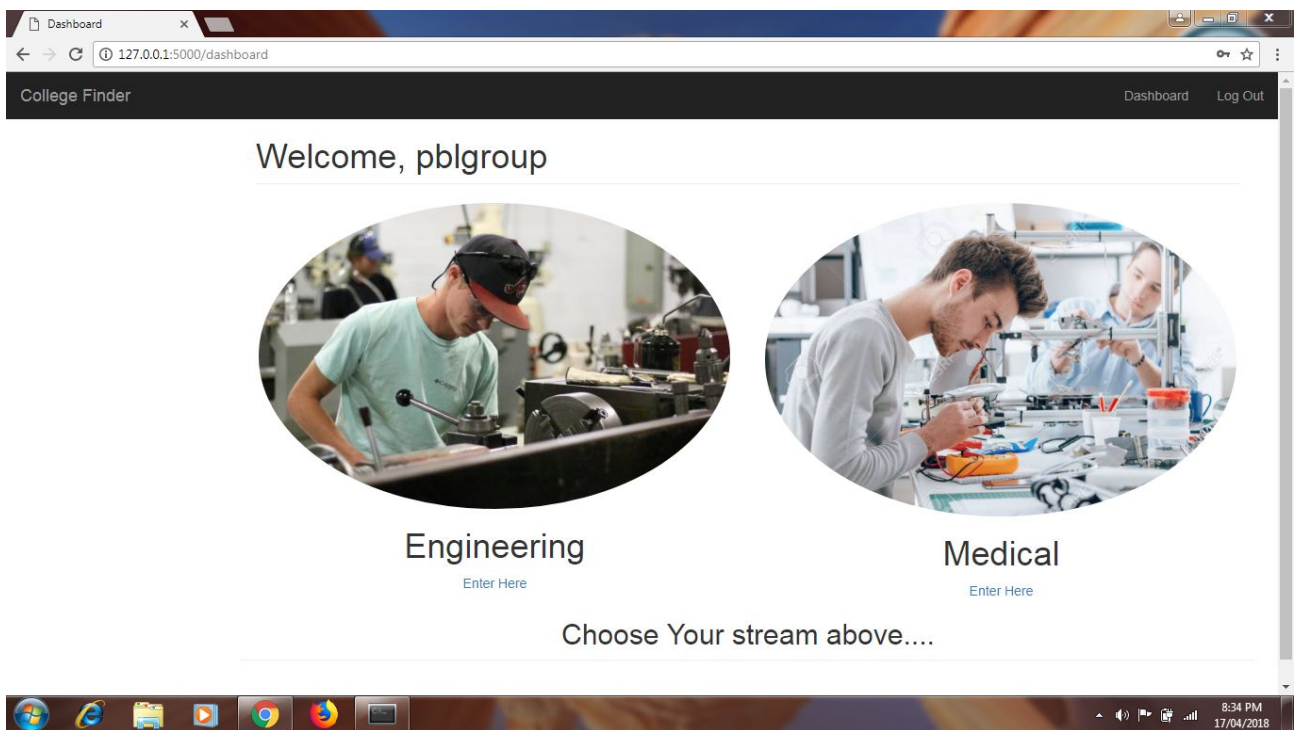
2.Registration Page.



