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.
* Using BeautifulSoup parse the response HTML content and get the *Browse apps* url
* Using above url make a request and get all the apps and the left pane has all the categories using BeautifulSoup
* Store all the categories in a list to be used later
* Parse the list of the categories and get the list of all sub-categories and store in a list/dict
* For each subcategory get the url with the page numbers and store them in a list.
* Now parse through each of the link stored in the above step and retrieve all the apps specific to a subcategory page and scrap the data of each url.
* Identify the tags to extract the information like Overview, Plans + Pricing, Rating + Reviews for each of the apps.
* Create a data frame using pandas and define the columns that must be used to identify the data that needs to be stored.
* Write the data frame to Excel sheet or Database based on the requirement.