

**SUMMER INTERNSHIP-II**

**(Innovation/Societal/Entrepreneurship-based Internship)**

(Academic Batch 2021-25)

**Submitted by**

by

|  |  |
| --- | --- |
| **NAME SUSHMITHA**  **USN 4SF21IS114**  **DEPARTMENT ISE** |  |

Under the Mentorship of

# Dr. Duddela Sai Prashanth

# November 2023

**CERTIFICATE**

This is to certify that the **SUMMER INTERNSHIP-II (Innovation/Societal/Entrepreneurship)** has been carried out by Ms. Sushmitha bearing the USN 4SF21IS114, bonafide student of Department of Information Science &Engineering , Sahyadri College of Engineering & Management, Adyar, Mangaluru, during the Academic Year 2022-23.

The internship report is verified as per the requirements of the Academic Statute and is recommended for the award of the Academic Credit for the said course.

|  |  |
| --- | --- |
| **Mentor** | **Head of the Department** |
| Mr. Vasudev Rao P V | Dr. Mustafa Basthikodi |

**Name of the Examiner Signature with Date**

1.............................................. ...............................

2.............................................. ...............................

**ACKNOWLEDGEMENT**

It is with great satisfaction and euphoria that I am submitting a report on internship carried out at ”**Sahyadri College of Engineering & Management**” provided by **Coding key LLP** in partial fulfillment of the requirements for the IV Semester of Bachelor of Engineering in Information Science & Engineering. I have completed the project entitled “**Developing a search engine for organization**” during the internship.

I am immensely grateful to our guide, **Dr. Duddela Sai Prashanth**, for his invaluable guidance, timely advice, and unwavering encouragement. His mentorship has been instrumental in our journey, and I sincerely express my heartfelt gratitude.

I express our sincere gratitude to **Dr. Mustafa Basthikodi**, Head & Associate Professor, Department of Information Science & Engineering for his invaluable support and guidance.

I sincerely thank **Dr. Sidramappa Shivanna Injaganeri**, Principal, Sahyadri College of Engineering & Management and Dr. D. L. Prabhakara, Director, Sahyadri Educational Institutions,who have always been a great source of inspiration.

SUSHMITHA(4SF21IS114)

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** | |
| Acknowledgement | ii |
| Table of Contents | iv |
| List of Figures | v |
| List of Tables | v |
| Chapter 1 week 1 | Page No |
| Chapter 2 week 2 | Page No |
| Chapter 3 week 3 | Page No |
| Chapter 4 week 4  Chapter 5 Final Solution | Page No |
| Chapter 6 Conclusion | Page No |
| References | Page No |

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| Figure 3.1 | Name | 13 |
| Figure 3.2 | Name | 15 |
| Figure 3.3 | Name | 15 |

**Chapter 1**

**Week 1**

**1.1 Introduction**

\* Understanding search engines is crucial in today's digital landscape. A search engine is

tool that helps users find information on the internet by indexing and organizing vas

amounts of web content. It employs algorithms to analyze and rank web pages, presenting

users with relevant results based on their queries.

**1.2 Day 1**

\* Clearly understanding the concept of search engine, how does it works and in which fields

We can apply this and in what way it will be useful for us. And what is its application in

Current situation.

**1.3 Day 2**

\* Understanding the concepts of crawling where crawling is the process where search engines

use automated bots to Navigate the web and gather information from web pages.

**1.4 Day 3**

**\*** Clearly understood about the crawling process and applied that crawling

Concept to collect the data from a particular website.

**1.5 Day 4**

**\*** Crawlers, also known as spiders or bots, navigate from one page to another through links,

collecting information for indexing. The bots crawl through the link which we provided

along with the code and this link can be added through the html using the anchor tag.

**1.6 Day 5**

**\*** Clear and logical URL structures can facilitate crawling and improve search engine

optimization.

\* It helps control the access of search engine bots to specific parts of a website.

**1.7 Day 6**

\* Understood Canonicalization concept in crawling where Canonicalization is the process

of selecting the preferred URL when multiple URLs point to the same or similar content.

Canonical tags help search engines understand which version of a URL to prioritize.

**1.8 Conclusion/ skill sets learnt**

\* web scraping and data extraction stands as a testament to our commitment to excellence.

From the initial steps of manual website downloads to the intricacies of implementing a

sophisticated web scraping code

**Chapter 2**

**Week II**

**2.1 Introduction**

\* Data Collection - Establish mechanisms to collect and update data about organizations,

including their names, industries, locations, contact details, leadership, financial data, and

articles relevant news

**2.2 Day 1**

**\*** Understanding the basics of html

Ex: Adding images, Adding URL of a particular website , How the Different tags in html

works and understood about the how does we can Make use of html in the concept of search

engine.

\* HTML plays a crucial role in search engine optimization (SEO) by providing the structure

and content of web pages.

**2.3 Day 2**

**\*** Understanding the basics of CSS .

\* Like, the way of styling html elements. Efficient use of CSS can contribute to faster page

load times.

