# PRINCIPLES OF DATA VISUALIZATION

## Unit 1

ABSTRACT

This course provides insight on data visualization, the art and science of turning data into readable graphics; design and create data visualizations based on data available and tasks to be achieved; data extraction, data modelling and data processing; map data attributes to graphical attributes, and strategic visual encoding based on known properties of visual perception.

**MAHE-MSIS-RAGHUDATHESH G P**

BDA 5132 – Common for AI&Ml and BDA

**Introduction to Web Scraping:** Web scraping models and techniques, Case study: BeautifulSoup, Scrapy, Selenium. **(4 Hrs.)**

***BY:***

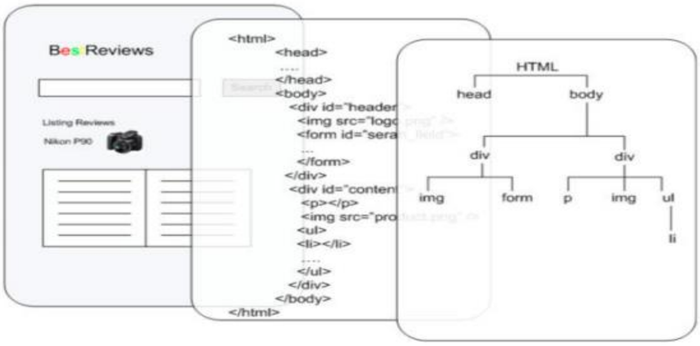
**RAGHUDATHESH G P, PRATHVIRAJ N**
**Assistant Professor – Senior Scale**
**Manipal School of Information Sciences (MSIS),**
**MIT Campus, MAHE, Manipal – 576104**
**Mail:raghudathesh.gp@manipal.edu, Prathviraj.n@manipal.edu**
**Website: raghudathesh.weebly.com**

## Quotes:

- Do good, it will come back to you in unexpected ways.
- 6 + 2 = 8 but so does 5 + 3. The way you do things is not always the only way to do them. Respect other people's way of thinking.
- We make a living by what we get. We make a life by what we give.
- Luxury and Lies have huge maintenance costs. However, Truth and Simplicity are self-maintained without any cost.
- The true mark of maturity is when somebody hurts you and you try to understand their situation instead of trying to hurt them back.

# 1 Introduction to Web Scrapping:

- Web Scraping is technique used for extracting unstructured data from the websites and transforming that data into structured entity.
- Web Scraping is also identified as web data extraction, web data scraping, web harvesting or screen scraping. Web scraping is a form of data mining.
- The basic and important aim of the web scraping process is to mine information from a different and unstructured websites, transform it into an comprehensible structure like spreadsheets, database, a comma-separated values (CSV) file.
- Data like item pricing, stock pricing, different reports, market pricing and product details, can be gathered through web scraping. Extracting targeted information from websites contributes to take effective decisions in business process.
- The first widely known scrapers were invented by search engine developers (like Google, AltaVista, and now Bing). These scrapers go through (almost) the whole Internet, scan every web page, extract information from it, and build an index that you can search.
- Web crawling creates a copy of what's there and web scraping extracts specific data for analysis, or to create something new.
- Web scraping is essentially targeted at specific websites for specific data, e.g. for stock market data, business leads, supplier product etc.
- Scraping is a technique used to crop information from web pages based on script routines. Web pages are documents written in Hypertext Markup Language (HTML), and more recently XHTML which is based on eXtensible Markup Language (XML). Web documents are represented by a tree structured called the Document Object Model, or simply the DOM tree and the goal of HTML is to specify the format of text displayed by Web browsers as shown in figure 1



```
<html>
      <head>
      ....
      </head>
      <body>
        <div id="header">
          <img src="logo.png" />
          <form id="serach_field">
          ...
          </form>
        </div>
        <div id="content">
          <p></p>
          <img src="product.png" />
          <ul>
          <li></li>
          ....
          </ul>
        </div>
      </body>
</html>
```

Figure 1.1: Three different outlook of web document - The document on web, the HTML code and the Document Object Model

- From the operation viewpoint, a web scraping look like manual copy and paste task. The difference here is that this job is done in a organized and automatic way, by a virtual computer script/agent.
- When an agent is following each link of a web page, it is actually performing the same operation that a human being would normally do when interacting with a web site. This agent can follow links (by issuing HTTP GET requests) and submit forms (through HTTP POST), browsing through many different web pages.
- Next step is, the parser follows user-specified paths inside the document to retrieve the desired information based on the data retrieved in previous step. These paths are specified by CSS selectors or XPATHs. They use both relative and absolute paths (based on the DOM tree) to point the parser to a specific element inside a web document.
- Normally web-scraping operations uses regular expressions to narrow or trim the located information, in order to retrieve data with a user specified granular size. This process is illustrated in figure 2.

