LIS590DCL Theory & Practice of Data Cleaning | Spring 2017
Dr. Bertram Ludäscher | April 10, 2017

Assignment 4 – Group Project

Due Dates: May 1 (Progress Update / Slide Presentation) and May 10 (Final Report)

Reading Assignments: See Moodle postings.

The goal of the last assignment is to combine a number of tools and techniques covered in this course together in a small end-to-end data cleaning workflow. Starting either from one of the provided datasets

- (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) a dataset of your own choosing.

You are encouraged to explore the web sites (a) and (b). For the assignment, reference versions of (a) and (b) are provided on the course Moodle.

If you plan to use your own dataset, you need to share information about it via Moodle by the end of the week (**April 16**), to allow (i) the instructor to evaluate the proposed dataset and (ii) other groups to possibly adopt that dataset as well.

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) use 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 as much as needed. 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 solution, e.g., using Python or R. Document your choice.

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 YW model more clearly describes what actually happened to what parts of the data. Create a visual version of your workflow using the YesWorkflow tool.

6. (Optional) Develop provenance queries (in Datalog / DLV) that show on which inputs and intermediate data and steps the outputs of your workflow depend.

Forming Groups. Groups should consist of 3 students. In exceptional cases groups of 2 or 4 students can be formed. For groups of size 4, dataset options (b) or (c) should be used and your group should also tackle the optional parts of the assignment.

How to submit. Additional information about the deliverables will be posted via the Moodle.