Assignment 06

Due Date: 04/03/2022,11:59pm Name: Sajal Shrestha

In [ ]:
```
# imports
import regex as re
```

1. Identify, and extract each of the email-ids from a given text.

In [ ]:
```
texts = '''
Valid email addresses are:
user@email.com
firstname.lastname@email.com
user.name+22@email.com
fitst-last.22@email.com
username@domain-test.com

Invalid emails are:
first.last.domain.com
first@last@domain,com
'''

emails = re.findall("[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9.]*", texts)
print(emails)
```

```
['user@email.com', 'firstname.lastname@email.com', 'user.name+22@email.com',
'fitst-last.22@email.com', 'username@domain-test.com']
```

2. Identify, and extract each of the url/web-addresses from a given text.

In [ ]:
```
texts = '''
The urls are
http://domain.com, https://domain.com, https://www.domain.com, https://domain.c
http://domain.com/index.html, http://www.domain-site.edu/site.asp,
'''

urls = re.findall("http[s]?://[a-zA-Z0-9.\-/?=&%]+", texts)
print(urls)
```

```
['http://domain.com', 'https://domain.com', 'https://www.domain.com', 'http
s://domain.com/resource/?filter=10&test=hello%20world', 'http://domain.com/ind
ex.html', 'http://www.domain-site.edu/site.asp']
```

3. Identify, and extract each of the phone-numbers from a given text.

In [ ]:
```
text = "My name is John, I have three phone numbers: 222-333-5555, 111-222-3333

phone_numbers = re.findall("\d{3}[-]\d{3}[-]\d{4}", text)
print(phone_numbers)
```

```
['222-333-5555', '111-222-3333', '222-333-4444']
```

4. Identify, and extract each of the zip-codes (5,9) from a given text.

In [ ]:
```
text = "The zip codes are 49302 49322-2222 49222-1111"

zip_codes = re.findall("\d{5}-\d{4}|\d{5}", text)
print(zip_codes)
```

```
['49302', '49322-2222', '49222-1111']
```

5. Identify, and extract each of the dates from a given text.

In [ ]:
```
text = "The dates format can be 05/02/2022, 5/2/2022, 05-02-2022 and 12/01/2022

dates = re.findall("[0-9]{1,2}[/-]+[0-9]{1,2}[/-]+[0-9]{2,4}", text)
print(dates)
```

```
['05/02/2022', '5/2/2022', '05-02-2022', '12/01/2022']
```

6. Identify, and extract each of the ip-addresses from a given text.

In [ ]:
```
text = "The ip addresses are: 192.168.1.10 127.0.0.1 255.255.255.255"

ip_addresses = re.findall("\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}", text)
print(ip_addresses)
```

```
['192.168.1.10', '127.0.0.1', '255.255.255.255']
```

7. Identify, and extract each of the substrings with beginning with “Hello”, and ending with “Bye”.

In [ ]:
```
text = """
This is a sample: Hello, my name is Sam. I am testing regex. Bye This is a test
Hello again. This a test, Bye
"""
expression = r"Hello(.*)Bye"
pattern = re.compile(expression)

print(re.findall("Hello.*Bye", text))

substrings = re.findall("Hello(.*)Bye", text)

print(substrings)

for item in substrings:
    pattern = re.compile("\s+")
    x = pattern.split(item)
    print(x)
```

```
['Hello, my name is Sam. I am testing regex. Bye', 'Hello again. This a test,
Bye']
[', my name is Sam. I am testing regex. ', ' again. This a test, ']
[',', 'my', 'name', 'is', 'Sam.', 'I', 'am', 'testing', 'regex.', '']
['', 'again.', 'This', 'a', 'test,', '']
```

8. Identify, extract, and then replace all the symbols and digits (non-alphabets) in a given text.

In [ ]:
```
text = """This message contains symbols like $ # @ !@ and digits 1 2 3 43 4120'

pattern = "[^a-zA-Z ]+"
new_text = re.sub(pattern, "PLACEHOLDER", text)

print(new_text)
```

```
This message contains symbols like PLACEHOLDER PLACEHOLDER PLACEHOLDER PLACEHO
LDER and digits PLACEHOLDER PLACEHOLDER PLACEHOLDER PLACEHOLDER PLACEHOLDER
```

9. Identify, extract, and then replace all the numeric values (salary, price, age etc.) in a given text.

In [ ]:
```
text = """
Hello, my name is Sam. I am earning 2000. I am from Michigan and my age is 25.
I recently bought a phone for 500
"""

new_text = re.sub("[0-9]+", "PLACEHOLDER", text)
print(new_text)
```

```
Hello, my name is Sam. I am earning PLACEHOLDER. I am from Michigan and my age
is PLACEHOLDER.
I recently bought a phone for PLACEHOLDER
```

10. Split any given string/text on white-space(s)/tabs and replace (join them back) using a hypen '-'.

In [ ]:
```
text = """This is an example"""

pattern = re.compile("\s+")
split_text = pattern.split(text)

result = "-".join(split_text)
print(result)
```

```
This-is-an-example
```

11. Partition all the email-addresses extracted from task-1 into username, domain-name, and domain-suffix values.

In [ ]:
```
data = {
    "username": [],
    "domain name": [],
    "domain suffix": []
}

for email in emails:
    expression = r"(.+)@(.+)\.(.+)"
    pattern = re.compile(expression)
    match = pattern.match(email)
```

```
    username, domain_name, domain_suffix = match.groups()

    data["username"].append(username)
    data["domain name"].append(domain_name)
    data["domain suffix"].append(domain_suffix)

    print("Username: ", username, "Domain: ", domain_name, "Domain Suffix: ", d
```

```
Username:  user Domain:  email Domain Suffix:  com
Username:  firstname.lastname Domain:  email Domain Suffix:  com
Username:  user.name+22 Domain:  email Domain Suffix:  com
Username:  fitst-last.22 Domain:  email Domain Suffix:  com
Username:  username Domain:  domain-test Domain Suffix:  com
```

12. Write the output from task-11 to a csv file into three different columns.

In [ ]:

```
import pandas as pd

df = pd.DataFrame(data)
df.to_csv("email_data.csv", index=False)
```