Technical Document

Web Scraper

**Introduction**

Web crawling is a technique to collect data from the web by finding all the URLs for a domain. It is a process of collecting information from different websites on the web in an automated process. This process is also called as data extraction, content scraping, data scraping, web crawling, data mining and data collection.

Web crawling is part of web scrapping and web scarping is the extraction of data like URLs, prices, what all the useful information

In general, huge chunks of data helps the organizations to understand the market and gain a useful edge over the competitors. Web scraping is used in the ecommerce industry a lot. The data collected using the web scraping can be used for

* Market Research
* Training Machine Learning Algorithms
* Price Intelligence
* Trend Analysis
* Advertisements
* E-Commerce
* Predictive Analysis

**Objectives**

Azure marketplace is one of the online stores that contains thousands of software applications and services that you can try, buy, and deploy. The catalogue includes solutions for different industries and the apps are segregated into categories each and we can browse these apps specifically for each category by going to Homepage -> More -> Apps

<https://azuremarketplace.microsoft.com/en-us/home>

We need to build a tool to get all the apps details like category, sub-category, name, URL, costing, ratings, reviews etc. of each app and store the details in a spread sheet or a database based on the requirement.

**Tools and Packages**

* Python
* Excel
* Python Packages: requests, pandas, BeautifulSoup
* Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.
* DataFrame is a 2-dimensional labelled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most used pandas object.

**Approach**

* Install all the required packages listed above using pip.
* Using python requests get the contents of the main page.
* With BeautifulSoup, parse the response HTML content and get the Categories list using the class ‘*spza\_filterContainer*’
* Create a list with URLs of all the Categories.
* Loop through the Categories URLs list to fetch all the subcategories in each category and create a list of URLs of all sub-categories.
* For each subcategory get the URLs of all pages using the class tag ‘*filteredGalleryContainer*’ and create a dictionary with the subcategory and apps info as <key,value> pair
* Parse through each page item in the dictionary and get the URL of each app and append into a list.
* Create a data frame with all the info figured out till now ( like Category, Subcategory, App Title, App URL, Short Description)
* Parse through the list all All URLs and make a GET request to fetch the content like detailed description, pricing, and review details.
* Append all the data to the dataframe
* Write the data frame to a excel sheet