Civera

Project Deliverable 1

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request. If data has already been collected for your project you must answer two questions.

Checklist

1. Collect and pre-process a preliminary batch of data.
   1. We have familiarized ourselves with the datasets.
   2. Four tables with millions of rows and 5 to 13 fields in SQL.
   3. Many duplicates.
   4. Many missing fields.
2. Perform a preliminary analysis of the data.
   1. After reviewing the data, we decided to:
      1. Write scripts to pull rows of the case\_action\_index table incrementally.
      2. Update the client’s legacy “brute-force” regex in PHP with spaCy.
      3. Normalize by the same primary id to match it one-to-one with the source table, where we’ll add the critical actor and actions fields.
3. Answer one key question.
4. Refine project scope and list of limitations with data and potential risks of achieving project goal.
   1. Given the difficulty and importance of the tasks defined above, we met with the client and amended the SOW. Basically due to the garbage-in, garbage-out philosophy of data cleaning.
5. Submit a PR with the above report and modifications to original proposal.