DSBDL Assignment.

Assignment 7

Title :- Visualize the data using R/python by plotting the graphs for assignment no. 6.

Objective :-

1. To understand & apply the Analytical concept of Big data using R/python.
2. To study detailed Concept R.

Software Requirements :-

1. Ubuntu 14.04 / 14.10
2. GNU C Compiler.
3. Hadoop
4. Java.
5. R studio.

Problem statement :-

Visualize the data using R/python by plotting the graphs for assignment no. 6 & 7.

Theory :-

Python provides various libraries that come with different features for visualizing data. All these libraries come with different features & can support various types of graphs. In this tutorial, we will be discussing four such libraries

- Matplot lib.
- Seaborn.
- Bokeh
- plotly.

* matplotlib :-

matplotlib is an easy-to-use, low level data visualizat library that is built on numpy arrays. It consist of various plots like scatter plot, line plot, histogram etc. Matplotlib provides a lot of flexibility.

* Scatter plot :-

scatter plots are used to observe relationships betn variables & uses dots to represent the relationship between them. The scatter() method in the matplotlib library is used.

* Line Chart :-

line chart is used to represent a relationship between two data X & Y on a different axis. It is plotted using the plot() function

* Bar chart :-

A bar chart is a graph that represents the category of data with rectangular bars with lengths & heights that is proportional to the values which they represent bar() method is used.

* Histogram :-

A histogram is basically used to represent data in the form of some groups. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives the information about frequency. The hist() function is used to compute & create a histogram

* Seaborn :-

Seaborn is a high level interface built on top the matplotlib. It provides beautiful design style & color palletes to make more attractive graphs.

* Bar plot :-

Bar plot in seaborn can be created using the barplot() method.

Conclusion :-

Thus we have learnt visualize the data using R/python by plotting the graphs.

DSBDA LAB

Assignment 1 (C)

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

objective ::
To understand the application & impact of big data.

Software Requirements ::
Beaultiful soup library.

Theorey ::

What is web scrapping ?

Web scraping is the process of gathering information from the internet. Even copying & pasting the lyrics of the your favorite song is a form of web scrappin

Challenges of web scraping -

A] Variety :
Every website is different. While you'll encounter general structures that appear repeat themselves each website is unique & will need personal treatmen if you want to extract the relevant information.

B] Durability :
websites constantly changes. Say you've built a shiny new web scraper that automatically cherry-picks what you want from your resource of interest.

steps for scrapping website.

step 1 :- i> Inspect your data source.

ii> Decipher the information in URLs.

iii> Every URL consist of two parts.

a> The base URL represents the path to the search functionality of the website.

b> The specific site location that ends with .html is the path to the job description's unique resource.

c> Every URL has three parts.

ex - https://au.indeed.com/jobs?q=software+develop &l=Australia.

i> start :- The beginning of the query

ii> information - The pieces of information constitut one query parameter are encoded in key-value pair.

iii> separators - Every URL can have multiple query parameters, separated by an ampersand symbol.

step 2 :- scrap the HTML content from a page.

i> Python's requests library is used.

ii> Install requests library.

In shell,

```
$ python -m pip install requests
```

iii> open new file & implement below code -

```
import requests.
URL = "Http://google.com"
page = requests.get(URL);
```

```
print (page-text)
```

step 2: Parse HTML code with Beautiful soup.

i> Beautiful soup is a python library for parsing structured data. It allow you to interact with HTML in a similar way to how you interact with a web page using developer tools. The library exposes a couple of intuitive functions you can use to explore the HTML.

install Beautiful soup:

```
$ python -m pip install beautifulsoup4.
```

ii> Then import the library in your python script & create beautiful soup obj.

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

iii> We can find various elements by using id, class names.

Conclusion:

Thus web scrapping is learnt & understood clearly through out the assignment.