The task requires to find relevant papers from a list of papers curated in the field of virology and epidemiology. This documents lists steps followed to perform the necessary asked tasks and listing some interesting findings along the way. Before proecedding with the tasks, for entries missing abstract, it was replaced with title of the publication

Task #1
It asks to implement a semantic NLP technique to filter out papers that do not meet the criteria making use of deep learning strategies. For this we use semantic similarity matching using a sentence transformer model, the emebddings for the (abstract+title) of the paper and term “deep learning” were compared on cosine similarity and based on thresholding (decided after manual inspection), relevant papers were filtered.

Q: Why is this approach better than manual keyword based filtering?

- Better contextual understanding of the entire abstract

- Manual keyword based filtering is likely to miss synonyms or related phrases leading to poor recall

- Embeddings created through sentence transformer models are better suited to adapt to longer and complex text sequences

- Reduced need for manual intervention to update the list of keywords to search from (compared to a lexicon based matching approach)

After this task, 7927 papers were identified as being relevant.

Task #2
This task asks to further classify the relevant papers into type of method used: computer vision, text mining, both, other

For this, we use zero shot classifiaction using bart model with specific class labels defined. The filtering criteria is based on “score” value output by the language model for each class label. Output of the pipeline wrapper by huggingface makes the process for deciding the final class label more streamlined.

Other approaches explored for this task and their limitations:
- Semantic similarity based on keyword matching, however its difficult to curate an exhaustive list for relevant keywords for both the domains.

Task #3

This task requires to identity the methods used in the relevant papers. This was the most challenging task purely because of complexity of scientific language used in the abstract.

For this we use a token classification based approach to identify all scientific methods used in the abstract. The language model is sci-bert and we use a finetuned version of that model. The details of the dataset on which it was fine-tuned is unknown, however most likely it is SemEval 2017 Task 10 (https://aclanthology.org/S17-2091/). The model outputs tokens according to BIO classification scheme, after which tokens were processed using a helper function. To deal with limited sequence length of bert: 512, the text passed as input was broken down into chunks if needed and finally the identified tokens were passed through a post-processing function to cleanup the tokens. Please refer to the code, it is sufficiently commented.

Other approaches that were explored for the task and their limitations:

- LDA topic analysis was explored with basic pre-processing to filter out stop words, however the converged topics did not align to scientific terms expected, also it was difficult to estimate the #topics and #keywords for this given dataset.

- Tf-idf based approach to identify important keywords, a severe limitation of this approach was to come up with a heuristic to filter scientific method based words from other terms identified.

The final code includes the columns that are added to the dataset which the tasks demands, namely Label that shows which category the paper belongs to, and Top_Methods listing scientific methods used in each relevant paper.