


Report: Understanding the CSV Data

Overview:

This CSV file provides detailed information about each web page scraped from the website <https://franchisesuppliernetwork.com>. It includes the page URL, the types and counts of resources referenced on each page, and the amount of text extracted from each page.

Please read this note before:  Note

Report Sheet(CSV Data):  Report

Columns Description:

1. **page**: The URL of the scraped web page.
2. **resources_referenced**: A dictionary detailing the types and counts of resources referenced on the page. These resources can include:
 - **page_urls**: Links to other pages on the website.
 - **png**: PNG image files.
 - **jpg**: JPG image files.
 - **svg**: SVG image files.
 - **pdf**: PDF documents.
 - **gif**: GIF image files.
 - **webp**: Video files.
3. **text_extracted**: Total number of texts paragraph extracted from each <p> tag

How to Understand the Data:

1. **page**:
 - This column lists the URLs of the web pages that were scraped. Each entry corresponds to a specific page on the website.
2. **resources_referenced**:
 - This column provides a breakdown of the resources each page references. The resources are listed in a dictionary format, indicating the type of resource and the number of occurrences.
3. **text_extracted**:
 - This column indicates the amount of text content extracted from the page under each <p> tag within the page.

Example Rows Explanation:

1. **Homepage (/homepage/)**:
 - **Resources Referenced**: {'page_urls': 60, 'png': 33, 'jpg': 17, 'svg': 10, 'jpeg': 6, 'webp': 3}
 - The homepage references 60 other pages, 33 PNG images, 17 JPG images, 10 SVG images, 6 JPEG images and 3 WEBP Video
 - **Text Extracted**: 24 units
 - 24 units of text under <p> tag were extracted from the homepage.

Conclusion:

The CSV file provides a comprehensive overview of the web scraping results, detailing the resources referenced by each page and the amount of text extracted. This data can be used to understand the structure and content of the website, as well as the distribution of resources across different pages.