**2.4 Day 3**

**\*** Used a website to collect the data.

First we collected only single data items.

import requests

import re

url = 'https://sahyadri.edu.in/'

try:

    response = requests.get(url)

    if response.status\_code == 200:

        html\_content = response.text

        phone\_number\_pattern = r'\(?\d{3}\)?[-.\s]?\d{3}[-.\s]?\d{4}'

        phone\_numbers = re.findall(phone\_number\_pattern, html\_content)

        for phone\_number in phone\_numbers:

            print(phone\_number)

    else:

        print(f'Failed to retrieve content. Status code: {response.status\_code}')

except requests.exceptions.RequestException as e:

    print(f'Error: {e}')

**Fig.2.4.1 Code to crawl website and collect all the phone numbers.**

**2.5 Day 4**

import os

import re

import requests

from bs4 import BeautifulSoup

os.makedirs('email', exist\_ok=True)

os.makedirs('contact', exist\_ok=True)

url = 'https://sahyadri.edu.in/'

response = requests.get(url)

if response.status\_code == 200:

    soup = BeautifulSoup(response.text, 'html.parser')

    email\_pattern = r'\b[A-Za-z0-9.\_%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,7}\b'

    emails = re.findall(email\_pattern, soup.get\_text())

    contact\_info\_pattern = r'\b(?:\d{3}[-.\s]?)?\d{3}[-.\s]?\d{4}\b'  # Example: Phone numbers

    contact\_info = re.findall(contact\_info\_pattern, soup.get\_text())

    with open('email/emails.txt', 'w') as email\_file:

        for email in emails:

            email\_file.write(email + '\n')

    with open('contact/contact\_info.txt', 'w') as contact\_file:

        for info in contact\_info:

            contact\_file.write(info + '\n')

    print("Email addresses and contact information scraped and saved successfully.")

else:

    print(f"Error: {response.status\_code}")

**Fig:2.5.1**

The following code crawl the given website through the URL specified and give the emails which are present in the website.

**2.6 Day 5**

import csv

import re

from selenium import webdriver

driver = webdriver.Chrome( )

url = 'https://sahyadri.edu.in/'

driver.get(url)

page\_source = driver.page\_source

email\_pattern = r'\b[A-Za-z0-9.\_%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,7}\b'

phone\_pattern = r'\b\d{3}[-.\s]?\d{3}[-.\s]?\d{4}\b'

emails = re.findall(email\_pattern, page\_source)

phones = re.findall(phone\_pattern, page\_source)

driver.quit()

csv\_filename = 'info.txt'

with open(csv\_filename, mode='w', newline='') as file:

    fieldnames = ['Email', 'Phone']

    writer = csv.DictWriter(file, fieldnames=fieldnames)

    writer.writeheader()

    for email in emails:

        writer.writerow({'Email': email})

    for phone in phones:

        writer.writerow({'Phone': phone})

print(f'Email addresses and phone numbers scraped and stored in {csv\_filename}')

**Fig:2.6.1**

**2.8 Conclusion/ skill sets learnt**

\* Understanding the codes which will scan the website and gives us the required

information.

**Chapter 3**

**Week III**

**3.1 Introduction:**

**\*** Embarking on a week-long exploration, our focus intertwined the realms of thoughtful

project design, dynamic user interfaces, and efficient database integration. Day 1 saw the

conception of a project blueprint - a dual-page system catering to both user and admin

interactions.

**3.2 Day 1:**

**\*** create 2 pages- user page and admin page. User page is the one where the user can enter

the url to be scraped. As soon as user clicks search button the informations(phone number,

email, social media links) will be fetched directly from the database so that user no need to

wait to get the info.

3.3 Day 2:

\* Created Admin page is the one where admin will be able to add a new url . As soon as

the admin adds a new url , that particular url has to be scraped and sould save the info in

the database.

* Also admin can update the database and fix number of days for a update and if it is not updated it will automatically will give a signal.

**3.3 Day 3:**

\* Created a User Interface in which that contains a enter url which is used to enter the new

url and scrap button which will scrape the website and store all the scraped data’s in the

database. Whenever the user entered the url which is already scraped the bot will

automatically give the data which is stored in the database.

**3.4 Day 4:**

**\*** Integrating an admin panel for efficient project management. Admin panel seamlessly

integrated with the Jango database, emphasizing the importance of user-friendly

interfaces.

**3.5 Day 5:**

**\*** Enhanced admin side functionalities in which admin can update and Admin can view all

the statistics which contains how many links have been scraped and how many users are

there and how many data’s has been scraped. And also admin can have the ability to update

the database.

**3.7 Conclusion/ skill sets learnt**

* As the week concludes, our journey through project development reflects a harmonious convergence of design ingenuity and technical prowess.
* The successful integration of an admin panel, streamlined database operations
* The functionalities of the admin and the working of User Interface.

**Chapter 4**

**Week IV**

**REFERENCES**

Times new roman font size 12 line spacing 1.5 (IEEE Format)