CS 598 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.

Forming Groups. You are encouraged to form teams of 2 or 3 students, but the project can also be conducted individually. Use Slack or Piazza to find teammates. Individuals would typically choose option (a) below (to leverage prior work), while groups are encouraged to make use of options (b) or (c), and should also try and tackle the optional parts of the project.

Start by reviewing the web sites (a) and (b) for which datasets to be cleaned exist, 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/`), or
- (c) Do you have access to another dataset that you'd like to work with?

Reference versions for (a) and (b) will be provided; if you choose your own dataset (c), use a dataset that is publically available (preferred) or that you are allowed to share (with your teammates and the instructor/TA-team). If you plan to use your own dataset, you should also share information about the dataset with the instructor and TAs, by using (i) a message[1] on Piazza, describing key information about the dataset, *and* (ii) sending email to `ludaesch@illionois.edu` with subject "CS-598: Project Option (c)". Email your overview and initial assessment of your dataset by July 16.

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

1. **Overview and initial assessment** of the dataset (*narrative* and *supplemental information*). 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: Are their use cases for which the dataset is *already* clean enough? Others for which it well *never* be good 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 result of this phase, both in *narrative* form and with *supplemental* 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 other ways.

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

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

4. **Develop a relational database 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!

5. **Create a workflow model** of your 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 YesWorkflow model more clearly describes what actually happened to what parts of the data. Create a visual version of your workflow using the YesWorkflow tool. Supplementary material to help with YW will be posted on Piazza.

6. (***Optional***) 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).

How to submit. Additional information and discussions about the project, including details of how to submit will be via Piazza.