Lab assignment 1

Group C

→ Title: Create a review scrapper of any e-commerce website to fetch real time comments, reviews, ratings, comment tags, customer name using python.

→ Objective: To understand the application and impact of big data.

→ Software requirements:

a) Ubuntu 14.10 / 14.04

b) Python

→ Theory:

a) Web scrapping -

- Web scrapping is the process of gathering information from internet.
- Even copying and pasting the lyrics of your favourite song is a form of web scrapping.

Challenges of web scrapping -

- Variety: Every website is different while you will encounter general structure that repeat themselves. Each website is unique and you will need personal treatment if you want to extract the relevant information.
- Durability: Website constantly changes and say you have built a new web scrapper that automatically understands what you want from your resource of interact.

b) Steps for scrapping website -

Step 1:

1) Inspect your data source.
2) Decipher the information in URL's.
3) Every URL consist of two parts.
   i) the base URL represents the path to search functionality.
   ii) the specific site location that ends with html is the path to the job description unique resource.
   iii) every URL has 3 parts -
   eg. https://all.indeed.com/job a=software+development=Australia
   start : begining of the query
   information : pieces of information constituting query parameters are encoded in key-value pairs
   separators : every URL can have multiple query parameters separated by ampersan symbol

Step 2 : Scrap the html content from a page

1) Python requests library is used.
2) Install request library in shell.
3) One new file and implement below code -

```
import requests
URL = "HTTP://google.com"
page = request.get(URL);
print(page.text)
```

Step 3: Parse html code with beatiful soup

1) Install beautiful soup library

```
$ python -m pip install beatiful soup 4;
```

2) To import library in your python script and create beatiful soup object.

```
import requests
from bs4 import beatiful soup
URL = "https://google.com"
page = requests.get(URL)
soap = beatiful soap(page.content, "html.parser")
```

3) We can also find various elements by using id, class, name.

→ Conclusion: Thus, web scrapping is learnt and understood clearly throughout the assignment.