3. Login Page



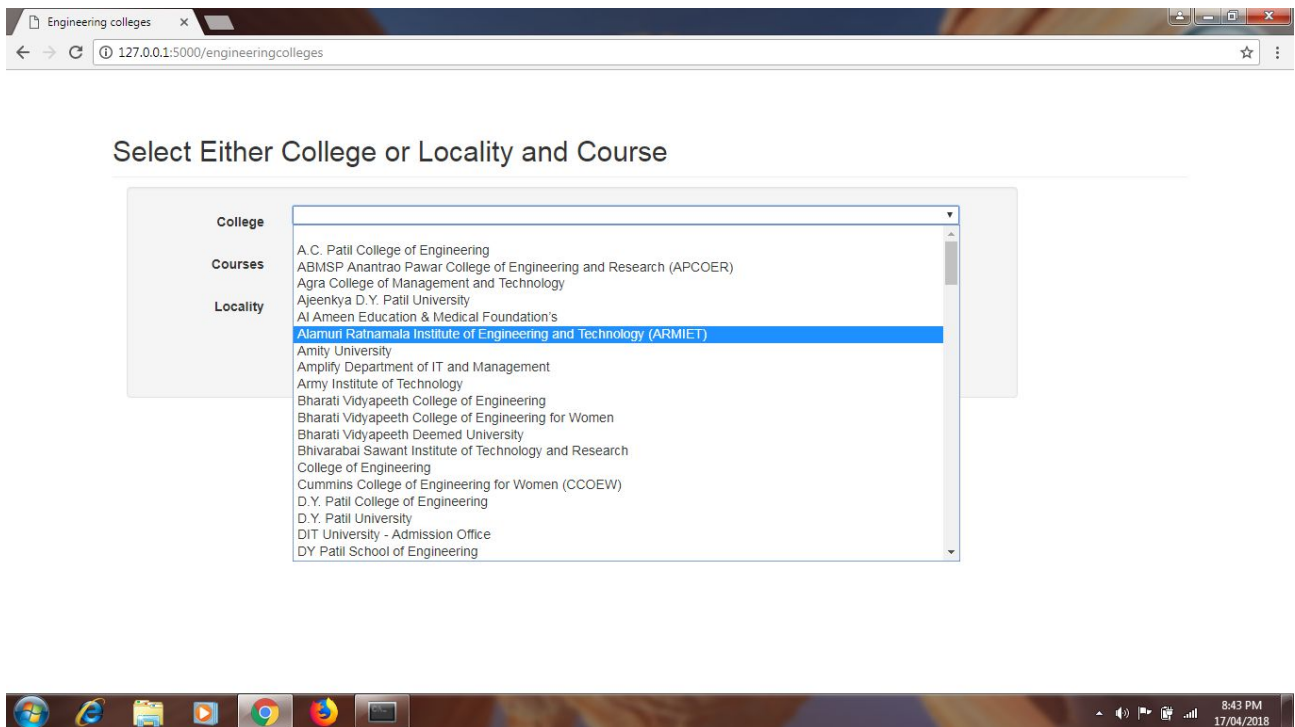
4. Dashboard

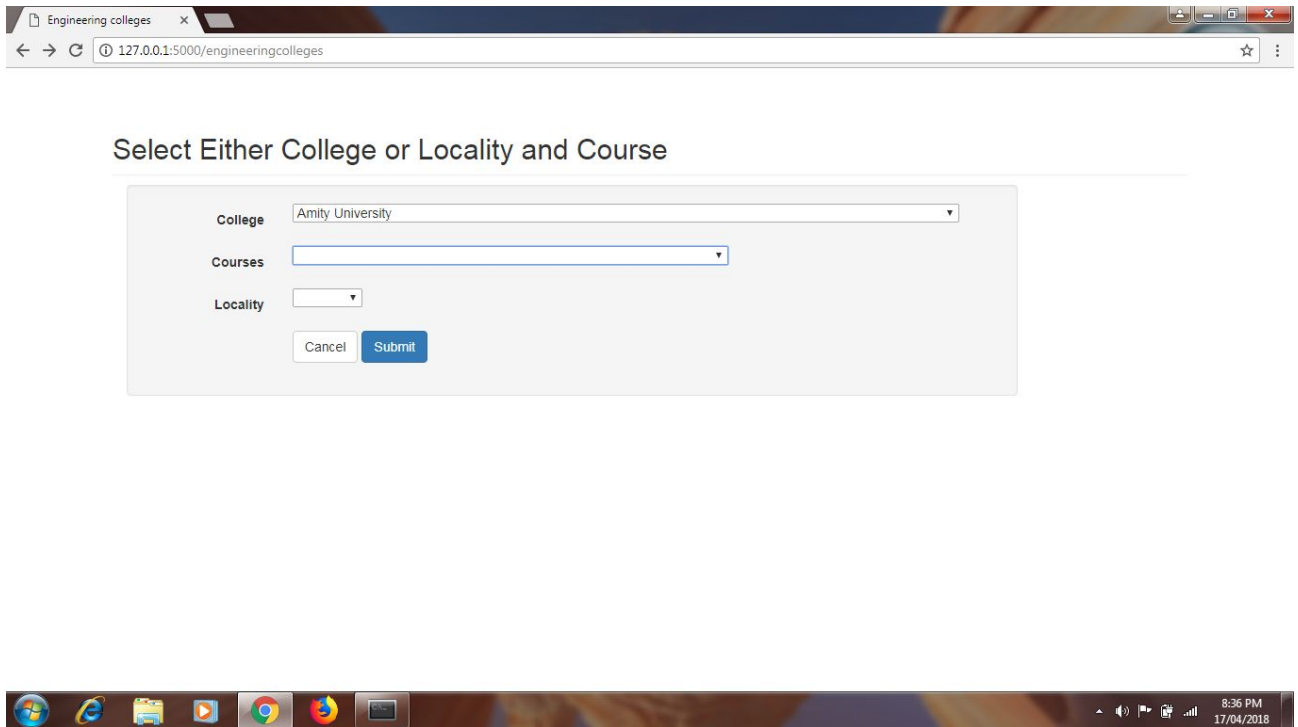


5.

A. Engineering

i. When Student selects a particular college.





o/p:

THIS IS WHAT WE FOUND FOR YOU...

Name	Locality	Course	Fees	Exam	Affiliation	Facilities
Amity University	Panvel, Mumbai	B.Tech. in Computer Science and Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Civil Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Aerospace Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Electronics and Communication Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Electrical and Electronics Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Civil Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium
