Training NER Process

Abbey performs OCR to generate text data

Tag NER data using spacy tagging tool & then save it text file

Abbey performs OCR to generate text data

Text extraction pipeline

Load files for verification of location and port

Identify page type

**Current flow**

1. Once file placed in folder UI path pickups file from it then pass it to Abbey OCR reader for converting it into text files.
2. To perform training process need to tag data using spacy tag editor then sve that file as txt file manually.
3. Run training script manually for 50 different models then evaluate each model performs based on f1-score for each entity then decide best model.
4. To perform prediction load all the necessary files like geolocation, vessel, port for different enitities.
5. Load bag of words file for port and city name for identifing confliction in names.
6. Generate date folder specific date files then start extracting export and buyer with other extra information if type is application.
7. If page type is not application then extract port, other location and person name related information.
8. Product name if product name is not present inside application page.
9. Fetch country name from NER model then compare with loaded data from file for cross validation.
10. Combine all port information into single dataframe like exporter, buyer and other locations.
11. Email extraction using regex.
12. After all data been extracted i creates 2 files complete and summerized excel file before summerization combine all organisation names.
13. All this data is then inserted into db.

**Proposed flow 1**

1. Once file placed in folder UI path pickups file from it then pass it to Abbey OCR reader for converting it into text files.
2. To perform training process need to tag data using spacy tag editor then save that file as txt file manually so insteaded of doing this we can using our geolocation, vessel, port file and salutation/regex for finding name then tag data automatically just cross check from user to save time.
3. Once this is done move data into specific folder then start training process automatically.
4. Run training script manually for 50 different models then evaluate each model performs based on f1-score for each entity then decide best model through automatly comparing previous model performs with current one. All model performs should be stored inside db.
5. To perform prediction fetch data from tables like geolocation, vessel, port for different enitities as per requirement.
6. Fetch data of bag of words file for port and city name for identifing confliction in names.
7. Build better folder structure to store every data getting generated through each steps like individual folder for each file pdf then ocr file, text files, operation performed files, result files should be stored in individual folder in that particular folder.
8. If page type is not application then extract port, other location and person name related information.
9. Product name if product name is not present inside application page.
10. Fetch country name from NER model then compare with loaded data from file for cross validation.
11. Combine all port information into single dataframe like exporter, buyer and other locations.
12. Email extraction using regex.
13. After all data been extracted i creates 2 files complete and summerized excel file before summerization combine all organisation names.
14. All this data is then inserted into db.
15. Provide it to end user through UI to cross check the data extracted through pipeline and possible option of editing extracted data so we can perform improvement on models and correct data is passed on to next process.
16. Maintain logs for each step to identify steps completed for each file and it’s failure which can be used by compliance team.

**Proposed flow 2**

1. Once file placed in folder UI path pickups file from it then pass it to Abbey OCR reader for converting it into text files.
2. Completely remove process of tagging but insteaded of doing this we can use our geolocation, vessel, port tables and salutation using db regex method for finding entities.
3. For general entities like geolocation, date, amount we can still use spacy + duckling libraries. We don’t need to train them use their based model for extraction.
4. To perform prediction fetch data from tables like geolocation, vessel, port for different enitities as per requirement.
5. Fetch data of bag of words file for port and city name for identifing confliction in names.
6. Build better folder structure to store every data getting generated through each steps like individual folder for each file pdf then ocr file, text files, operation performed files, result files should be stored in individual folder in that particular folder.
7. If page type is not application then extract port, other location and person name related information.
8. Product name if product name is not present inside application page.
9. Fetch country name from NER model then compare with loaded data from file for cross validation.
10. Combine all port information into single dataframe like exporter, buyer and other locations.
11. Email extraction using regex.
12. After all data been extracted i creates 2 files complete and summerized excel file before summerization combine all organisation names.
13. All this data is then inserted into db.
14. Provide it to end user through UI to cross check the data extracted through pipeline and possible option of editing extracted data so we can perform improvement on models and correct data is passed on to next process.
15. Maintain logs for each step to identify steps completed for each file and it’s failure which can be used by compliance team.