Question 1 Documentation

Task 1([task1.py](https://github.com/shubham-rr/HIT137_Assignment2/blob/main/Question%201/task1.py)):

For this task, we created a script to extract and clean text data from multiple CSV files. The goal was to take text from the 'TEXT' and 'SHORT-TEXT' columns, remove any non-alphabetic characters, and convert everything to lowercase for consistency. The cleaned text is then saved into a single file, extracted.txt.

Firstly, we made a files/ directory to store the original datasets, which are too large to upload to GitHub, so we made sure to ignore that folder in the repository by listing it in our .gitignore file.

The script checks if the output directory exists, and if not, it creates it. Then, it reads all the CSV files in the specified folder (“files/”), processes the text from the relevant columns, and writes the cleaned text into the output file.

Task 2:

For Task 2, we focused on installing the necessary libraries for extracting biomedical entities from text, like drugs and diseases.

We started by installing SpaCy and SciSpacy along with two specific models: en\_core\_sci\_sm and en\_ner\_bc5cdr\_md. These models help us detect medical terms, but there was an issue with the latest version of SciSpacy due to the nmslib package. To fix this, we had to use scispacy==0.5.2 and set up a virtual environment with Python 3.8.9, which worked well.

After setting up the environment, we also installed Hugging Face Transformers and BioBERT to detect entities from the biomedical text. These models add another layer of precision in identifying drugs and diseases.

We created a requirements.txt file to track all dependencies. So, to install everything, you can just run pip install -r requirements.txt, and if any new packages are needed, we update the list with pip freeze > requirements.txt. You can view the full list of packages we used in our [github repository](https://github.com/shubham-rr/HIT137_Assignment2/blob/main/requirements.txt):

<https://github.com/shubham-rr/HIT137_Assignment2/blob/main/requirements.txt>

Task 3 ([task3.py](https://github.com/shubham-rr/HIT137_Assignment2/blob/main/Question%201/task3.py)):

For Task 3, we tackled two different methods for counting words and tokens in a text file.

In **Task 3.1**, we used Python’s built-in libraries to count how often each word appears in the text. After cleaning the text to remove stopwords and unnecessary words, we identified the top 30 most frequent words and saved them in a CSV file called top\_30\_words.csv.

For **Task 3.2**, we used the **AutoTokenizer** from the Transformers library. This allowed us to tokenize the text, count unique tokens, and then filter them to remove subwords (like those starting with ## from the tokenizer). The top 30 tokens were saved to top\_30\_tokens.csv.

Task 4 ([task4.py](https://github.com/shubham-rr/HIT137_Assignment2/blob/main/Question%201/task4.py))

For this task, we have used Named-Entity Recognition (NER) to extract ‘diseases’ and ‘drugs’ entities from a sample text file named test.txt instead of the full extracted.txt file due to processing time constraints. We utilized two models: SciSpacy (with ‘en\_core\_sci\_sm’ and ‘en\_ner\_bc5cdr\_md’) and BioBERT. The text was cleaned and chunked to meet the models' input requirements. We compared the entities detected by each model in terms of total counts and the most common entities found. The comparison also included identifying common entities across all models. The results, including the total counts and a summary of the most frequent entities, were compiled into CSV files.