CS 513 Theory & Practice of Data Cleaning

Final Project: End-to-End Data Cleaning Workflow

The goal of the final project is to use various tools and techniques covered in this course together in a small end-to-end data cleaning workflow. This document contains important information about the final project, including team formation, dataset options, recommended workflow, project deliverables, grading criteria, and sample projects. Please periodically check Piazza for additional information and new updates about the project.

Team Formation

You are strongly encouraged to form teams of **2 or 3** students. In exceptional cases, you are allowed to form a team of 4 students or do an individual project. Exceptions (i.e., teams of 4 students or individual projects) require permission from the instructor: Please send a private message to the TA-team on Piazza explaining why you can't be part of a regular team (2 or 3 students). The TAs/instructor will then follow-up via Piazza or email.

To find teammates, use Slack or Piazza!

Dataset Options

There are three options for choosing a dataset that you will clean in the project. Start by reviewing the web sites (a) and (b) for which datasets to be cleaned already exist (the reference versions of the datasets are currently available under Piazza resources: https://piazza.com/illinois/summer2020/cs513/resources), or think about a dataset of your own choosing (c):

(a) US Farmers Markets:
https://www.ams.usda.gov/local-food-directories/farmersmarkets
(b) New York Public Library's crowd-sourced historical menus:
http://menus.nypl.org/
(c) Your own dataset of choice! (see below)

We strongly recommend individuals to choose option (a), since the dataset is relatively small. We expect teams to choose either option (b) or option (c). Please note that, for

teams of 3 or more students, the optional phases in the Recommended Workflow (see the next section) are mandatory.

Further Instructions for Option (c). If you choose your own dataset, please use a dataset that is publically available (preferred) or that you are allowed to share (with your teammates and the instructor/TAs). If you plan to use your own dataset, you should also share information about the dataset with the instructor and TAs, by sending (i) a message on **Piazza**[1], describing key information about the dataset, and (ii) an email to **ludaesch@illinois.edu** with the subject "CS-513: Project Option (c)". Email your overview and initial assessment (described in the next section) of the dataset no later than **July 10**.

Recommended Workflow

The recommended overall workflow for the project should include the following phases:

1. **Overview and initial assessment of the dataset.** You should describe the *structure* and *content* of the dataset and *quality issues* that are apparent from an initial inspection. You should also describe a (hypothetical or real) *use case* of the dataset and derive from it some *data cleaning goals* that can achieve the desired fitness for use. In addition, you should answer the following questions: Are there use cases for which the dataset is *already* clean enough? Are there use cases for which the dataset will *not* be clean enough? You can speculate a bit here – but the rest of the project should focus on a "middle of the road" use case that requires a practically feasible amount of data cleaning.

2. **Data cleaning with OpenRefine.** In this first hands-on part of the project, you should use OpenRefine to clean the chosen dataset—either (a) or (b) or your own (c)—as much as needed for the use case. Document the process and result of this phase, both in *narrative form* along with *supplementary information* (e.g., which columns were cleaned and what changes were made?). Can you quantify the results of your efforts? Also, provide provenance information from OpenRefine. Pay close attention to what OpenRefine includes and does not include in its operation history! If important information is missing in the latter, provide that information in narrative form.

[1] You can use a private or public message. The latter allows other groups to also consider your favorite dataset for their project.

3. **[Optional] Data cleaning with other tools.** If you find that certain data cleaning steps are not well suited for OpenRefine (e.g. due to scalability or other issues), consider using an alternative, more suitable solution, e.g., Python, R, or other tools such as Trifacta Data Wrangler, Tableau, etc. Document your choice and answer the same questions as in Step 2.

4. **Developing a relational schema.** Develop a relational schema for your dataset. What logical *integrity constraints* (ICs) can you identify? Load the data into a SQLite database with your target schema. Use SQL *queries* to profile the dataset and to check the ICs that you have identified! You can also use other query languages such as Datalog to profile the dataset and check the ICs, but you should not use a procedural language such as Python, R, etc.

