Code Walkthrough:

- Run **pre_process.py** to convert the .pdf/.csv file to a .txt file.
    - Input docs should be present in the following format:
        - If .pdf file, it should be located in **data/pdf_data**
        - If .csv file, it should be present in **data/csv_data**
    - The text files are stored in **data/text_data**
- The text data is uploaded to **label-studio** where it is tagged/annotated
- Annotated data is downloaded as a .json/.jsonl file and stored in **data/train_data/train/** and **data/train_data/dev**/ folders respectively
- The **training.py** file picks this annotated data from the train and dev folders and creates an equivalent "*.spacy*" file and stores them in data/spacy_files and then saves the model(best and last) to the folder specified by the user
- The **predict.py** file accesses the saved model and accesses the test data in the folder **data/test_data/** in a "*.txt*" file, performs NER, and saves the file as a .csv in **output/** folder