Big Data Analytics

Lab Practical and date – Practical 1, Monday 19th July 2020

Name and Roll Number- Het Shah, 17BIT103

Practical Objective- Introduction to R/Python programming Language. Write program to read the data from any online website, excel file and CSV file.

Steps Involved-

We perform data scarping using python by reading data from different file format such as excel, csv and also performing web scarping by extracting data from a website

Background

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

Data scraping, also known as web scraping is the process of importing information from a website into a spreadsheet or local file saved on your computer.

File Formats Used

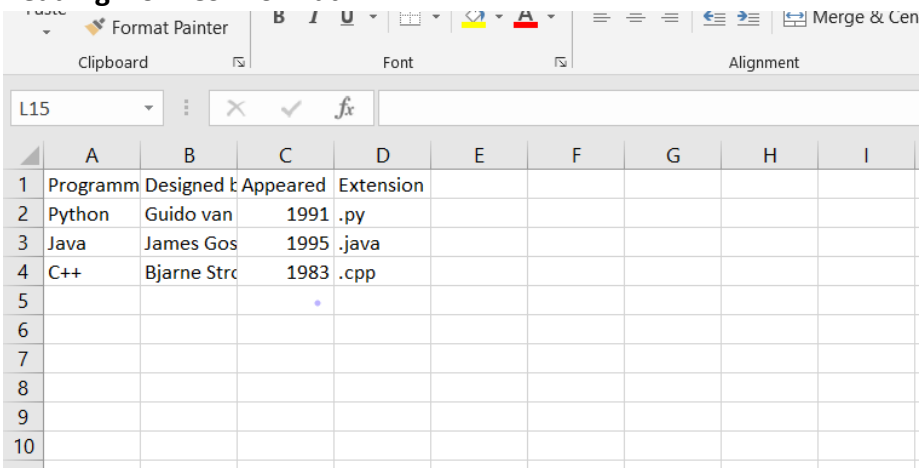
1) **Excel**- excel is spreadsheet application developed and maintained by MircroSoft. Excel organizes data in rows and columns format and they intersect at a space called cell.
2) **CSV**(comma separated values)-A minimal format compatible with many spreadsheet applications. Rows of data are represented by lines in text file, with columnar breaks delimited by a single character, usually a comma.

Libraries used-

1) **Pandas**-pandas is a software library written for the Python programming language for data manipulation and analysis
2) **Requests**- it is a python HTTP library, allows to send HTTP requests easily
3) **Csv**-The csv module implements classes to read and write tabular data in CSV format.
4) **Lxml.html**- lxml provides a very simple and powerful API for parsing XML and HTML. It supports one-step parsing as well as step-by-step parsing using an event-driven API

Outputs

1) **Reading from CSV format**



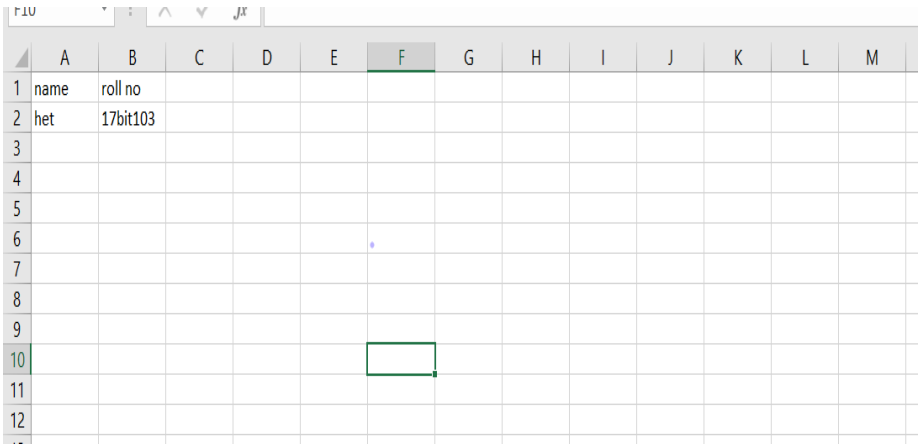
	A	B	C	D
1	Programm	Designed b	Appeared	Extension
2	Python	Guido van	1991	.py
3	Java	James Gos	1995	.java
4	C++	Bjarne Stro	1983	.cpp