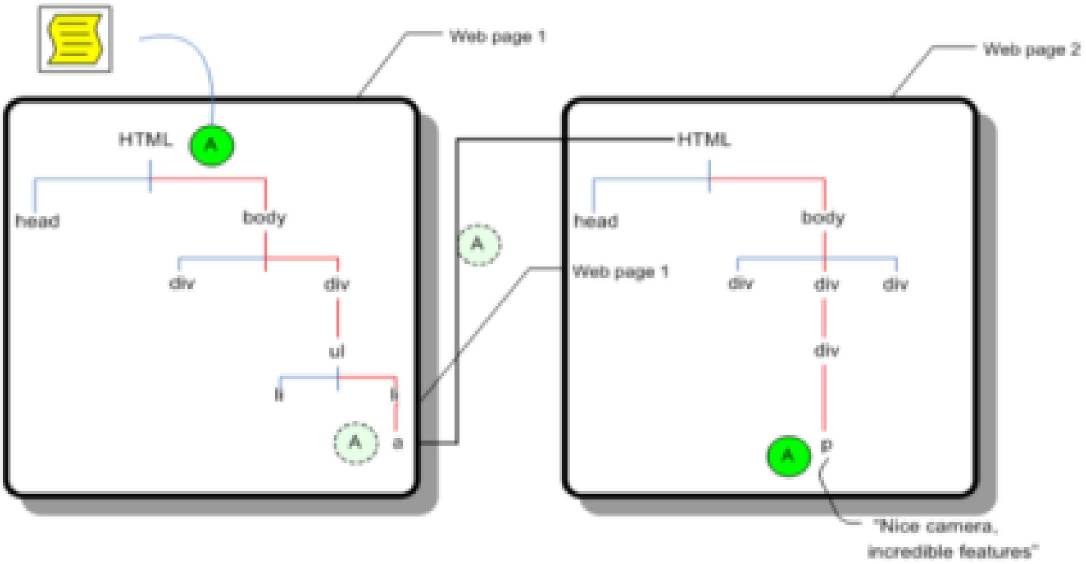


Figure 1.2: Last step of web scraping- Gathering information by scraping agent

- A web scraping agent gathering information from web pages - The dotted circle represents a web scraping agent traversing a DOM tree. The red lines are XPATHs to a desired element within the document. The agent reaches the hyperlink in web page 1 and proceeds to web page 2 until it finds the information enclosed by element p (paragraph).

# 2 WEB SCRAPER USES

- Web Scrapers are also being used by Online Marketers to pull data privately from the competitor's websites such as high targeted keywords, valuable links, emails & traffic sources. Some of the area where web scraper techniques are mostly uses are:
  1. Change detection on website
  2. Product Price comparison
  3. Weather broadcasting and data monitoring
  4. Research analytics
  5. Machine Learning applications to build data sets
  6. Analyze data in graphics
  7. Web Indexing & rank checking
  8. Advertisement analysis
  9. Market Analysis

# 3 WEB SCRAPING TECHNIQUES

1. **Classical copy and paste:**
   - The human's manual examination and copy and-paste method is the best and the workable web scraping technology.
   - But it appears to introduce or contribute to mistakes, with the user needing to analyze and store loads of data sets using tiresome technology.
2. **Hypertext Transfer Protocol (HTTP) Programming:**
   - By using this technique user can be extract information from static and dynamic web pages.
   - Data can be retrieved by posting HTTP requests to the remote web server using socket programming.
3. **Hyper Text Markup Language (HTML) Parsing:**
   - Semi-structured data query languages, like the Hyper Text Query Language (HTQL), can be used to parse HTML pages and to retrieve, transform page content.
4. **Document Object Model (DOM) Parsing:**
   - By embedding a full-fledged web browser, such as the Chrome or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts.
   - These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages.

5. **Web Scraping Software:**
   - Now a days many software tools are available, that can be used to customize web-scraping solutions.
   - This software may attempt to automatically recognize the data structure of a page or provide a recording interface that removes the necessity to manually write web scraping code, or some scripting functions that can be used to extract and transform content, and database interfaces that can store the scraped data in local database.
6. **Computer vision web-page analyzers:**
   - There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might.

# 4 WEB SCRAPING TOOLS

- Web Scraping Software are the computerized program that are used to make the manual copy paste work automatically.
- It also collect large amount of information from websites like directory sites, real estate sites, classified websites and job boards and stored on local server.
- Suppose you want to scrape real estate property details of India then you need to appoint few guys to copy and paste details from websites to excel by visiting each property agent pages. This way it will take days and even months to get your property data ready to use.
- Web scraping can automate the manual work programmatically by visiting each page and extract data from pages and parsing the html pages.
- There are number of Web Scraping Software that available in market that can help you to scrape data from any website you want.
- Web scraping software work as, to bot or web crawler access the web data directly using the Hypertext Transfer Protocol, or through a web browser and extract the precise data from that web page. This extracted data is then store into a central local database or spreadsheet for later use or analysis.
- Following are the list of some scraping tools:
- **Mozenda:**
  - ➢ It is a Business Intelligence Software. It is allows user to extract data from documents as the same way user can extract data from web page.

