Notes on replication materials

Peter Norlander

December 18, 2022

- Stata_files.zip:
    - The .do file takes the output of CRAML and makes the tables and figures in the paper "**Streetlight Effects in the Study of Employer Collusion in the Labor Market: On the Variety and Extent of Anti-Competitive Clauses in Franchise Documents**"
        - Uses 2251.csv and 28542.csv (described below), and other rule-generated output from CRAML (in the "replicate" folder to make figures and tables.

- The files below can be unzipped and placed in a "project" folder by a user of CRAML to replicate the text-to-data process using ML or rules. The files contain the extracted text; the full corpus will be posted to another repository.
- ca_ml_franchise_nopoach.zip and mn_ml_franchise_nopoach.zip contain:
    - settings.json – needs to be configured on local machine
    - keywords.json
    - folders:
        - csvs –
            - extracted text from CA or MN
        - db – CRAML output:
            - 2251.csv – CA metadata with ML classification
            - 28542.csv – MN metadata with ML classification
        - logs – information CRAML records for user
        - rules - rules files to classify text with exact matches and build training data
        - sample – a sample of the corpus containing keywords
        - train – ML classifiers
            - output of CRAML – classified text chunks
            - combine_augment.py merges data from 10% sample + 100% sample with nopoach==1 only.
            - Augment_aug25 is the result, but this was pruned to augment
            - MN data is classified using CA ML model