Amity University	Panvel, Mumbai	B.Tech. in Aerospace Engineering	8.40 Lakh	Amity JEE	Offered by Amity University, Mumbai	Moot Court (Law), Design Studio, Library, Cafeteria, Hostel, Sports Complex, Gym, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium

ii. When Student selects a particular course and locality.

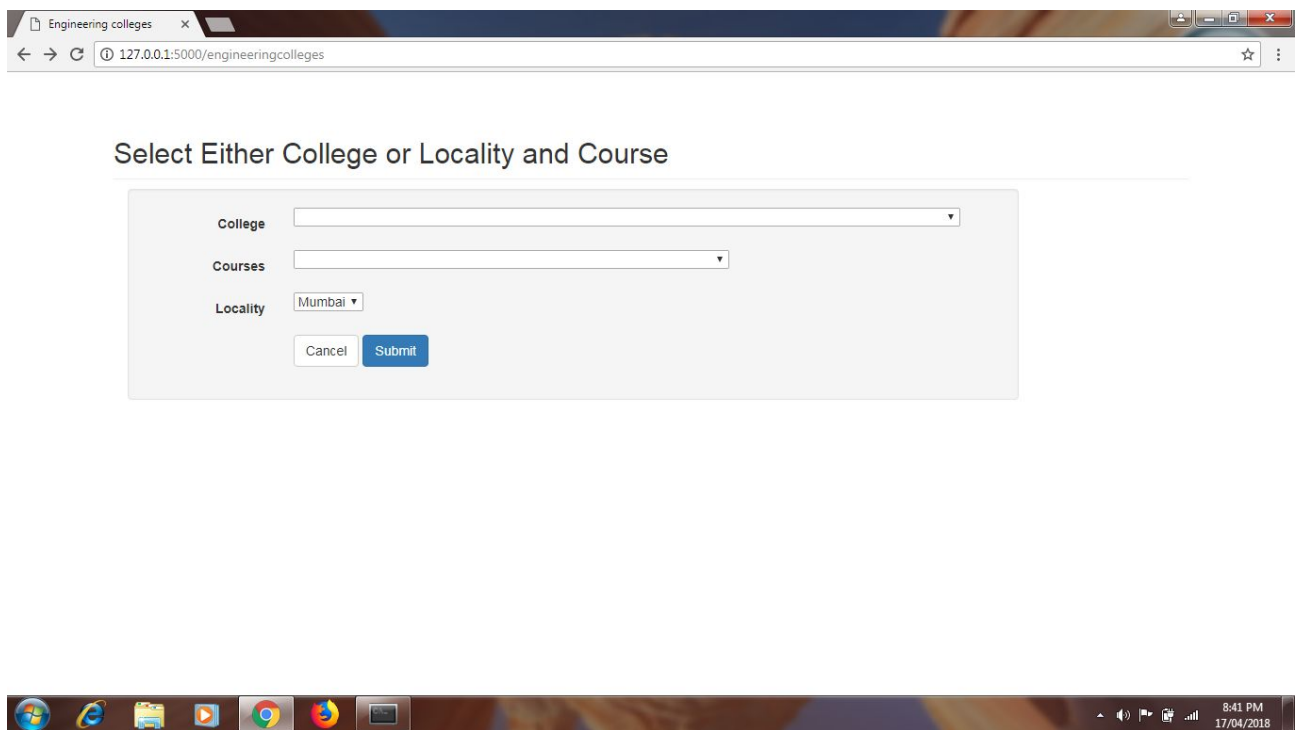
The screenshot shows a web browser window with the address bar displaying "127.0.0.1:5000/engineeringcolleges". The page title is "Engineering colleges". The main heading is "Select Either College or Locality and Course". Below this, there is a form with three dropdown menus: "College" (empty), "Courses" (selected: "B.E in Civil Engineering"), and "Locality" (selected: "Mumbai"). At the bottom of the form are two buttons: "Cancel" and "Submit".