  - Also user can be combine data from multiple sources into a single data set.
  - Mozenda currently support documents scraping for several popular formats like World, Excel, PDF etc.
  - Mozenda will automatically detect names and associated values and build robust data sets with minimal configuration.
- **Web Content Extractor (WCE):**
  - It has good wizard that guide user to setup scraper.
  - User can scrape data from website with few clicks and WCE is self intelligent for putting data into different formats like Excel, text, HTML formats, Microsoft Access database, Structured Query Language(SQL) Script File, MySQL Script File, Extensible Markup Language (XML) file, HTTP submit form and Open Database Connectivity (ODBC) Data source.
- **Import.io**
  - Import.io is a web based tool for extracting data from website without writing code.
  - If user want a fast result then he/she will try for this tool so that he can convert website in short time.
  - For extracting data, user enter URL and application automatically extract data which user want to need, if user does not interested in the automatic extraction, the point and click interface help to select data fields on website.
  - As the data extraction is over, the extracted dataset is store on Import.io cloud server and farther downloaded in CSV, Excel, JSON format.
- **Scrapy**
  - An open source and collaborative framework for extracting the data you need from websites.
  - Scrappy is designed to scrape web content from sites that are composed of many pages of similar semantic structure.
  - The system is implemented as a Firefox browser extension, and works in three main stages to scrape web data.
    - First, a user navigates to a page that he would like to scrape and generates a template for the content that he would like from that page.
    - Next, the user selects a set of links that point to pages matching the content template defined by the user.

        - Finally, the user selects an output data format and Scrappy crawls the links specified by the user and scrapes content corresponding to the user’s template.
    - Scrapy written in Python and runs on Linux, Windows, and Mac.
- **BeautifulSoup:**
    - Beautiful Soup is a Python library for getting data out of HTML, XML, and other markup languages.
    - Say you’ve found some webpages that display data relevant to your work/research, such as date or address information, but that do not provide any way of downloading the data directly. Beautiful Soup helps you pull particular content from a webpage, remove the HTML markup, and save the information.
    - It is a tool for web scraping that helps you clean up and parse the documents you have pulled down from the web.
- **Selenium:**
    - Selenium is an automation testing framework for web applications/websites which can also control the browser to navigate the website just like a human.
    - Selenium uses a web-driver package that can take control of the browser and mimic user-oriented actions to trigger desired events.
    - Selenium is a Python library and tool used for automating web browsers to do a number of tasks. One of such is web-scraping to extract useful data and information.
    - Selenium is the perfect tool to automate web browser interactions where dynamic websites load the data from a data source.

# 5 Steps to automate web scraping:

- Automating web scraping involves setting up a process that automatically extracts data from websites at regular intervals or as needed. Here are the general steps to automate web scraping:
    1. Identify the target website:
        - Determine which website or websites user want to scrape data from.
        - Check if there are any legal restrictions or terms of service that prohibit web scraping.
    2. Write a scraper:
        - Use programming languages, such as Javascript or Python, to extract the data from the website.
        - The code will usually involve sending requests to the website, parsing the HTML content, and extracting the relevant data using CSS selectors or XPath expressions.
    3. Host your scraper with a dedicated server:
        - Use a hosing service to host your scraper script on a server. This step ensures that your scraper is always running and collecting data, even if your computer is turned off.
        - Running a scraper may require high computing power, and hosting it on a dedicated server is faster and more efficient.
            - Ex: Acho,
    4. Set up a scheduler for your scraper:
        - Use a task scheduler, such as Crontab, to schedule the scraper to run at regular intervals or as needed.
    5. Store the scraped data:
        - Once the data has been extracted, store it in a file or database for later use, or process it further, depending on your requirements.
    6. Build an application to utilize the data:
        - Once automating your scraper to collect data, you can build an interface based on your scraped data.
        - Here are some examples of applications or analyses that you can build:
            - **Sentiment analysis:** Web scraping can be used to extract customer reviews, feedback, and social media posts related to a

particular product or service. This data can then be analyzed using sentiment analysis techniques to gauge customer sentiment and improve customer experience.

- **Predictive modeling:** Web scraping can be used to extract historical data on sales, customer behavior, or other relevant data points. This data can then be used to create predictive models that can help businesses forecast future trends, identify potential risks, and make informed decisions.
- **Competitive analysis:** Web scraping can be used to extract data on competitors, such as pricing, product information, and marketing strategies. This data can then be analyzed to identify competitive advantages and develop effective strategies.
- **Marketing analytics:** Web scraping can be used to extract data on website traffic, search engine rankings, and social media metrics. This data can then be analyzed to identify patterns and trends, and develop effective marketing strategies.
- **Supply chain optimization:** Web scraping can be used to extract data on suppliers, shipping times, and inventory levels. This data can then be analyzed to optimize supply chain operations and reduce costs.