Final Project

End-To-End Data Cleaning Workflow – NYPL Data

by

Sudhir Behani (sbehani2@illinois.edu)

Bollam Raja Shekar

Table of Contents

Introduction and Overview

1. Dish.csv

Information about all the dishes from all the menus transcribed by the project are stored in data.csv. In this data file, a dish is represented by a row of values. Columns identify attributes of a dish. One of these attributes is an identifier, which identifies the dish. However, the identity of a dish appears to be based on the exact form of the string labeled "name." Thus, dishes with variant orthographic forms of their names, e.g. “half chicken”, “Half Chicken” and “chicken [half]”) are treated as separate entries with different identifiers.

2. Menu.csv

Information about the menus as physical objects, including historical information about their origins, uses, and formats are stored in menu.csv

3. MenuItem.csv

This is the largest data file. A "MenuItem" represents a single instance of a dish appearing somewhere on a menu page image. “MenuItem.csv” is useful as a mapping between multiple other data files/tables.

4. MenuPage.csv

Information about individual pages of the menus is stored in "Menu.csv". Pages are modeled here as digital images produced as a result of digitization by the NYPL. Menus often have multiple pages.

5. Missing Data

We note the presence of missing data points in the column “missing values.” However, there are also strings in the present data that point to missing information (e.g. “unknown” or “?”), which do not formally appear as null values.

Other Information

The data in these files comes from several different sources. We reflect this in the “generated by” category of the tables below. Some of the data is supplied by “volunteer transcribers,” by which we mean people who have participated in the project through the What’s On the Menu? site. Some of the information is generated by “web application.” This means that some of this information was automatically created as the database supporting the application was constructed and populated (e.g., various ids); some is created as the web application runs (e.g. timestamps as data values are updated). Finally, a lot of information is generated from “NYPL metadata.” This metadata comes from many places and reflects the long history of the project and the many parts of New York Public Library involved in it. Much of the data “supplied by NYPL metadata” in the menu spreadsheet is from the catalog cards made by Frank E. Buttolph in the early twentieth century.

Initial Assessment of the dataset

Data Cleaning methods and Process

Open Refine Tasks

1. Menu

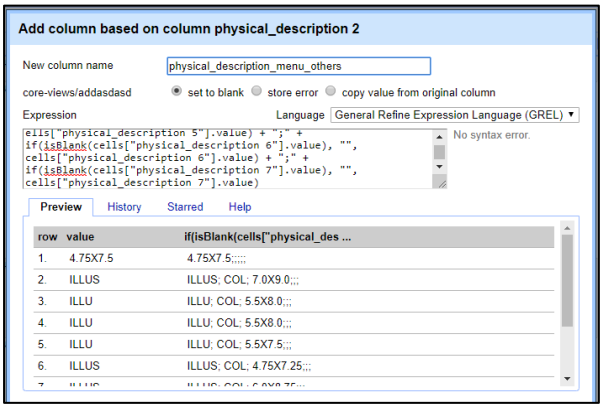
   Activities performed:

- Trim leading and trailing white spaces
- Collapse consecutive white spaces
- Convert columns to number type
- Convert columns date to format YYYY-MM-DD
- Remove special characters “()[]{}\?#!%” from the text columns. Also, remove characters at the end of the string like semicolon, comma and hyphen “;,-”

  Note: Can’t remove (semicolon) character in the middle of string as it signifies

  multiple options- example multiple sponsors, event, venue, place etc
- Remove continuous hyphen characters from the text columns
- Convert all values in columns with string data type to upper case so that problem in clustering similar values is reduced considerably
- After removing special characters again perform white spaces removal for each the string data type columns
- Clusters columns based on method key collision and keying function as fingerprint. Accepted the default values for new cell value. Merge cluster and then perform ngram-fingerprint using ngram size as 2. Metaphore3 and cologne-phonetic were mostly creating clusters with somewhat different names too so were not used.
- Physical_description seems to contain information about the menu hard copy format and its size. Menu could be in different forms – accordian fold, book, booklet, broadside, card, folder, tri-folder, two cards joined by ribbons etc

  Split the physical_description column with separator as semicolon. It is split into 7 columns. Rename the first column as “physical_description_menu_type” and join other columns as rename it as “physical_description_menu_others”



