Anote

A CASE STUDY

Data Scraping for Hardy Riggings

CUSTOMER Hardy Riggings
TASK TYPE Data Scraping

INTRODUCTION

In this case study, we will explore the process of scraping data from 79 URLs associated with Hardy Riggings. Our objective is to extract valuable information such as product titles, descriptions, variants, prices, product manuals, and spec sheets. The extracted data will be formatted into a CSV file suitable for importing into WooCommerce, a popular WordPress plugin for e-commerce.

SYSTEM COMPONENTS & FUNCTIONS

1 SCRAPING TEXT FROM URLS

We utilized the Python library "tika" along with the "requests" module to retrieve the content from each URL. By using Apache Tika, we ensured proper handling of various document formats.

2 CHUNKING TEXT & GENERATING

The text obtained from each URL was chunked and converted into embeddings. This step helps provide context to the AI model, enabling it to generate more accurate predictions.

3 PROMPTING THE MODEL FOR INFORMATION EXTRACTION

We prompted an AI model with specific questions for each URL to extract the desired information. For example: What is the product title?

4 CSV OUTPUT GENERATION

The model's responses to our prompts were compiled into a CSV file. Each row in the CSV file represents a URL, while each column corresponds to a specific question. The cells contain the AI model's predicted answers for the respective URL and question.

CONCLUSION

By following the outlined steps, we successfully extracted data from the 79 URLs associated with Hardy Riggings. The resulting CSV file provides a comprehensive overview of each product, including essential details such as product title, description, variants, price, product manual and certifications, and spec sheet. The accuracy and completeness of the extracted information enhance the efficiency of WooCommerce integration, enabling streamlined product management within the WordPress ecosystem.

Contact

https://upreach.ai/
vidranatan@gmail.com
https://www.linkedin.com/company/anote-ai/
https://www.youtube.com/@anote-ai