CS513: Theory & Practice of Data Cleaning

Final Project Task Details

ROHIT, GOUTAM, DEB

TARGET DATE: JULY-3-2022

GOAL: Conduct end-to-end data cleaning projects using various tools.

TOOLS:

Regular Tools: RegEx,
OpenRefine,
Datalog,SQL,
Python,
YesWorkflow ,
OpenRefinecompanion tools such asor2yw

Commercial tools: Trifacta Wrangler,
Tableau

TASKS - Phase 1:

1. Identify dataset: Farmers Market Data.

2. Develop target/main Use Case(U1).

   How do you specify data analysis use cases? You can simply explain the use case in a short paragraph. You might also want to be more specific and phrase use cases as questions: What is it that we want to know from (or about) the data?

   In particular, a use case may be a set of database queries $Q_1, \ldots, Q_n$ against the dataset D(e.g., how many farmers markets offer bakery goods in addition to vegetables and fruits?) On the other hand, use cases may also be more general, e.g., you could state that you'd like to develop a web application that serves a particular purpose.

3. Two minor use cases.
   a. U0: "Zero data cleaning" i.e good enough as it is.

   b. U2: “never (good) enough”, i.e., no amount of data cleaning or wrangling will make data usable.

4. Creating Conceptual Model
   a. ER Diagram: entity types and relationship types
   b. Ontology: illustrates the main classes and their relationships
   c. Database Schema: illustrates and explains the structure and contents of the dataset

5. Short Narrative: One or more paragraphs in English to describe the origin of the data and any relevant metadata (e.g., a temporal or spatial extent). A dataset about farmers markets, e.g., can be described with a relational schema (e.g.,CREATE TABLEstatements); the narrative would then explain what the different columns (attributes) mean. Other metadata may describe, e.g., the spatial extent of the data (only Illinois markets? All the Midwest? Or the US?), and the temporal extent (for which period is the data correct?)

6. List Obvious Data Quality Problem: This should show the need for data cleaning to achieve U1. You need to support this claim by documenting data quality problems that your inspection has revealed and that need to be addressed before U1 can be tackled.

   One simple way is to include(copy-pasted) snippets of “dirty data” in your Phase-I report (you can also use screenshots for illustration) and then explain what the problem is in narrative form.

7. Devise an Initial Plan: Outlines how to clean the data for Phase 2.

   short list of your planned stepsS_1, . . . , S_5willdo during Phase-I.

   a. S_1: description of dataset D and matching use case U1.

   b. S_2: profiling of D to identify the quality problems P that need to be addressed to support U1.

   c. S_3: performing the data cleaning process using one or more tools to address the problems P (here you should describe which tools you are planning to use, e.g., OpenRefine;Python; etc.)

d. S_4: checking that your new dataset D_1 is an improved version of D, e.g., by documenting that certain problems P are now absent and that U1 is now supported.

e. S_5: documenting the types and amount of changes that have been executed on D to obtain D_1.

8. assignment of tasks to team members (who does what).