o/p:

The screenshot shows a web browser window with the address bar displaying "127.0.0.1:5000/engineeringsubmit". The page title is "Short Listed Colleges". The main heading is "THIS IS WHAT WE FOUND FOR YOU...". Below this, there is a table with 7 columns: Name, Locality, Course, Fees, Exam, Affiliation, and Facilities. The table contains 3 rows of data.

Name	Locality	Course	Fees	Exam	Affiliation	Facilities
Thakur College of Engineering and Technology	Kandivali East, Mumbai	B.E in Civil Engineering	4.86 Lakh	JEE Mains & MHTCET	Affiliated to University of Mumbai, ...	Library, Cafeteria, Hostel, Sports Complex, Hospital / Medical Facilities, Wi-Fi Campus, Shuttle Service, Auditorium, A/C Classrooms, Labs
Saraswati College of Engineering	Navi Mumbai	B.E in Civil Engineering	3.73 Lakh	MHTCET & JEE Mains	Affiliated to University of Mumbai, ...	Library, Cafeteria, Sports Complex, Gym, Hospital / Medical Facilities, Shuttle Service, Auditorium, Labs
Sahyadri Valley College of Engineering and Technology	Mumbai Pune Road, Pune	B.E in Civil Engineering	-	JEE Mains & MHTCET	Affiliated to Savitribai Phule Pune ...	-

iii. When user just selects locality and ignores course selection.



