City, University of London

Msc Data Science

Yumi Heo

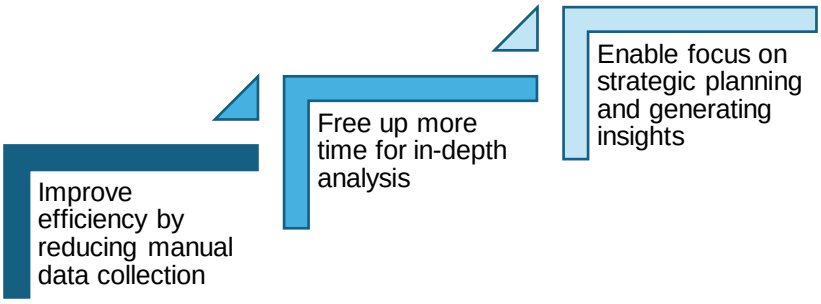
Automated Data Scraping

Project Progress and Key Updates

October 2024

1. Objectives for Automated Data Scraping

- To maintain a consistent database of medicine guidance from regulatory agencies and HTAs for analysis and statistics
- Stepwise Objectives



2. How AI Works in This Project

We used a Large Language Model (LLM), specifically GPT-4o mini from OpenAI, to streamline data extraction from PDFs and web pages. LLMs are pre-trained models that function similarly to neural networks in the brain, recognizing patterns, summarizing, predicting, and generating content based on extensive datasets that those models learned. We chose GPT-4o mini for its efficiency, balancing high performance with lower cost, given the time constraints of the 3-month project. The model identifies relevant data from unstructured sources such as PDF files (e.g. procedural steps PDF files from EMA) or web pages (e.g. EMA website or NICE website). This data is the model's response from our queries or questions. Then, the data was saved to create a database.

3. Project Progress

- EMA data scraping: The basic framework for scraping EMA data has been completed. It can successfully collect the requested information and save it in an Excel file. However, further improvements are needed.
  **Data Accuracy*: The model's ability to accurately interpret and derive answers from the unstructured sources needs to be refined. Improving the precision of the answers generated is a key focus.

- NICE scraping: We have been reviewing the range and format of data that can be extracted via the API service provided by NICE. Currently, we are testing the model's ability to respond accurately through simple queries based on this data or relevant web pages in NICE.
  **Work Expectation*: It is estimated that it will take approximately 3 months to finalize the script for scraping NICE data, as this database involves collecting a broader and more diverse set of information compared to EMA.

4. EMA Data Scraping
   - Script Location: Script Link
     **How to use*: Save the script to your personal Google Drive and open it. Detailed instructions for script execution can be found within the file itself.
   - Output: The script generates an Excel file named '`final_EMA_dataset`'.
   - Excel Features: Excel File Link
   - Notes:
     - The script scrapes only the most recent *meeting highlights* news from the EMA's news page, filtered through CHMP.
     - The code in the script is functionally interconnected, meaning it is highly dependent on the source data from which values for the dataset's features (e.g., therapy area, medicine name, brand name) are extracted.
     - The scraping process captures data directly from the web page (e.g., medicine name, news dates, brand name, etc.), extracting the relevant text from the site.
     - To enhance data collection efficiency, particularly for indications, we used a large language model to help organize the data, enabling the model to respond to our queries based on information gathered from multiple sources.

5. NICE Data Scraping
   - The script is still under development and will rely on decisions regarding the sources from which the values for the features of the NICE dataset will be collected. We are currently evaluating which approach is most efficient for each feature to ensure higher accuracy in the obtained values.

6. Current Limitations
   - Script Limitations
     - Manual Execution Required: The script collects data and generates an Excel file, but it is not a standalone software application. It requires manual execution through Google Colab, followed by downloading the generated file.
     - Ongoing Query Testing: Ongoing testing of appropriately prompted queries is needed to improve the accuracy of the model's responses.
     - Model Accuracy: The model's answers are not always 100% accurate, meaning the values for the features provided by the model require additional review and validation.
     - HTML Structure Dependency: The script depends on the current HTML structure of EMA or NICE web pages. If the structure changes, manual inspection and updates to the script are necessary to ensure continued functionality.

   - External Limitations:
     - Ongoing Costs: There are continuous costs associated with using OpenAI's model for data processing.
     - Data Display Changes by EMA: If EMA changes how it displays the data updates (e.g., indication updates in the "variation" page, procedure steps PDF, or other sections), the query or input file for the model may need to be adjusted accordingly. Currently, the extension of indication can be found in a separate web page called 'variation' in the URL or in the procedure steps PDF, but sometimes the extension of indication exists in both ways or exists in only one of the two. Sometimes EMA removes a previously existing variation page or doesn't update or upload the procedure steps PDF for a medicine.
     - NICE API Stability: If the NICE data servers experience instability (e.g., 500 error in the server means the internal server error), the script for NICE data scraping will be temporarily unusable until the issue is resolved.

7. Future Work

Basic groundwork is required to collect NICE data, which is expected to involve both feature connection through the API and web page scraping. Additionally, model prompt testing and accuracy assessments are needed to further enhance the EMA data scraping script.