14-12-2020

Chapter-4
Collecting Text Data from Web.

Objectives :-

- Extract and process data from web pages
- Describe different kinds of semi-structured data, such as JSON and XML
- Extract real time data using Application programming Interfaces
- Extract data from various file format.

⇒ Semi Structured Data

1.) JSON
Java Script object notion

- Data is stored as key-value pair
- Datatype of values can be :
  - A string
  - A number
  - Another JSON object
  - An array
  - A Boolean
  - Null.

DATE
PAGE No

16-12-2020

2) XML

Extensible markup language

- Human readable and extensible
- May or may not have a declaration, but if there's a declaration, it has to be the starting line
- The declaration has three parts → version, encoding, standalone.

- Can be represented as a tree called XML tree.

⇒ Using API's to retrieve real time data.

Application programming interface

~~Some~~

Some websites do not provide their data directly. Instead they provide API's using which we can extract data from them.