- Remove semilon at the end of the string using expression

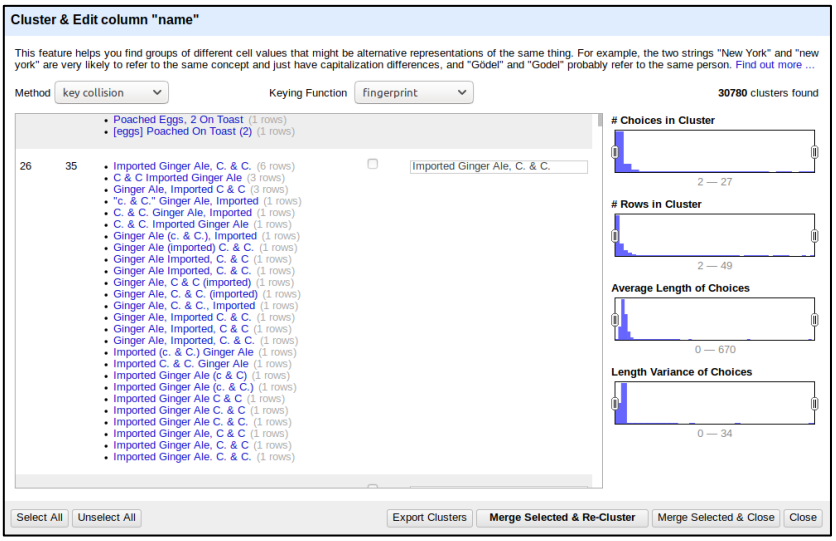
Not updated columns:

Columns which are not updated: id, name, keywords, language, location_type, currency, currency_symbol

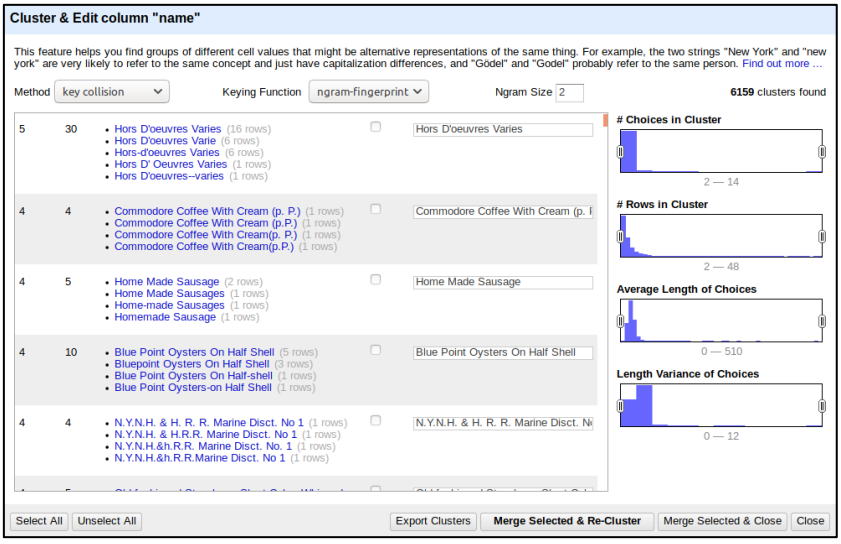
1. Dish

Activities performed:

- Trim and Collapse white spaces in “name” column.
- Convert column “name” to title case.
- Make a facet and perform the cluster operation using the “key-collision” method and “fingerprint” as shown below.



- Repeat step c with “ngram-fingerprint” as shown below.



- Remove “description” column, since most of the entries are empty.

