STATISTICS FOR DATA SCIENCE

Reading Files & Web Scraping

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Reading Files

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Reading Files

For any data scientist or data engineer, dealing with different formats can become a tedious task.

In real-world, people rarely get neat tabular data. Thus, it is mandatory for any data scientist (or a data engineer) to be aware of different file formats, common challenges in handling them and the best / efficient ways to handle this data in real life.

As a data scientist, you need to understand the underlying structure of various file formats, their advantages and disadvantages.

Unless you understand the underlying structure of the data, you will not be able to explore it. Also, at times you need to make decisions about how to store data.

Source: https://www.analyticsvidhya.com/

Reading from .CSV file

Reading data from CSV(comma separated values) is a fundamental necessity in Data Science.

Often, we get data from various sources which can get exported to CSV format so that they can be used by other systems.

The Pandas library provides features using which we can read the CSV file in full as well as in parts for only a selected group of columns and rows.

How to read .CSV Files

Click for Demonstration of reading .csv files.

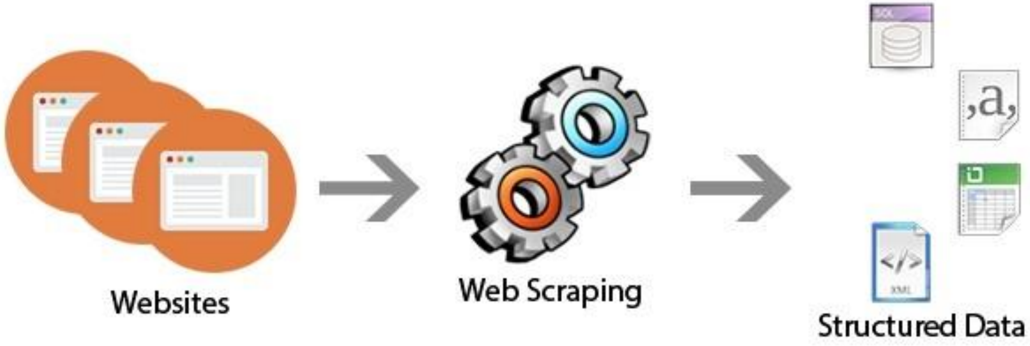
STATISTICS FOR DATA SCIENCE

Web Scraping

What is Web Scraping?

- **Web scraping** is the process of gathering information from the Internet.

- The incredible amount of data on the Internet is a rich resource for any field of research or personal interest.



Source: realpython
Image Source: Quora.com

Extraction of Data from Website

There are mainly two ways to extract data from a website:

- Use the API of the website (if it exists). For example, Facebook has the Facebook Graph API which allows retrieval of data posted on Facebook.

- Access the HTML of the webpage and extract useful information/data from it. This technique is called web scraping or web harvesting or web data extraction.

Source:DataCamp

Why Web Scraping?

- Data displayed by most websites can only be viewed using a web browser.
- They do not offer the functionality to save a copy of this data for personal use.
- The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete.
- Web Scraping is the technique of automating this process, so that instead of manually copying the data from websites, this will perform the same task within a fraction of the time.

Source: https://www.webharvy.com/

What is done with Web Scraping!!!!!

- Web scraping is typically used for different reasons like change detection, market research, data monitoring, and in some cases, theft.
- Web developers sometimes scrape their own websites when testing for broken links and images within each page
- Scraping can also done for unlawful purposes, such as copying a website and republishing it under a different name.
- This type of scraping is viewed as a copyright violation and can lead to legal prosecution.
- Web scraping is considered malicious when data is extracted without the permission of website owners.

Steps involved in Scraping

Steps involved in web scraping:

The web scraping process follows the below 3 steps.

1. Request-Response
2. Parse and Extract
3. Transform the data

Third-party python library called **Beautiful Soup** is used for pulling data out of HTML and XML files.

Source:RealPython

Is it Legal to Scrape?

- Not all web scraping acts are considered as legal.
- Python Web scraping services that extract **publicly available data are legal**.
- But, at times, it may cause legal issues, just the way any tool or technique in the world can be used for good as well as for bad.
- For example, web scraping non-public data, which are not accessible for everyone on the web, can be unethical, and also it can be an invitation to legal trouble.

- Check robots.txt – displays the pages that can be scraped.

Source: intellipad.com

Applications

- Search Engines
- Price Monitoring
- Sales and Marketing
- Content Aggregators
- Sales intelligence
- Training datasets for Machine Learning
- Data for Research

How is Web Scraping Done??

- Web Scraping can be done efficiently using Python because it is flexible and powerful.

- Includes Python libraries like request and BeautifulSoup4 which helps us to fetch URL and pull out information from web pages.

- In addition, re, numpy and pandas could help us clean and process the data.

Source: https://www.octoparse.com/

Demonstration

Lets do it practically.

Click here for Demonstration of Web Scraping of an Amazon Product Review.

To do - Assignment

Scrape the following for any product of your wish.

1) Rating for the product.
2) Review Content of the product (along with image if any) and date of review.
3) Comments for that review by other users.

THANK YOU

Dr. Uma D
Prof. Silviya Nancy J

Department of Computer Science and Engineering