This is the CSV file from which we want to extract the data

```
In [29]: runfile('E:/Desktop/prac 1.py', wdir='E:/
Desktop')
csv output
['ï»¿Programming language', 'Designed by',
'Appeared', 'Extension']
['Python', 'Guido van Rossum', '1991', '.py']
['Java', 'James Gosling', '1995', '.java']
['C++', 'Bjarne Stroustrup', '1983', '.cpp']
```

This is the python program reading the csv file and displaying it to the console

2) Reading from Excel Sheet



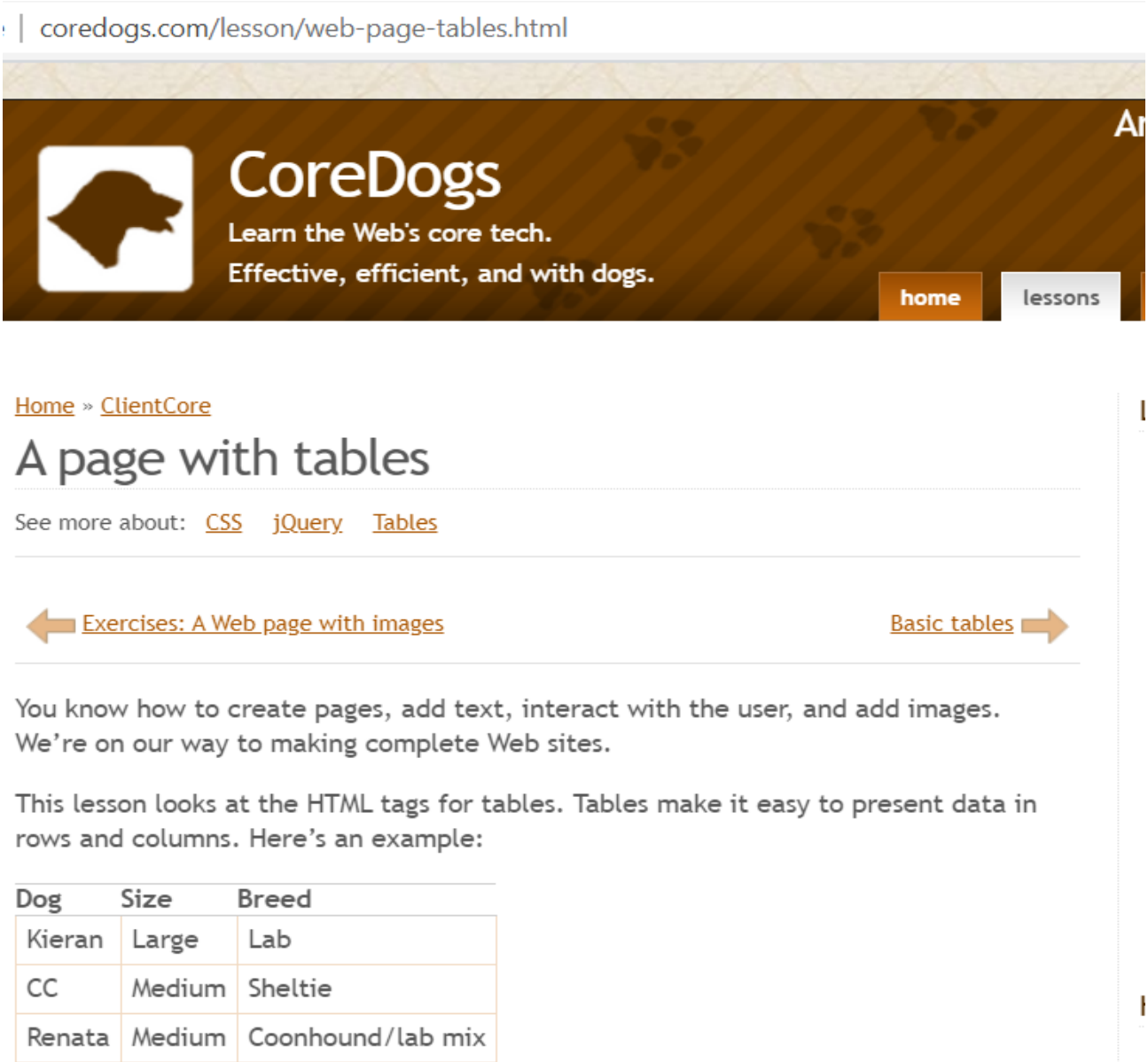
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	name	roll no											
2	het	17bit103											
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													