2. MenuPage

- Convert columns to number type

There are no other columns cleaned

3. MenuItem

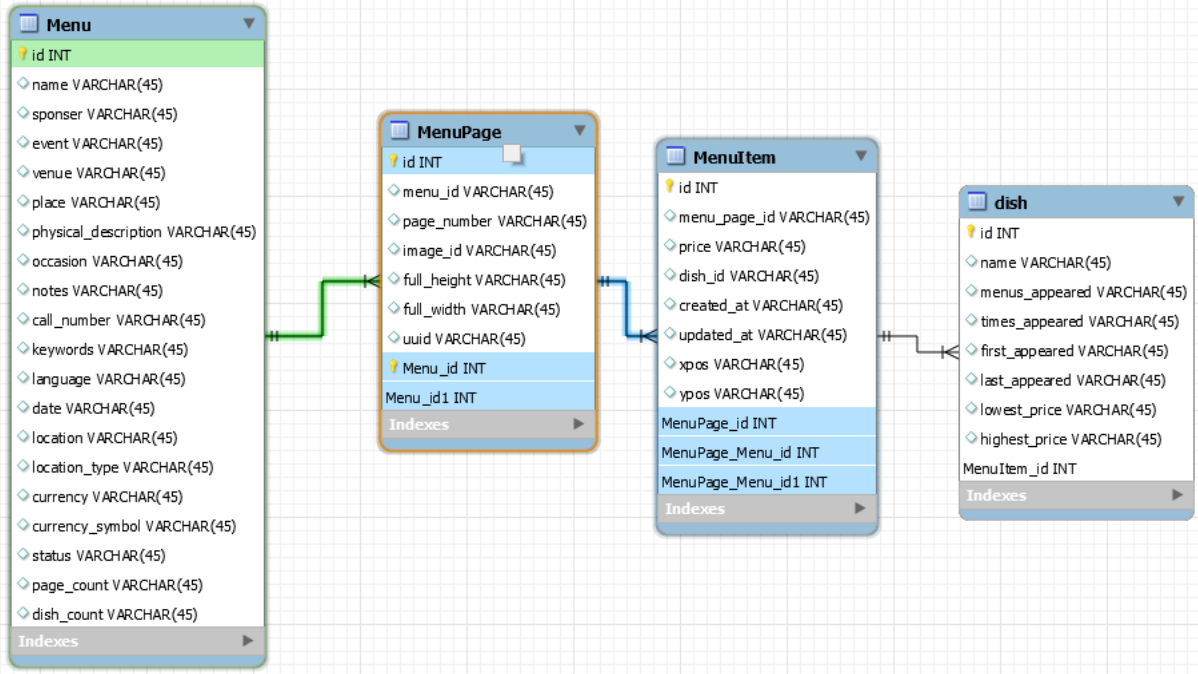
Following data cleaning steps were performed on the Dish dataset.

- “high_price” column is removed
- “created_at” is transposed into “toDate()”
- “updated_at” is transposed into “toDate()”

