**PYTHON PROGRAMMING LAB PROJECT REPORT**

**ON**

**“WEB SCRAPING USING BEAUTIFULSOUP”**

**SUBMITTED**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF**

**BACHELOR OF TECHNOLOGY0**

IN

**INFORMATION TECHNOLOGY**

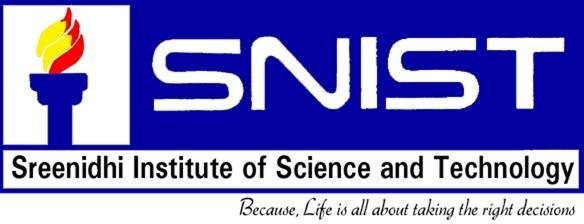
BY

**SAI DEEPESH** [Reg. No. 17311A1207]

**UNDER THE SUPERVISION OF**

**Mr. SREEDHAR**

**ASSISTANT PROFESSOR, IT DEPARTMENT**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**SREENIDHI INSTITUTE OF SCIENCE & TECHNOLOGY**

**(AUTONOMOUS)**

**DECLARATION AND ACKNOWLEDGEMENT**

We hereby declare that the work described in the Project report, entitled “**WEB SCRAPING USING BEAUTIFUL SOUP**” which is being submitted by us in partial fulfillment for the award of **Bachelor of Technology** in the Dept. of **Information Technology, Sreenidhi Institute Of Science & Technology** affiliated to Jawaharlal Nehru Technological University ,Hyderabad, Kukatpally, Hyderabad (Telangana) is the work on our own effort and has not been submitted else where.

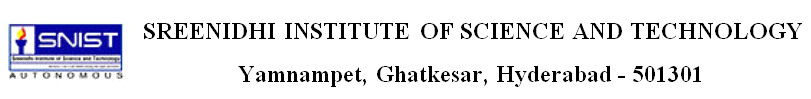
We are very thankful to **Mr. SREEDHAR** , IT Dept., Sreenidhi Institute of Science and Technology, Ghatkesar for providing an initiative to this OS lab project and giving valuable timely suggestions over the work.

We convey our sincere thanks to **Dr.Balaram**, Head of the Department (IT), Sreenidhi Institute of Science and Technology, Ghatkesar, for his kind cooperation in the completion of this work. We even convey our sincere thanks to **Dr. P. Narsimha Reddy** ,Executive Director and **Dr. M.Shiva Reddy**, Principal, Sreenidhi Institute of Science and Technology , Ghatkesar for their kind cooperation in the completion of the OS lab project.

Finally we extend our sense of gratitude to all our friends, teaching and non-teaching faculty, who directly or indirectly helped us in this endeavor.

Place: **Hyderabad.**

Date: **18-11-2019.** SAI DEEPESH



**CERTIFICATE**

**Date: 18-11-2019**

This is to certify that the Project report entitled **"WEB SCRAPING USING BEAUTIFULSOUP"** being submitted by **SAI DEEPESH** [Reg. No.17311A1207] In partial fulfilment for the award of **Bachelor of Technology** degree in **Information Technology** to Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar [Telangana], is a report of review work carried out by them during academic year 2019-2020 under our guidance and supervision.

**Mr. SREEDHAR**

Assistant Professor

IT Department

**Dr. BALARAM**

Professor

HOD, IT Department

**CONTENTS :**

* ABSTRACT
* INTRODUCTION
* LIBRARIES USED
* SAMPLE CODE
* OUTPUT

**ABSTRACT:**

The primary objective of the program is to effectively scarp the “stackoverflow.com/tags” website and find the most number of tags posted on the site and also visualize the data.

**INTRODUCTION:**

**Web Scraping** is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

I am using python as programming language to write code for my project because it is easy to code with, has large no of libraries/packages. I am using the following below libraries from python for my project.

**LIBRARIES USED:**

* BeautifulSoup ( web scraping)
* Requests (http requesting )
* pandas (Dataframes)
* matplotlib (Data Visualization)

CODE:

**scraper.py**

import requests

from bs4 import BeautifulSoup

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib.style as style

df = {'Technology':[],'Tag\_Count':[]}

def extract\_tagged(url):

#print('Extracting Content')

content = requests.get(url)

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

for tag in soup.find\_all('a',attrs={'class':'post-tag'}):

df['Technology'].append(tag.text)

for count in soup.find\_all('span',attrs={'class':'item-multiplier-count'}): df['Tag\_Count'].append(count.text)

for i in range(1,3):

extract\_tagged('http://stackoverflow.com/tags?page='+str(i)+'&tab=popular')

df['Tag\_Count']=[int(i) for i in df['Tag\_Count']]

#end of scraper.py

**bar.py:**

#Bar Chart for Technologies and Tagged\_Count

from scraper import \*

df2= pd.DataFrame(df)

print(df2)

df2.sort\_values(by=['Tag\_Count'],ascending=False).head(10).set\_index('Technology').plot(kind='bar')

plt.title('10 Most Tagged Technologies On Stackoverflow.com')

plt.show()

#end of bar.py

**\* pie.py**

from scraper import \*

df4= pd.DataFrame(df,columns=['Technology','Tag\_Count']).head(20)

print(df4)

df4.plot(labels=df['Technology'],y='Tag\_Count',kind='pie',legend=False,autopct='%1.1f%%',startangle=90)

plt.title('20 Most Tagged Technologies On Stackoverflow.com')

plt.show()

# end of pie.py

**\* line.py**

from scraper import \*

df3= pd.DataFrame(df,columns=['Technology','Tag\_Count']).head(10)

print(df3)

df3.plot(x='Technology',y='Tag\_Count',kind='line')

plt.title('10 Most Tagged Technologies On Stackoverflow.com')

plt.show()

OUTPUT:

