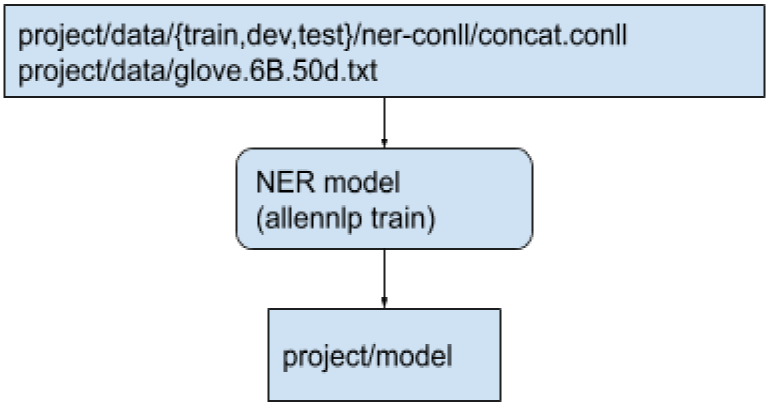
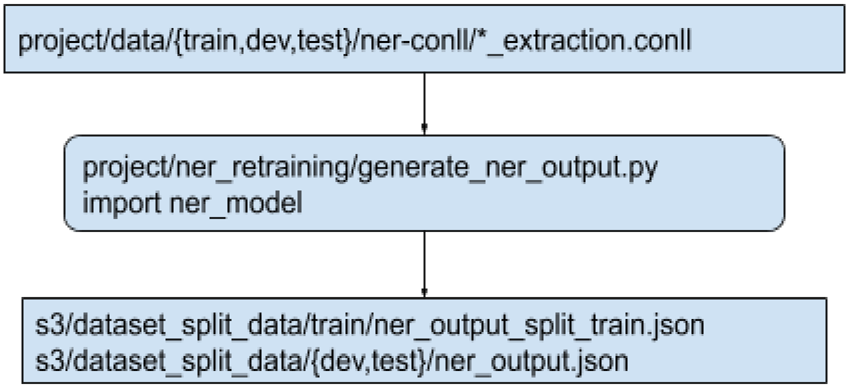
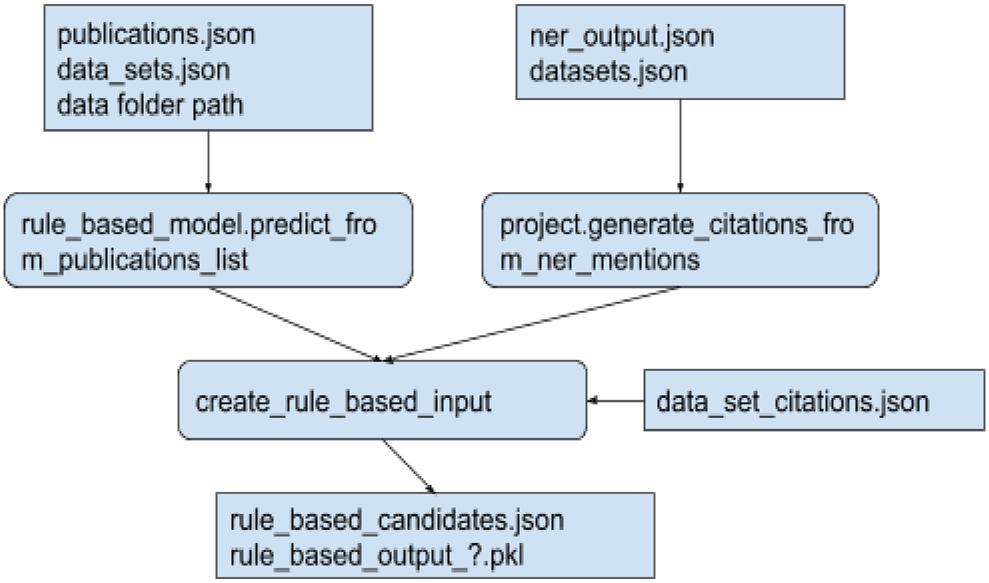
1. **project/create_dataset_split_folder.py** creates project/dataset_split_data folder with dev, train and test sub-folders
2. **project/to_conll.py** converts all of the publications in train/dev/test to conll format, both for NER and linking. The output will be in folders called ner-conll and linking-conll.
3. Train CRF tagger Named Entity Recognition (NER) model using glove word embeddings with paper texts. Configuration is in project/ner_model/allennlp-ner-config.json or project/ner_model/tweaked_parameters_config.json.


project/data/{train,dev,test}/ner-conll/concat.conll
project/data/glove.6B.50d.txt
NER model
(allennlp train)
project/model

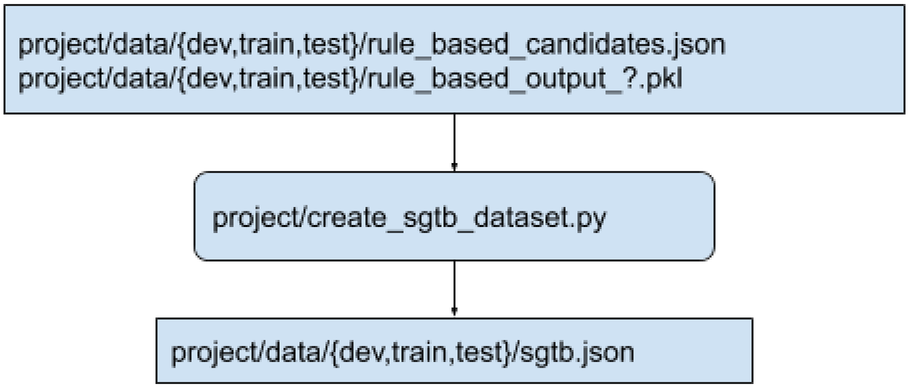
4. Create NER produced mentions. The output is a set of quadruplets of publication_id, mention, score and instance.


project/data/{train,dev,test}/ner-conll/*_extraction.conll
project/ner_retraining/generate_ner_output.py
import ner_model
s3/dataset_split_data/train/ner_output_split_train.json
s3/dataset_split_data/{dev,test}/ner_output.json

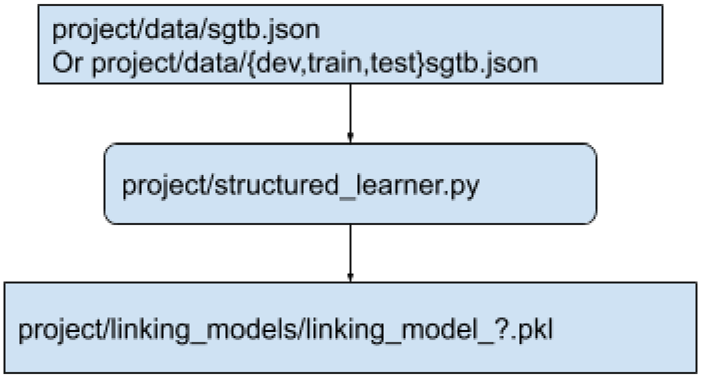
5. Create rule based dataset candidates using RuleBasedModel, golden dataset citations (tagged by humans) and mention predictions produced by NER. Code for **project/create_linking_dataset.py**:



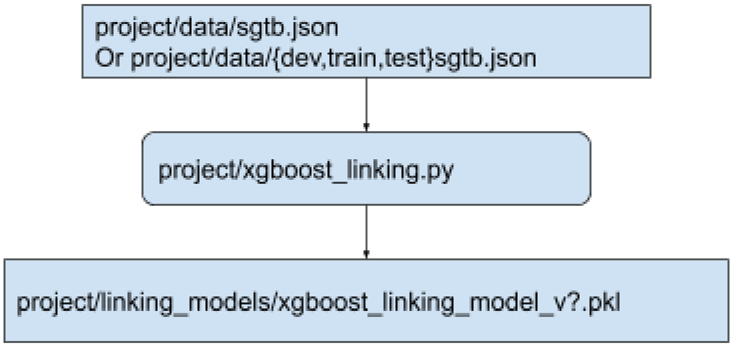
6. Create the featurized dataset of mentions and candidates


project/data/{dev,train,test}/rule_based_candidates.json
project/data/{dev,train,test}/rule_based_output_?.pkl
project/create_sgtb_dataset.py
project/data/{dev,train,test}/sgtb.json

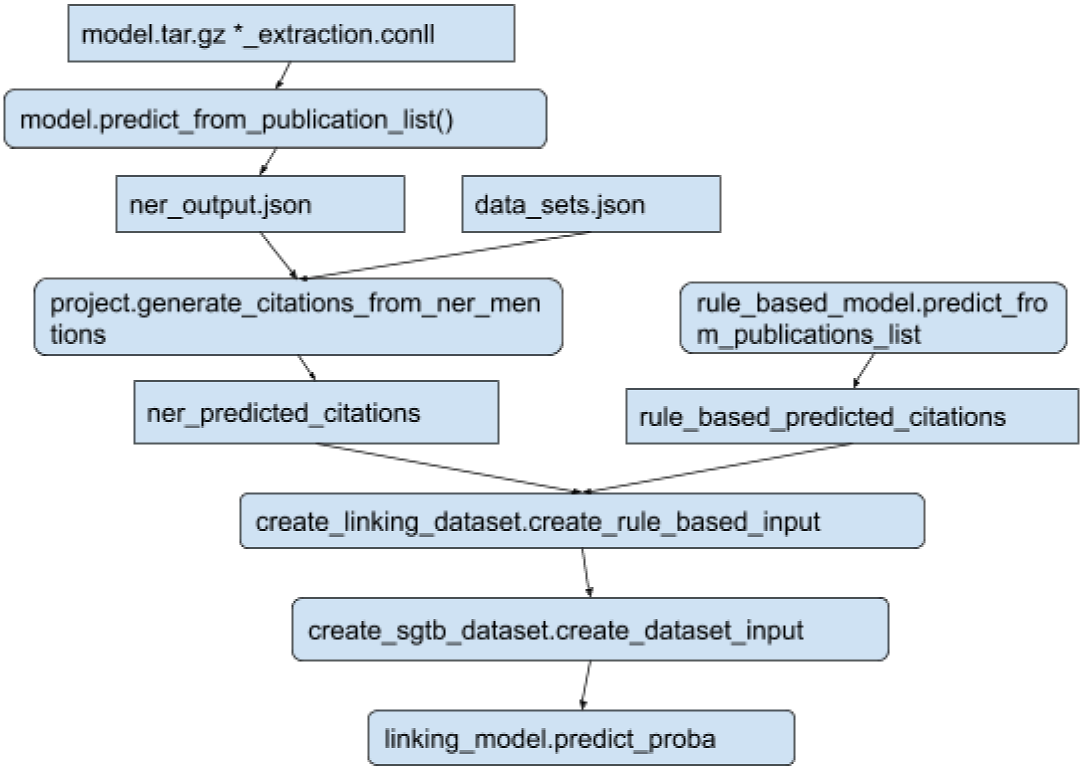
7. Train the Structured Gradient Tree Boosting model


project/data/sgtb.json
Or project/data/{dev,train,test}sgtb.json
project/structured_learner.py
project/linking_models/linking_model_?.pkl

8. Train the XGBoost model


project/data/sgtb.json
Or project/data/{dev,train,test}sgtb.json
project/xgboost_linking.py
project/linking_models/xgboost_linking_model_v?.pkl

Evaluation of the approach: project/project.py


model.tar.gz *_extraction.conll
model.predict_from_publication_list()
ner_output.json
data_sets.json
project.generate_citations_from_ner_men tions
rule_based_model.predict_fro m_publications_list
ner_predicted_citations
rule_based_predicted_citations
create_linking_dataset.create_rule_based_input
create_sgtb_dataset.create_dataset_input
linking_model.predict_proba

Function project.generate_citations_from_ner_mentions takes NER generated mentionas from ner_output.json and dataset title and produces citation candidates using TF-IDF weighted overlap with dataset titles, then it returns list of top rated citation candidates for each mention.

Example from **data_set_citations.json**:

```
                {
                "citation_id": 1945,
                "publication_id": 105,
                "data_set_id": 311,
                "mention_list": [
                "Deutsche Bundesbank's balance of payments statistics"
                ],
                "score": 1.0
                },
```

Example from **publications.json**:

```
        {
        "publication_id": 116,
        "unique_identifier": "bbk-15",
        "title": "Why do banks bear interest rate risk?",
        "pub_date": "1969-01-01",
        "pdf_file_name": "116.pdf",
        "text_file_name": "116.txt"
        },
```

Example from **data_sets.json**:

```
                {
                "data_set_id": 1,
                "unique_identifier": "10.3886/ICPSR07213",
                "title": "ANES 1952 Time Series Study",
                "name": "ANES 1952 Time Series Study",
                "description": "This study is part of a time-series collection of national surveys
        fielded continuously since 1948. The election studies are designed to present data on
        Americans' social backgrounds, enduring political predispositions, social and political
        values, perceptions and evaluations of groups and candidates, opinions on questions of
        public policy, and participation in political life. The 1952 National Election Study gauges
        political attitudes in general, along with attitudes and behaviors directly relevant to the
        1952 presidential election. The interview schedule contained both closed and
        open-ended questions designed to collect data on a wide range of issues. Most
        respondents were interviewed both before and after the date of the election. The
        pre-election survey tapped attitudes toward political parties, candidates, and other
        specific issues, and inquired about the respondents' personal and political background.
        The post-election interview focused on the actual vote and voting-related behaviors.
        Additionally, a sub-sample of 585 respondents was administered a Form B re-interview
        obtaining further information about organizational affiliations, personal data, and
        non-political opinions and attitudes. A special emphasis was placed on the perception of
```

```
group behavior, especially the perceived political preferences of family, friends, and associates.",
	"date": "2016-09-20 00:00:00+00:00",
	"coverages": "",
	"subjects": "candidates,congressional elections,domestic policy,economic conditions,foreign policy,government performance,information sources,national elections,political affiliation,political attitudes,political campaigns,political efficacy,political issues,political participation,presidential elections,public approval,public opinion,special interest groups,Truman Administration (1945-1953),trust in government,voter expectations,voting behavior,United States,1952-09--1952-12",
	"methodology": "",
	"citation": "",
	"additional_keywords": "ICPSR",
	"family_identifier": "",
	"mention_list": [
	"ANES study",
	"ICPSR",
	"SRC data",
	"Surveys conducted by the Survey Research Center and the Center for Political Studies of the University",
	"eight SRC-CPS presidential election surveys",
	"eight SRC-CPS presidential election surveys con- ducted between 1952 and 1980",
	"eight presidential election surveys conducted by the Survey Research Center and the Center for Political Studies (SRC-CPS)",
	"time series"
	],
	"identifier_list": [
	{
		"name": "ICPSR data ID (dataId)",
		"identifier": "10.3886/ICPSR07213"
	}
	]
	},
```