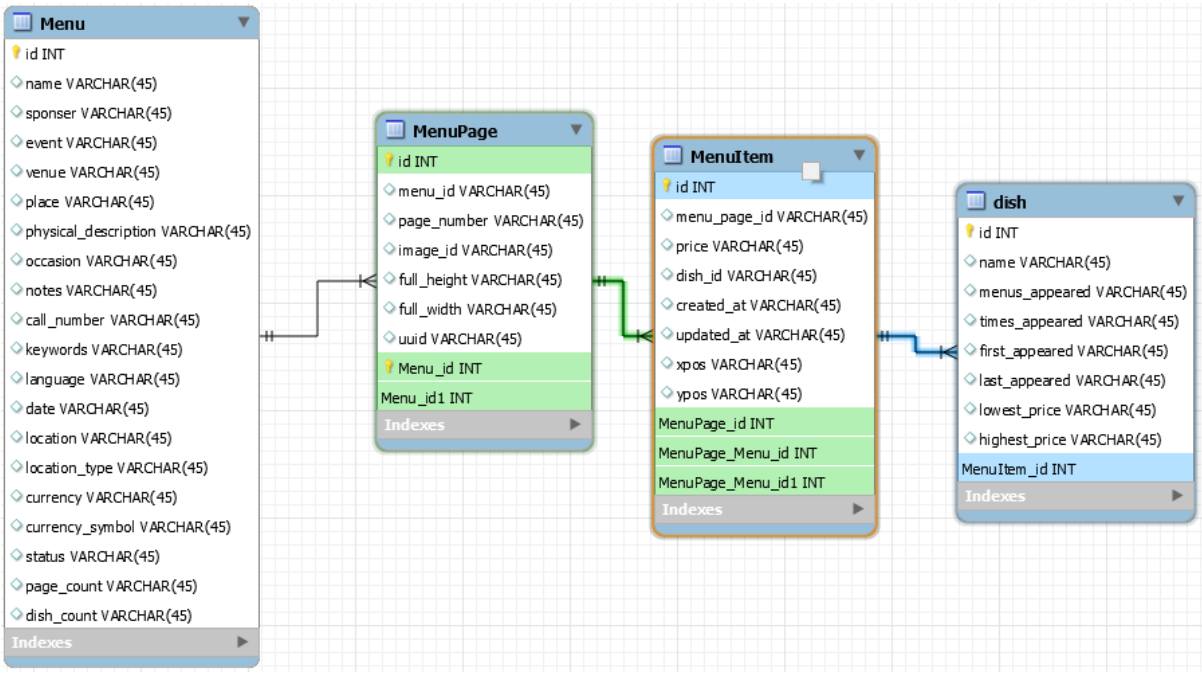
Data Cleaning Results

Relation Database Schema

ER Diagrams



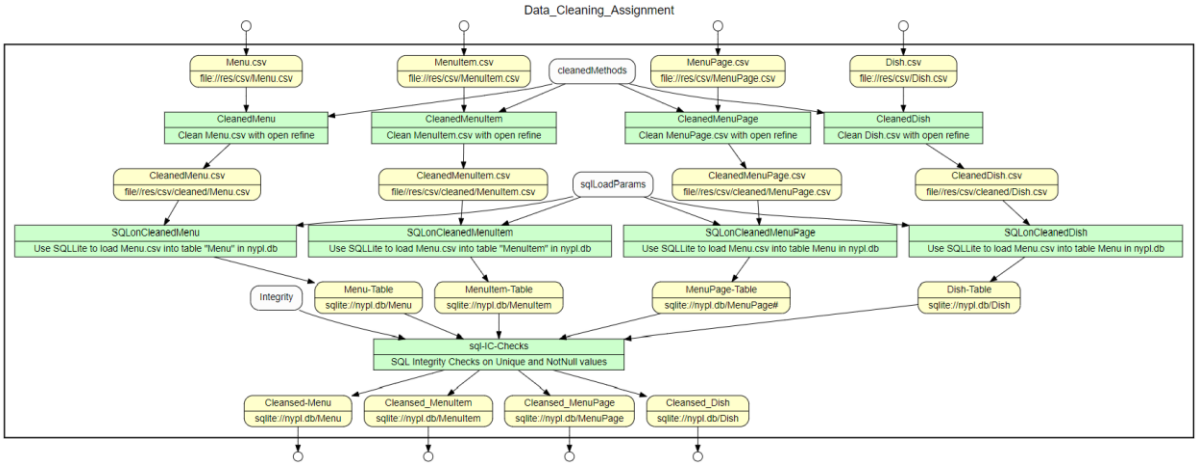
<<<Add text >>>



Integrity Constraint Checks

Workflow Model

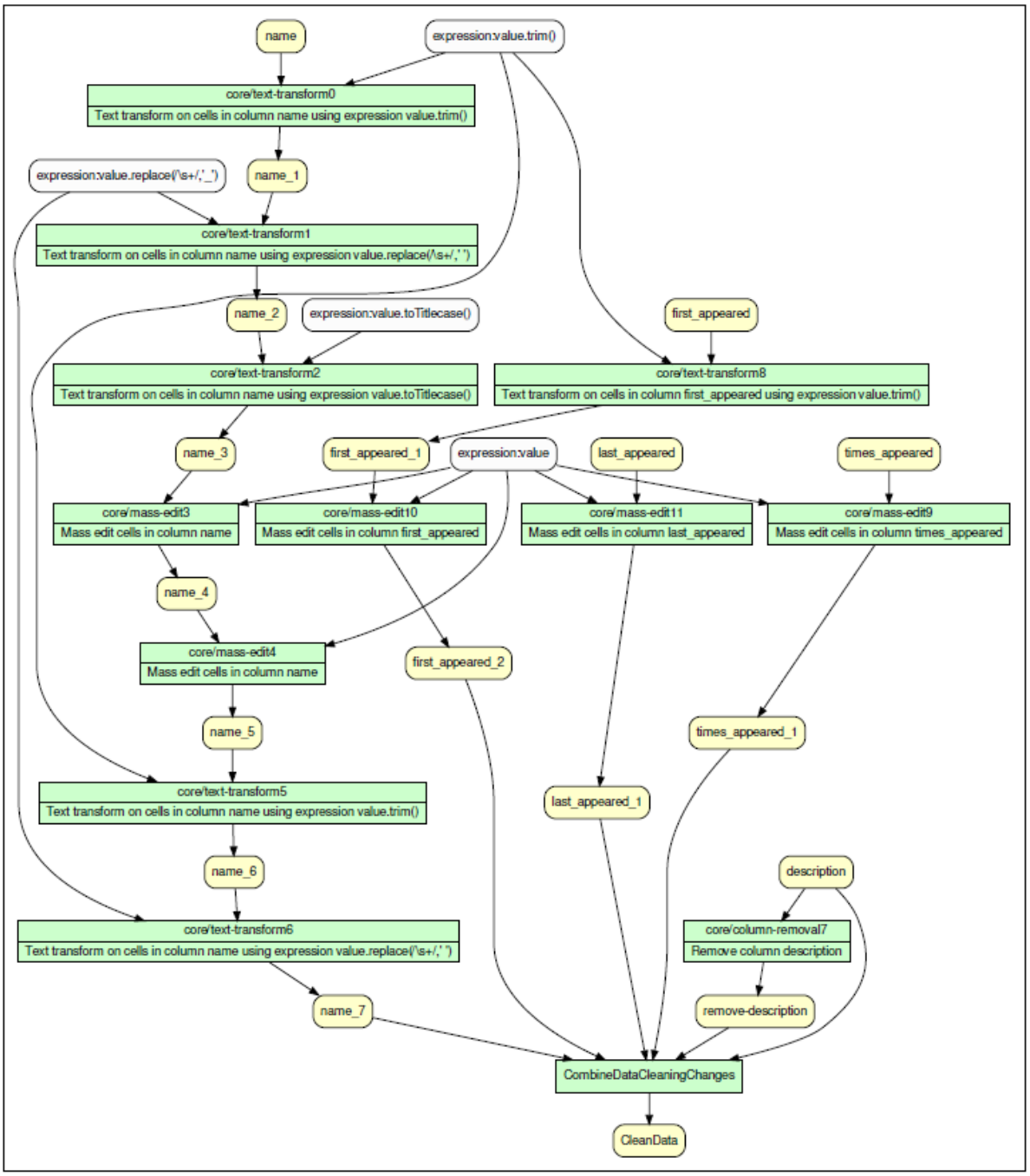
Or2yw overall workflow



Or2yw open Refine csv

Following graphs are generated using or2yw tool based on the recipes generated from the openrefine too.

1. Dish.csv



2. Menu.csv

Overview of changes

Conclusion and Future Work

Data cleaning course helped us to understand the importance of having a cleaned dataset. NYPL being a crown sourced dataset is full of issues and needs good amount of cleaning activities. Majority of the cleaning works were required for string-based columns in the dataset files. As textual information could be written in different ways by different people.

Main outcome of our cleaning project is to create a cleaned dataset. Though this is a huge task in hand given the time and effort required to bring data into a completely consistent state so that this could be useful for further tasks like machine learning or natural language processing. We also tried our best to create a yesworkflow visualization so that its easier for clients to get visual details about the cleaning activities performed.

Important takeaways from this cleaning activity is to first get an understanding about the importance of cleaned dataset. This was understood once we investigated the NYPL dataset and cleaned data was user readable and one could develop better understanding of the information. A simple removal of leading and trailing whitespaces and removal of consecutive white spaces bring so much consistency to the dataset. It is utmost importance to create a cleaned dataset. From NYPL dataset it was also understood that crowd source dataset needs a lot of cleaning activities.

There were number of challenges ….

Problems or Challenges:
Additional steps:
Certain foreign letters which needs to be take care by studying the words in more detail. These foreign characters seems to occur in menu table under columns – sponsor and location.

Future additional tasks:
For Menu we could use status column having two values “Complete” and “Under Review” and perform classification task. We could try to predict which features in menu table led to the value Complete and Under Review.

Appendix A
Link to project deliverables: