|  |  |
| --- | --- |
| **Programming 1 (PRG1)**  Diploma in IT / DS / CSF / IM / CICTP  Year 1 (2023/24) Semester 1 | Week 15 |
| **Part 2** |
| **Activities: Applications of Programming II -  Regular Expressions, Web Scraping** | |

**OBJECTIVES**

At the end of this exercise, students should be able to:

* Use Regular Expressions to perform data validation and search/replace characters in a string.
* Use Web Scraping to extract data from web site.

Lecture Part 2

Activity 1 – Use Regular Expression to perform Data Validation

Write a program to:

* Prompt user to enter an Email
* Check if the email entered is a valid email address.
  + When the email entered is a valid, display the message “Valid email”.
  + Otherwise, display the message “Invalid email”.

Refer to the sample output:

Text

Description automatically generated

**Activity 2 – Collect email addresses and phone numbers from any web page using RegEx and Web Scraping**

Write a program to collect email addresses and phone numbers from any web page of your choice

Hint: Refer to the “**Example – Find email addresses”** in the slides.

The regular expression to find phone numbers can be: “\d{4}\s\d{4}”

**Activity 3 – Display academic calendar using Web Scraping**

Modify the following program to exclude the year 2022.

Hint: use row.text

# to install beautifulsoup,

# pip install beautifulsoup4

import requests

from bs4 import BeautifulSoup

def display\_calendar():

tables = soup.find\_all('table')

for table in tables:

rows = table.find\_all('tr') # find all rows in the table

for row in rows:

cells = row.find\_all('td') # find all <td> tags

for cell in cells:

print(cell.text) # get <td> text

session = requests.Session()

my\_headers = {"User-Agent":'Mozilla/5.0'}

url = 'https://www.np.edu.sg/admissions-enrolment/guide-for-prospective-students/academic-calendar'

response = session.get(url, headers=my\_headers)

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

display\_calendar()