ML Engineer Assignment

Table of Contents

Note: for additional information please refer to the presentation.

Introduction

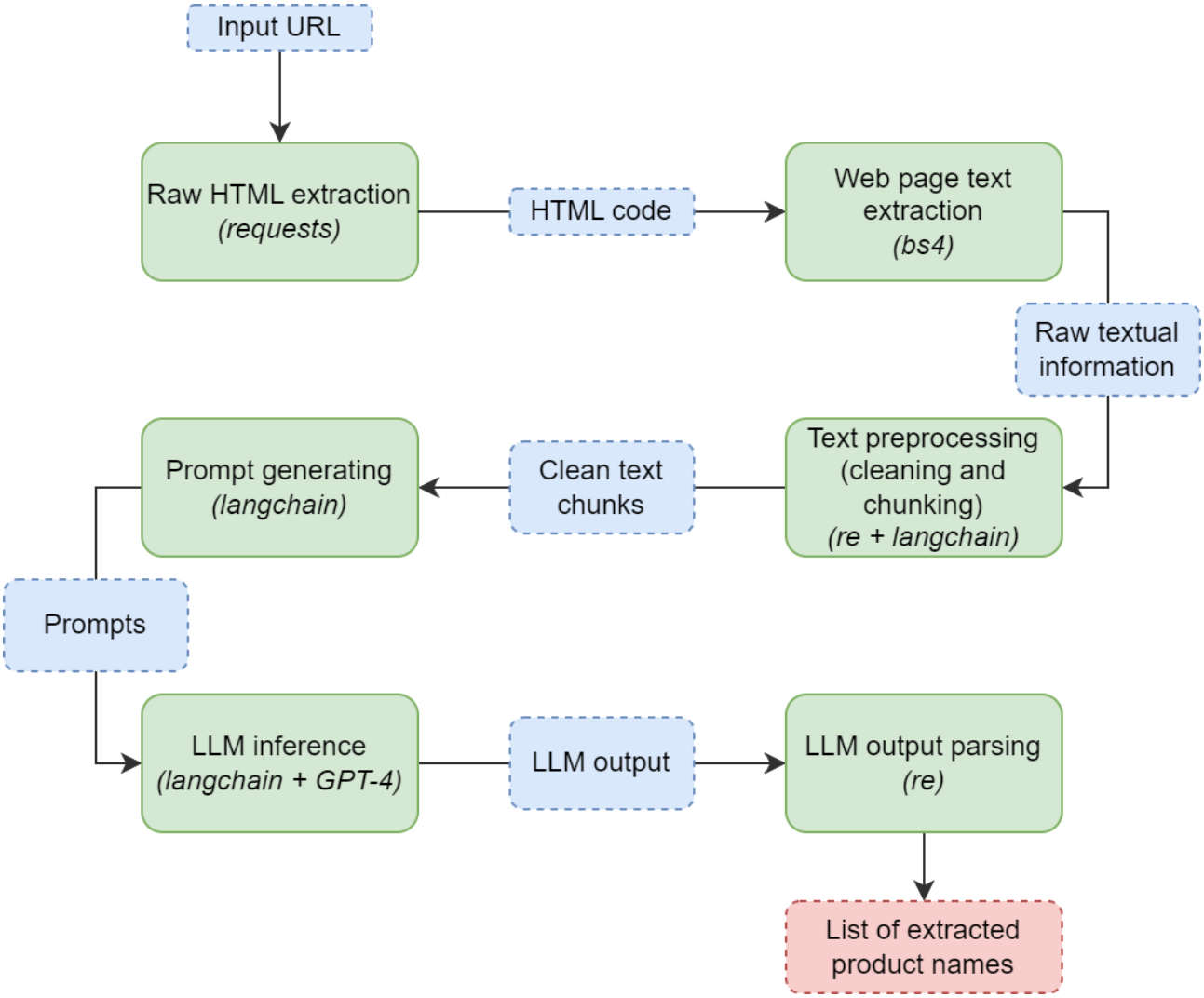
The purpose of this simple application is to extract product names from online furniture stores. Our approach is based on using LLM and prompt engineering for entity recognition and extraction. The main reason we choose this method is because any other approaches, such as custom Named-Entity Recognition pipeline, require additional training or fine-tuning and, most importantly, data labeling.

