Data Preparation: New York Public Library Menu Dataset

The NYPL Menu dataset is a recorded menu in a venue/restaurant that has been existed through time in many places. This dataset is compiled from human input or scanned from the physical menu. The fields that we cleaned on this project:

- name: stores the name that appear in the menu;
- sponsor: stores sponsor name, this field looks mostly the same as the name field. The sponsor name should be more general than the name itself in which will be part of the data cleaning process;
- event: stores which event that can be served in the particular place, for ex: breakfast, dinner, or lunch;
- venue: stores which kind of event that can be served in the place, for example, commercial, group, private party;
- occasion: stores in which occasion this menu was used, for example, new year, Easter, daily menu, because different occasion may affect the menu;
- place: place field stores the street, city, or country,
- location: location stores the specific building

We want to do some data visualization and aggregation through the dataset, but during our exploration, we found that there are several inconsistencies in the values and decided to do these manual data cleaning steps using OpenRefine

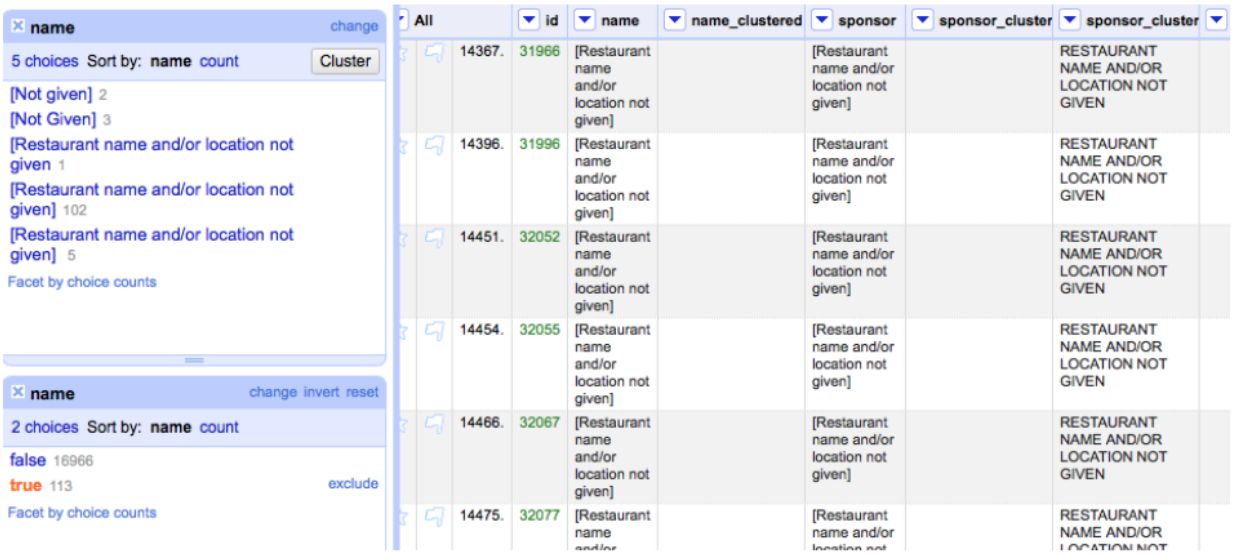
1. The consensus of our data cleaning execution is, for each field that we want to clean we made a new column base on that column (copy the column to a new column) to keep the original column from changing and or compare what we have done with the original one side by side. The aforementioned new columns are named the same as the original column's name with clustered in the back, for example, name -> name_clustered, sponsor -> sponsor_clustered.

2. Basic cleaning operations are performed on the field of interests following these procedures:
    - trim leading and trailing space because some data really have problem with these spaces; after that, we collapse consecutive whitespace to minimize error of typing; and we made all the letter capital in order to make it easier to cluster and check the data for merging operation.
    - In addition to the basic transpose column, we also eliminated unusual characters using custom transformation using regular expression following

this function value.replace(/[%@#!\\\[\]\"\']/,'') in which it will clean most of the characters that aren't needed in the data clustering.

- After done those operations, we do facet and cluster operations in order to find some correlations between the value in one column. For some clusters result that have confidence correlations, we merged them basically making a canonical vocabulary for the column of interests.

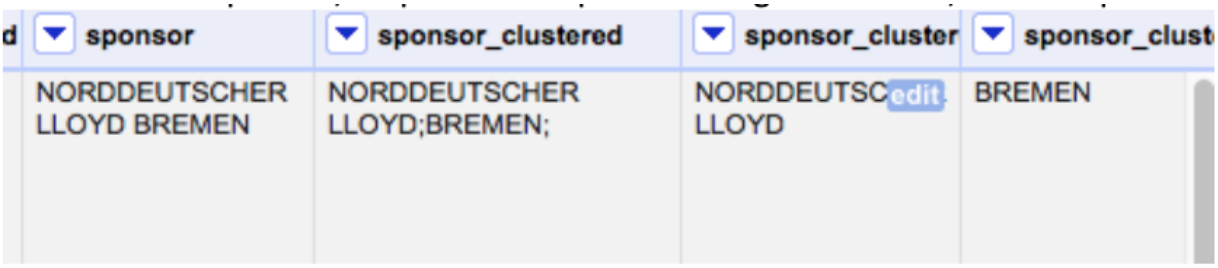
3. The name_clustered column has a rather interesting value. Firstly, it has data contain "[NOT GIVEN]", "[RESTAURANT NAME AND/OR LOCATION NOT GIVEN]". We prefer to replace these values with a blank (null) to make it consistent and easier to distinguish real data and data not given. To make sure we change the right data I apply a custom facet using this function, value.match(/[\[\]].*/) != null, which will give me a list of data that have "[" and "]" resulted like in the picture below



4. After done cleaning the NOT GIVEN value, we do clustering and merging for the column and specifically for some value that followed by HOTEL, RESTAURANT, or CAFÉ, I prefer to change the value following the GENERAL to SPECIFIC pattern, therefore ASTON HOTEL will be HOTEL ASTON, to make it easier to classify the name type.

5. For sponsor_clustered. In general, this column looks the same as the name column but the data stored here are more detailed. Just like the name column, we change NOT GIVEN value in this column following the exact procedure that

we have done to the name column. An additional finding that we found in the sponsor_clustered data is most of the rows have a pattern of “sponsor name – sponsor location – sponsor city” pattern. Thus we divided this column using semicolon “;” like this picture below:



d	sponsor	sponsor_clustered	sponsor_cluster	sponsor_clust
	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD;BREMEN;	NORDDEUTSC edit LLOYD	BREMEN

6. On Venue_clustered column: firstly it contains “?” character which is equal to the blank value, therefore we simply replaced the data that have only “?” value in the cells into blank. Similar to the sponsor, we also found the semicolon pattern in this field; therefore, we divide them using a semicolon (“;”) to standardize and make it easier for aggregation/grouping.

After all these data cleaning transformations, we can ingest the new dataset to a visualization/aggregation tool such as Tableau / Microsoft Excel for reporting.