IS-537: Theory & Practice of Data Cleaning: Project Guidelines

The final project should be written as a narrative, i.e., in paragraph format. Bullet points can be used for enumeration purposes to provide additional detail, but not to replace the narrative. Screenshots of relevant parts of the overall data cleaning workflow are welcome (but pages and pages of screenshots, possibly with sparse text in-between isn't welcome.)

Here is a typical workflow with some of the key phases that you should include for your final project. Note: if you have ideas that *dramatically* differ from this recommended workflow and the associated structure of the final project report, please check with the TA team and/or the instructor.

1. **Overview and initial assessment of the dataset.** You should describe the *structure* and *content* of the dataset and the *data quality issues* that are apparent from an initial inspection. You should also describe (hypothetical or real) *use case(s)* 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 (no need to touch the data for those use cases)? Are there use cases for which the dataset will *never* be clean enough? That is, are there use cases that seem initially plausible for the given data, but that on closer inspection seem infeasible (e.g., because key information is missing, or because the low data quality makes a satisfactory repair impossible). Apart from these two kinds of use cases, the main focus of your project should be a "*middle of the road*" use case, i.e., one that requires a practically feasible amount of data cleaning.

2. **Data cleaning with OpenRefine.** You can use OpenRefine to clean your dataset 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?) and *screenshots* of key cleaning steps. Can you quantify the results of your efforts (e.g. in a summary table)? 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.
   Data cleaning with other tools. If you find that your 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 Prep, etc. Document your choice and answer similar questions as mentioned above.

3. **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! In particular, write "denial queries" that report IC violations (if any). 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. unless you can justify your choice.

4. **Creating a workflow model.**
    a. **Overall workflow**: Create a workflow model of your ***overall*** data cleaning processes. Here you may want to model the "big picture" of each phase in your project (e.g. data profiling phase, OpenRefine phase, IC checking phase). Create a visual representation of your overall workflow using YesWorkflow or another diagramming tool. Hint: think about what the key inputs and outputs of your workflow are, and what dependencies exist between different workflow steps.
    b. **OpenRefine (or analogous) workflow:** If you have used OpenRefine, then create a visual representation for those OpenRefine cleaning steps. You can use the 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). If you've used other tools rather than OpenRefine, create a workflow model specific to the steps you've taken in section 2. You can use other diagramming tool(s) of your choice as well.

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

6. **Conclusions.** Can you showcase (in any way you like) a summary of your dataset before vs. after cleaning? What are the takeaways of the project? Have you encountered any problems or challenges along the cleaning process? Describe them.

7. **[For extra credit] Data analysis**. Data analysis or visualizations after all the cleaning processes that addressed your use cases.

Project Deliverables

You need to submit the following deliverables (see class page for deadlines etc.)

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 except your datasets).

    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) or 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 the beginning of your final report.

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 [25%]
    c. Summary of the results of cleaning (e.g., provide a table of changes along with appropriate quantification) [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. A visual representation of your overall workflow, e.g., using YesWorkflow or other diagramming tools [7%]
    b. A visual representation of your OpenRefine workflow, e.g., using OR2YWTool [3%]
5. Other factors [10%]
    a. Conclusions [5%]
    b. Clarity of final report [5%]