General pipeline

The general pipeline is the following:

- First, raw HTML is obtained by sending a GET request to the specified URL.
- All textual information is then extracted by BeautifulSoup4 library.
- Obtained raw text must be preprocessed, so we remove any unnecessary whitespaces and linebreaks.
- In order not to exceed the input token limit, we split the texts into several overlapping chunks using CharacterTextSplitter from Langchain framework.
- For each chunk we generate a prompt for GPT-4. According to the prompt, model should return a list of product names or “None” if no product is mentioned. Interface for GTP-4 is provided by Langchain framework.
- Run LLM inference for each generated prompt.
- Using some magic of regular expressions, we extract product names from the LLM output. Product names is then printed to console.

This pipeline is also described in the image below.



Results

Without labeled data we cannot measure the quality of our approach, so from the first 48 URLs we've obtained 19 working web pages, that contained products names, and then we've manually extracted entities of interest (this data is stored in the `/data/scraped_products.json`). Finally, we've computed precision, recall and f1-score by comparing our results with the output of the described pipeline.

Precision	0.676
Recall	0.742
F1-score	0.708

We can see that results are not quite perfect: low precision means that LLM extracts unnecessary entities that are not presented in labeled data (number of false positives is high); low recall means that LLM doesn't extract all entities of interest (number of false negatives is also high). However, the main reason of this is that our data labeling is subjective. For example, LLM can extract correct entities, but with longer (or shorter) names compared to our labels. This fact explains low values of precision and recall respectively. To measure the actual quality metrics, we had to recompute them manually. Approximate values are given in the table below.

Precision	0.911
Recall	0.995
F1-score	0.951

As we can see, these results are much better than previous! You can compare LLM's output and ground-truth product names by yourself, all data is stored in the *results.txt*. If you want to see automatically computed metrics, run *metrics.py*.

Usage example

Note: you must have access to GPT-4 model deployed in Azure OpenAI Service.

To use the app, set up a virtual environment, install packages from *requirements.txt*, and create a *.env* file with the following content:

```
API_KEY = <YOUR_AZURE_OPENAI_API_KEY>
API_TYPE = azure
API_VERSION = 2023-05-15
ENDPOINT = <YOUR_MODEL_ENDPOINT>
LLM_DEPLOYMENT = <YOUR_MODEL_DEPLOYMENT_NAME>
```

Once you have this, you can use the app in the command line by running:

```
python main.py --url <your_url_here>
```

For example:

```
python main.py --url  https://www.factorybuys.com.au/products/euro-top-mattress-king
```

Output:

```
Factory Buys 32cm Euro Top Mattress - King
Savannah Grey Bed Frame Fabric Gas Lift Storage - King
Azalea LED Bed Frame PU Leather Gas Lift Storage - Black King
Florence Metal Bed Frame Base Platform - Black King
Lucca Bed Frame Fabric Gas Lift Storage - Grey King
KING Sheets Fitted Flat Bed Sheet Pillowcases - Summer Grey
KING Mattress Topper 100% Wool Underlay Reversible Mat Pad Protector
KING Mattress Topper Bamboo Fibre Pillowtop Protector
```

Known issues

- As we've already mentioned, it's hard to estimate the actual quality of the pipeline, since it requires data labeling that can be subjective.
- Sometimes LLM extracts entities that are not visible on the web page, so with actual product names we can get some junk (this problem explains not so high precision score).

- This app cannot be used “out-of-the-box” because it requires access to GPT-4 and Azure OpenAI Service. However, experiments with more accessible LLMs can fix this issue.