5. **Creating a workflow model.** Create a workflow model of your overall data cleaning workflow: What are the key inputs and outputs of your workflow? What are the dependencies? Note: Here you may want to model the various steps you have executed with OpenRefine as parts of the workflow. This way, the workflow model more clearly describes what actually happened to what parts of the data. Create a visual representation of your overall workflow using YesWorkflow or other diagramming tools. Supplementary material to help with YesWorkflow will be posted on Piazza. Also create a visual representation of your OpenRefine workflow using OR2YWTool (https://pypi.org/project/or2ywtool) or other appropriate tools. The OR2YWTool provides an auto-parsing method from Openrefine Operation History JSON file to YesWorkflow model (developed by Lan Li and Nikolaus Nova Parulian). Please include both overall workflow and OpenRefine workflow in your project report.

6. **[Optional] Developing provenance.** Develop provenance queries (in Datalog / DLV) that show on which inputs and intermediate data and steps the outputs of your workflow depend (cf. Provenance Assignment).

Project Deliverables

You need to submit the following deliverables through Coursera by **Aug 6**:

1. **Project Report.** The project report (a single **PDF** file) should contain all items mentioned in the Grading Criteria. In addition, it should contain the **name and netid** of each team member (in the front page) and the **contribution** of each team member (at the end of the report).

2. **Supplementary Materials.** In addition to the project report, you need to provide the following supplementary materials (as a single **ZIP** file).

    a. **Operation History:** A copy of the OpenRefine operation history (copy-paste it into a json file named **Open_Refine_History.json**). If you are using an alternative tool instead of OpenRefine, please provide an analogous history file (Other_Tool_History.json) and other provenance information (as available for that tool).
    b. **Queries:** A copy of the queries written in SQL or Datalog to profile the dataset and check the integrity constraints (copy-paste them into a plain text file named **Queries.txt**)
    c. **Workflow Model:** For the overall workflow model (using YesWorkflow or other diagramming tools), provide the file that has the annotations (e.g., **Overall_Workflow.txt**), and the generated Graphviz or DOT file (e.g., **Overall_Workflow.gv**). For the OpenRefine workflow, provide similar files.
    d. **Raw and Cleaned Dataset:** Please **DO NOT** provide the datasets in the ZIP file. Rather, upload the raw and cleaned datasets in a Box folder and share the link in a plain text file (**Data_Link.txt**).

Grading Criteria

The grade of the project would depend on the following parameters:

1. Overview and initial assessment of the dataset [25%]
    a. A clear description of the structure and content of the dataset [3%]
    b. A comprehensive list of data quality issues [7%]
    c. Identifying a feasible use case and the essential data cleaning goals [10%]
    d. Identifying use cases for which the dataset is already clean and use cases for which it will never be clean enough or usable [5%]
2. Data cleaning with OpenRefine (and other tools) [40%]
    a. Identifying the appropriate data cleaning steps for the use case [10%]
    b. A clear description of the data cleaning steps with supplemental information [20%]
    c. Quantifying the results of cleaning (e.g., provide a table of changes along with appropriate quantification) [5%]
    d. Provenance information from OpenRefine and/or other tools [5%]
3. Developing a relational schema [15%]

    a. Identifying the appropriate integrity constraints [5%]
    b. Loading data into a database with proper schema (show whether you load data directly from the SQLite prompt, via a script, or using a GUI) [3%]
    c. Writing queries to check the integrity constraints [7%]
4. Creating a workflow model [10%]
    a. Identifying the key inputs, outputs of your workflow along with the dependencies [3%]
    b. A visual representation of your overall workflow, e.g., using YesWorkflow [4%]
    c. A visual representation of your OpenRefine workflow, e.g., using OR2YWTool [3%]
5. Other factors [10%]
    a. Clarity and presentation of the report [5%]
    b. Further analysis/takeaways/challenges [5%]

Past Project Reports

Here are a few project reports from the previous offerings of the course:

1. https://drive.google.com/file/d/1WQ81ZFmG3LwlPOGyncL2udR7pJzXaGvg/view?usp=sharing
2. https://drive.google.com/file/d/17r7z1nj5UXganjzdr7n8Gbf0XtqmQ1rU/view?usp=sharing

N.B.: The purpose of sharing the past project reports is to provide examples. Following the methodology or presentation of past reports do not guarantee a full score. In fact, the grading criteria for this offering (of the course) is different from the past offerings.