Engineering colleges

127.0.0.1:5000/engineeringcolleges

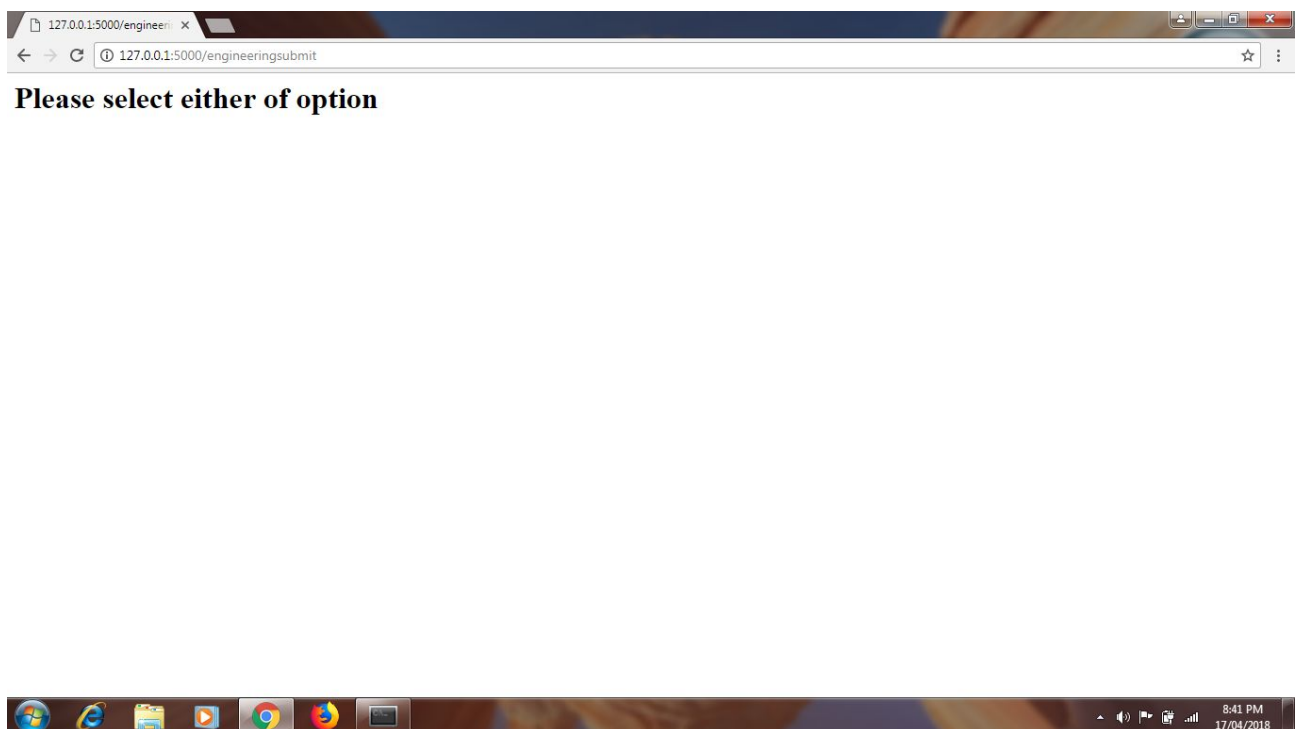
Select Either College or Locality and Course

College

Courses

Locality

o/p:



127.0.0.1:5000/engineer

127.0.0.1:5000/engineersubmit

Please select either of option

B.Medical

i. When student selects specific Course and Locality.

Medical Colleges

127.0.0.1:5000/medicalcolleges

Select Either College or Locality and Course

College

Courses

Locality

o/p:

Short Listed Colleges

127.0.0.1:5000/medicalssubmit

THIS IS WHAT WE FOUND FOR YOU...

Name	Locality	Course	Fees	Exam	Affiliation	Facilities
KJ Somaiya Medical College and Research Centre	Sion, Mumbai	M.B.B.S.	☐ -	NEET	Affiliated to Maharashtra University...	Library, Hostel, Hospital / Medical Facilities, Auditorium, Labs
Seth Gordhandas Sunderdas Medical College	Parel, Mumbai	M.B.B.S.	☐ -	NEET	Affiliated to Maharashtra University...	-

APPLICATION AND FUTURE SCOPE:

- To add data of colleges from more cities.
- To make our website more interactive.
- To add a option for students to bookmark courses and colleges for their future references.
- To get Placements for the applicable courses.

PROJECT OUTCOME:

- We got to learn how data is pulled from the internet using web scraping.
- We learned how to Automate tasks using Selenium.
- We learned how to create a website using Flask and host it on the local server.

ACKNOWLEDGEMENT:

We would like to thank all the faculties for guiding us throughout the project, Mr.Dhiraj Amin sir, Mrs.Gayatri Hegde ma'am, Mrs.Madhura Vyawahare ma'am, Mrs.Smita Joshi ma'am, Mr.Krishnan Babu sir, Mrs.Neethu P. ma'am for helping us throughout the project.

A Special Thanks to Dr. Vasudevan Pillai sir and our Principal Dr.SANDEEP JOSHI sir. Head of the IT department Mrs. Sharvari G. ma'am for giving us an opportunity to learn many new concepts by conducting The Project Based Learning.