PoC related to data extraction from URL

derived from the e-commerce sector

The main goal of the solution is from a URL to extract

if they are present at all

- 1. Furniture objects (products)
- 2. Properties of the extracted objects such as:
 - colour
 - size
 - shape
 - what the product is made of (fabric, metal, wood)
 - o
- 3. Name of the product

State-of-the-art Generative AI extraction methods have been applied

Generative Pre-trained Transformers, commonly known as GPT, are a family of neural network models that uses the transformer architecture. In this solution in-context learning (ICL) has been applied on a small subset of manually processed URLs. It can be considered as part of the so called "Prompt Engineering".

More precisely, the Web-application gives the option to the user to experiment with different URLs in the left side-bar:

- · Put the URL
 - Example: https://home-buy.com.au/products/bridger-pendant-larger-lamp-metal-brass
- Preprocess it to clean some unsignificant information such as the domain of the URL through regular expressions

Example: bridger pendant larger lamp metal brass

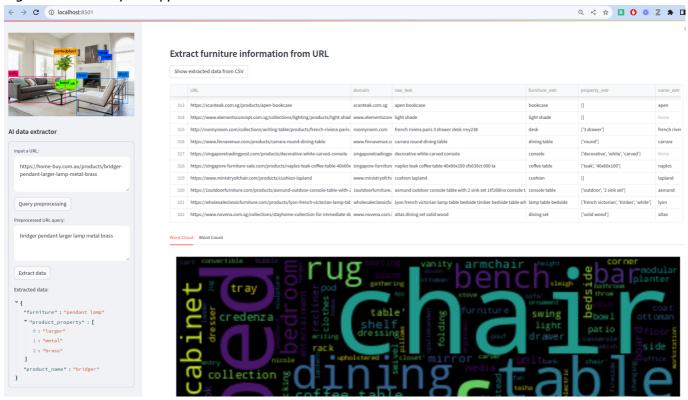
- · Get the extracted data in JSON format
 - "furniture": "pendant lamp"
 - "product property": ["larger", "metal", "brass"]
 - "product_name": "bridger"

The user is given the option to load and visualize preprocessed CSV file with sample URLs by pushing the button "Show extracted data from CSV".

The extracted fields are: furniture_extr, property_extr, name_extr

The user is also given some simple data visualization of the extracted furniture objects in the form of "Word Cloud" and "Word Count" histogram. *

Fig 1. Screenshot of the app



Tech-stack

- 1. For the AI (data extraction) part
 - LangChain: framework for developing applications powered by language models
 - OpenAI/GPT-3: Transformers' powered search, conversation, text completion, and other advanced AI features