TOP BOOKS

Basic Plots

Data Collection ,

WEB SCRAPING USING KNIME
STATISTICS USING JASP

Web Scraping is an automatic way to Collecting and parsing raw unstructured data From a websites and store them in a Structured format

Web Scraping is used for getting data. Access to relevant data, having methods to analyze it and performing intelligent actions based on analysis can make a huge difference in the success and growth of most businesses in the modern world.

WEB SCRAPING PROCESS

- Identify the target website
- Collect URLs of the pages where you want to Extract data from.
- Make a request to these URLs to get the HTML of The page
- Use locators to find the data in the HTML
- Save the data in a JSON or CSV file or some other Structured format

The following are few of the many uses of Web Scraping:

1. In e- Commerce, Web Scraping is used for competition price monitoring.

2. In Marketing, Web Scraping is used for lead generation, to build phone and email lists for cold outreach.

3. In Real Estate, Web Scraping is used to get property and agent/owner details.

4. Web Scraping is used to collect training and testing data for Machine Learning projects.

WEB SCRAPING using KNIME

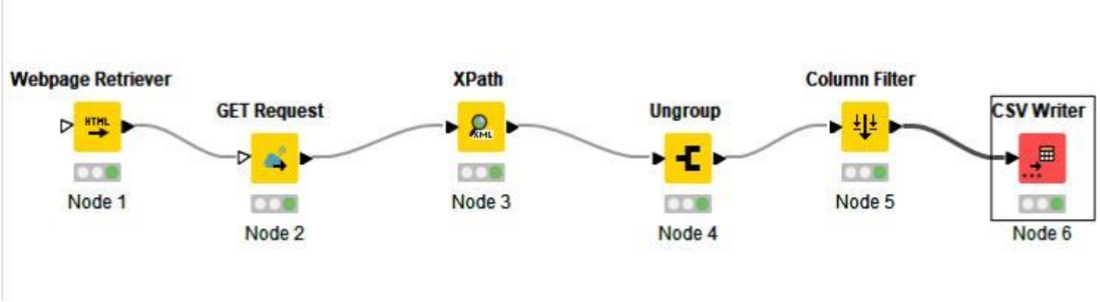
Here we used KNIME software for Web scraping (data scraping)

- Website –www.theguardian.com

•URL - https://amp.theguardian.com/books/2019/sep/21/best-books-of-the-21st-century

We are scrapped Top Books List using **KNIME SOFTWARE**..

The Used Nodes are,



•WEB Retriever [Node1]

This node can be used to retrieve webpages by issuing HTTP GET requests and parsing the requested HTML webpages. For parsing, soup is used as library which implements the WHATWG HTMLS specification. The parsed HTML will be cleaned by removing comments and, optionally, replacing relative URLs by absolute ones Node Workflow Coach

By default, the output table will contain a column with the parsed HTML converted into XHTML. However, you can specify to get the parsed HTML as string output instead node recommendations only available with usage data Reporting. The node allows you to either send a request to a fixed URL (which is specified in the dialog) or to a list of URLs provided by an optional input table. Every URL will result in one request which in turn will result in one row in the output table. You can define custom request headers in the dialog.

•Get Request [Node2]

This node can be used to issue HTTP GET requests GET requests are used to retrieve data from a web service without sending any data other than (optional) request parameters the node allows you to either send a request to a fixed URL (which is specified in the dialog) or to a list of URLS provided by an optional input table Every URL will result in one request which in turn will result in one row in the output table. You can define custom request headers in the dialog node recommendations only available with usage data reporting. By default the output table will contain a column with the received data, its content type, and the HTTP status code. The node tries to automatically convert the received data into a KNIME data type based on its content type.

•XPATH [Node3]

The node takes the XML Documents of the selected column and performs XPath queries on them. The node supports XPath 1.0

•UNGROUP [Node4]

Creates for each list of collection values a list of rows with the values of the collection in one column and all other columns given from the original Row. Rows with an empty collection are skipped, as well as rows that contain only missing values in the collection cell with the “Skip missing values option enabled.

•COLUMN FILTER [Node5]

This node allows columns to be filtered from the input table while only the remaining columns are passed to the output table. Within the dialog. columns can be moved between the Include and Exclude list.

•CSV WRITER [Node6]

This node writes out the input data table into a file or to a remote location denoted by an URL The input data table must contain only string or numerical columns. Other column types are not supported. This node can access a variety of different tile systems More information about file handling in KNIME can be found in the official File Handling.

STATISTICAL ANALYSIS Using JASP

Statistical analysis is the collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modeling or designing surveys and studies.

Here, We were used Scraped data set for Statistical Analysis by

Using "BASICPLOTS"

Basic Plots,

- Distribution Plots
- Q-Q Plots
- Correlation Plots
- Display Density