Bharatiya Vidya Bhavan's
SARDAR PATEL INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated To University Of Mumbai
Munshi Nagar, Andheri (W) Mumbai 400 058

C.I.T.L. EXPERIMENT 8

Submitted By:

Akash Panicker	2021300089
Mahesh Patil	2021300095
Rohit Phalke	2021300100
Adwait Purao	2021300101

Submitted To:
Prof. Sunil Ghane

Inventory Management System

Aim:

Demonstrate the behavior of Web Crawlers/ spiders (use XPATH,CSS PATH),extract information and store it in the database.

Problem Statement:

Develop an inventory management system for a retail store that efficiently tracks and manages the inventory of products. The system should provide real-time updates on stock levels, generate alerts for low stock items, enable easy addition and removal of products, and offer insights into sales trends to optimize restocking decisions.

Theory:

Web Crawler Bots:

- Web crawlers, also known as spiders or search engine bots, are software programs that systematically browse the internet to download and index content from web pages.
- The primary purpose of web crawlers is to gather information about web content, which is later used to provide relevant search results when users query search engines like Google or Bing.
- Web crawlers are mainly operated by search engines and are responsible for organizing and cataloging the vast amount of information on the internet.
- They follow hyperlinks from one webpage to another, effectively "crawling" the web and discovering new content.

Search Indexing:

- Search indexing is the process of creating a structured database (index) of web content.
- It allows search engines to quickly retrieve relevant information in response to user search queries.
- Indexing involves analyzing the text and metadata of web pages, including titles, descriptions, and keywords.

- Search engines add the words on web pages to their index, excluding common stop words such as "a," "an," and "the."

How Web Crawlers Work:

- Web crawlers start from a seed, a list of known URLs, and then crawl the webpages at those URLs. They discover new URLs from the pages they visit and add them to their list for future crawling.
- Crawlers prioritise pages based on factors like the number of inbound links, visitor traffic, and other indicators of a page's importance and relevance.
- They revisit web pages periodically to ensure the index reflects the latest content updates.
- Web crawlers also respect the robots.txt protocol, a text file that specifies which pages bots can or cannot crawl on a website.

Web Crawler Access to Web Properties:

- Decisions about whether to allow or disallow web crawlers to access web properties depend on the website owner's preferences and technical considerations.
- Web crawlers consume server resources and bandwidth, so website operators may limit indexing to avoid overloading their servers or incurring excessive costs.
- Some web pages are intentionally blocked from indexing by adding "no index" or "disallow" tags in the page or the robots.txt file.

Difference Between Web Crawling and Web Scraping:

- Web crawling is the automated process of systematically browsing the internet to index web content for search engines.
- Web scraping involves extracting specific data from web pages without permission, often for purposes like data extraction or scraping content for malicious use.
- Web crawlers follow links and continuously discover new content, while web scrapers are more focused on specific pages or websites.
- Web crawlers respect server resource limitations and follow robots.txt rules, whereas web scrapers may not consider server load and may disregard rules, potentially overloading web servers.
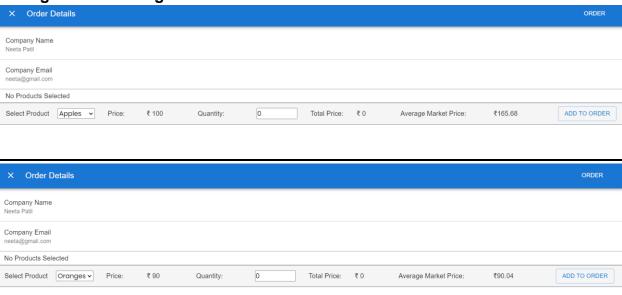
Code:

```javascript
const cheerio = require('cheerio');
const fuzzyset = require('fuzzyset');


router.get('/prices', async (req, res) => {
  try {
    // console.log("Inside Prices");
    const url = 'https://www.numbeo.com/cost-of-
living/country_result.jsp?country=India';

    // Attempt to make the request
    let response;
    try {
      response = await axios.get(url);
      console.log("Response Status:", response.status);
    } catch (error) {
      console.error("Error making request:", error.message);
      throw error; // Rethrow the error to trigger the catch block below
    }

    const $ = cheerio.load(response.data);

    // Update the selector according to the structure of the webpage
    const itemSelector = '.tr_highlighted';
    const priceSelector = 'td.priceValue';

    const prices = [];

    $('.data_wide_table tr').each((i, element) => {
      const item = $(element).find(itemSelector).text().trim();
      const price = $(element).find(priceSelector).text().trim();
      console.log("Item: Price:", item, price)
      if (item && price) {
        prices.push({ item, price });
      }
    });

    // console.log("Prices:", prices);
```

```
    await PriceModel.insertMany(prices);
    res.json(prices);
  } catch (error) {
    console.error("Error:", error.message);
    res.status(500).json({ message: error.message });
  }
});
```

Implementation Screenshots:

We get the Average Market Price of the Products from the data we fetched through web crawling.

Website Used For Web Crawling:
https://www.numbeo.com/cost-of-living/country_result.jsp?country=India

Conclusion:

In conclusion, web crawlers play a vital role in content discovery and indexing on the internet. Website owners should be aware of how these automated agents operate to ensure that their web content is effectively accessed and indexed. Monitoring web crawler behavior, adhering to best practices, and considering ethical and legal aspects are key to maintaining a harmonious online ecosystem.