This is the Excel file from which we want to extract the data

```
excel output
  name    roll no
0  het   17bit103
```

This is the python program reading the csv file and displaying it to the console

3) Reading data from an online Website



This is the website http://coredogs.com/lesson/web-page-tables.html from which we extracted the table

```
table parsed
      Dog    Size              Breed
0  Kieran   Large                Lab
1      CC  Medium            Sheltie
2  Renata  Medium  Coonhound/lab mix
```

The python code was able to parse the table from the website and be able to display it

4) Running Queries on the Parsed Data

```
table parsed
      Dog    Size              Breed
0  Kieran   Large                Lab
1      CC  Medium            Sheltie
2  Renata  Medium  Coonhound/lab mix

selecting dog breed with medium size
1        CC
2    Renata
Name: Dog, dtype: object

selecting dog breed with name size more than 5
0    Kieran
2    Renata
Name: Dog, dtype: object

selecting distinct dog breed sizes
['Large' 'Medium']

insert a row - Sunny, small, pug
      Dog    Size              Breed
0  Kieran   Large                Lab
1      CC  Medium            Sheltie
2  Renata  Medium  Coonhound/lab mix
3   Sunny   Small                pug

delete  a row - with dog CC
      Dog    Size              Breed
0  Kieran   Large                Lab
2  Renata  Medium  Coonhound/lab mix
3   Sunny   Small                pug
```

In this step, we mainly performed CRUD(create, read, update and delete) and select operations using the pandas framework and we displayed the results as shown.

Conclusion

In this practical, we learned about data parsing from various file formats such as CSV and excel and also how to parse data from websites. Later, we used the parsed data and performed CRUD operations on them.

```python
# -*- coding: utf-8 -*-
"""
Created on Mon Jul 20 14:19:50 2020

@author: HETSHAH
"""


import csv
import pandas as pd
import requests
import lxml.html as lh

#read from csv
print('csv output')
with open('data.csv','rt')as f:
  data = csv.reader(f)
  for row in data:
        print(row)

print()
print()
print('excel output')
#read from excel
df = pd.read_excel (r'book1.xlsx')
print (df)



print()
print()

#read online

url='http://coredogs.com/lesson/web-page-tables.html'
page = requests.get(url)
doc = lh.fromstring(page.content)
tr_elements = doc.xpath('//tr')


tr_elements = doc.xpath('//tr')
col=[]
i=0

```

```python
#getting the table headings
print('heading names')
for t in tr_elements[0]:
    i+=1
    name=t.text_content()
    print('%d:"%s"'%(i,name))
    col.append((name,[]))


#storing the data
for j in range(1,len(tr_elements)):
    T=tr_elements[j]
    i=0
    for t in T.iterchildren():
        data=t.text_content()
        col[i][1].append(data)
        i+=1

print()
print('table parsed')
Dict={title:column for (title,column) in col}
df=pd.DataFrame(Dict)
print(df.head())

print()
print( 'selecting dog breed with medium size')
print(df[df.Size == 'Medium'].Dog)

print()
print( 'selecting dog breed with name size more than 5')
print(df[df.Dog.str.len() >=5].Dog)


print()
print( 'selecting distinct dog breed sizes')
print(df.Size.unique())

print()
print( 'insert a row - Sunny, small, pug')
df = df.append({'Dog': 'Sunny', 'Size': 'Small', 'Breed':'pug'}, ignore_index=True)
print(df)


print()
```

```
print( 'delete  a row - with dog CC')
df= df[df.Dog != 'CC']
print(df)

```