Table of Contents

Introduction

This report details the process and outcome of an assessment designed to evaluate my ability to apply machine learning skills, particularly in natural language processing, to solve a real-world problem. The task involved developing an API that extracts meaningful attributes from an HTML block of an e-commerce website using an open-source large language model (LLM) and returns the relevant and meaningful information in JSON format.

Large language models (LLMs) are advanced machine learning models trained on vast amounts of text data. They are capable of understanding and generating human-like text, making them suitable for various natural language processing tasks such as text extraction, summarization, translation, and more. In this assessment, I utilized an LLM to process HTML content and extract structured data.

Objective

The objective of this assessment was to create an API capable of:

- Extracting meaningful attributes from an HTML block of an e-commerce website.
- Returning the relevant information in JSON format.
- Utilizing an open-source LLM.

Methodology

Model Selection

For this task, I selected the Llama 2 model, specifically Llama-2-7B-Chat-GGML. This model was chosen due to the following reasons:

- Open-source: Llama 2 is freely available and suitable for the requirements of this assessment.
- Performance: Llama 2, particularly the 7B variant, is known for its robust performance in natural language processing tasks.
- Resource Efficiency: Given that I have a low-end PC, the Llama-2-7B-Chat-GGML model is lightweight enough to run locally without requiring high computational resources.
- Adaptability: The model can be fine-tuned and adapted to specific tasks, making it ideal for extracting structured data from unstructured HTML content.

Implementation Steps

1. Model Setup: I set up the Llama-2-7B-Chat-GGML model in my local environment.
2. Data Preparation: HTML blocks from the e-commerce website daraz.com.np were collected and preprocessed for input into the model.
3. Model Training: The model was fine-tuned using a subset of the HTML data to enhance its ability to identify and extract relevant attributes.
4. API Development: I developed a Flask API to interface with the model, process the HTML input, and return the extracted information in JSON format.
5. Testing and Validation: The API was tested with various HTML inputs to ensure accuracy and reliability of the extracted data.

Results

The developed API successfully extracted meaningful attributes from the HTML blocks and returned the information in JSON format. The Llama-2-7B-Chat-GGML model demonstrated high accuracy in identifying relevant information such as product names, prices, descriptions, and other attributes.

Conclusion

The use of the Llama-2-7B-Chat-GGML model proved effective for the task of extracting meaningful attributes from HTML content. The open-source nature, performance capabilities, and resource efficiency of Llama 2 made it an excellent choice for this assessment, particularly given the constraints of using a low-end PC. The developed API can be a valuable tool for e-commerce platforms like daraz.com.np to automatically extract and organize product information, enhancing data management and user experience.