User guide for Steel coatings pipeline

DIRECTORY STRUCTURE

steel_coatings_pipeline
    search_annotation tool
    prepare_dataset
    machine_learning_models

SEARCH_ANNOTATION_TOOL

Consists the query tool, for further instructions look up the user guide inside this folder.

PREPARE_DATASET

Folders

Steel_coatings_text: The original dataset
Section_wise_text: contains the research text divided section wise.
ip_to_tagger: Dataset with the text from the 'experiment' section extracted out.
sentence_ip_to_tagger: Dataset with extra newlines from text
sentence_wise_normalised_ip_to_tagger: Dataset with normalized sentences.
consolidate: contains files and code to extract AMCPW structures.
ip_to_docaano: input to doccano
ip_to_docaano_concatenated: contains files where all the sentences are concatenated into one dictionary
ip_to_docaano_concatenated_clean: cleaned dataset
op_doccano: output from doccano
chemicaltagger-chemicalTagger-1.5.0

Code

sentence_ip_to_tagger.py: removes newlines from the files in ip_to_tagger folder.
sentence_normalise: normalizes sentences from files in sentence_ip_to_tagger
ip_to_doccano: creates dataset suitable for performing annotation on doccano.
clean_json: cleans the json files in ip_to_doccano_concatenated.

Workflow

1) run sentence_ip_to_tagger.py
2) run sentence_normalise
3) run ip_to_doccano
4) run clean_json

MACHINE LEARNING MODELS

Consists of code for the various approaches tried for NER
Demo: contains demo code to test the models. Just run the .py files.