

CS513: Theory & Practice of Data Cleaning Project

Phase II - Report

University of Illinois at Urbana-Champaign Summer 2023

Team 140 - Akriti Sinha, Christopher Buja, Tianhao Luo

Team 140 information :

Members: Akriti Sinha (akritis5)
Christopher Buja (cbuja2)
Tianhao Luo (tluo3)

1. Actual data cleaning workflow

Step I: Python (stage1.ipynb)

Input file : Menu.csv, Dish.csv, MenuPage.csv, MenuItem.csv

1. In the string columns (such as occasion), delete the following characters
() [] ; ?
2. Also, for the string columns, convert NaN to "Unknown"
3. Every single name is converted to title case
4. This transformation has been applied to
 - a. 'name' column of 'dish' dataset
 - b. 'name', 'sponsor', 'event', 'venue', 'place', 'occasion', 'notes', 'location' columns of 'menu' dataset

Output file : Menu_stage1.csv, MenuPgae_stage1.csv, MenuItem_stage1.csv, Dish_stage1.csv

Step II: OpenRefine (TBD)

Cleaning of Menu_stage1.csv

For column: sponsor and event

- 1) Trim leading and trailing white spaces
- (2) Collapse consecutive white spaces
- (3) Convert all column values to upper case
- (4) Remove special characters using GREL (% , # , ! , / , (,) , [,] , ?)
- (5) Replace “,” with a space instead and then trim leading/trailing white spaces
- (6) Make a facet and perform the cluster operation using the key-collision method and fingerprint function. Next merge the selected clusters.
- (7) Repeat the previous step with n-gram, fingerprint, meta-phone3, and cologne-phonetic methods.

Cluster and edit column "sponsor"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method Keying function n-Gram size 75 clusters found

Count	Choices	Key
2	<ul style="list-style-type: none">SOCIETA LA PIEMONTESESOCIETA'LA PIEMONTESE	<input checked="" type="checkbox"/> SOCIETA LA PIEMON
3	<ul style="list-style-type: none">HOTEL DUPONT (2 rows)HOTEL DU PONT	<input checked="" type="checkbox"/> HOTEL DUPONT
6	<ul style="list-style-type: none">A.H. MEYER RATHSKELLER (3 rows)A.H.MEYER RATHSKELLER (3 rows)	<input checked="" type="checkbox"/> A.H. MEYER RATHSK
13	<ul style="list-style-type: none">NIPPON YUSEN KAISHA - S.S.KASUGA (9 rows)NIPPON YUSEN KAISHA - S.S. KASUGA (4 rows)	<input checked="" type="checkbox"/> NIPPON YUSEN KAIS
8	<ul style="list-style-type: none">LAUREL IN THE PINES (4 rows)LAUREL-IN-THE-PINES (4 rows)	<input checked="" type="checkbox"/> LAUREL IN THE PINE
2	<ul style="list-style-type: none">S. S. PRESIDENT WILSONS.S. PRESIDENT WILSON	<input checked="" type="checkbox"/> S. S. PRESIDENT WIL
2	<ul style="list-style-type: none">BELLEVUE - STRATFORDBELLEVUE-STRATFORD	<input checked="" type="checkbox"/> BELLEVUE - STRATF

Choices in cluster
2 — 3

Rows in cluster
0 — 790

Average length of choices
0 — 49

Length variance of choices
0 — 2

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

For column: physical_description

- (1) Split the columns using ‘;’
- (2) Then rename the first column: physical_description_type
- (3) Use GREL to join ‘physical_description 1’, ‘physical_description 2’, ‘physical_description 3’, and ‘physical_description 4’.

(4) One column 'physical_description 1' renamed to physical_description_type.

(5) 'physical_description 2', 'physical_description 3', and 'physical_description 4' are combined using ' - ' to separate the values from the different columns (Remember the space before and after the dash) to column physical_description_additional.

For column: date

(1) Convert date format to YYYY-MM-DD

For column: call_number and id

(1) Trim leading/trailing white spaces

(2) Collapse consecutive white spaces

Unchanged columns: name, keywords, language, status, page_count, dish_count

Output file : Menu_clean.csv

Cleaning of MenuPage_stage1.csv

There was nothing to be refined here so the

Unchanged columns: id, menu_id, page_number, image_id, full_height, full_width, uuid

Output file : MenuPage_clean.csv

Cleaning of MenuItem_stage1.csv

For column: created_at

(1) Convert date format to YYYY-MM-DD

For column: updated_at

(1) Convert date format to YYYY-MM-DD

Unchanged columns: id, menu_page_id, price, high_price, dish_id, created_at, updated_at, xpos, ypos

Output file : MenuItem_clean.csv

Cleaning of Dish_stage1.csv

For column: name

(1) Use key-collision to cluster values.

Output file : Dish_clean.csv

Step III : Develop Relational Database Schema

Schema : Below, is the schema in which there are four tables (one for each input file): dish, menuitem, menupage, and menu for the cleaned data.

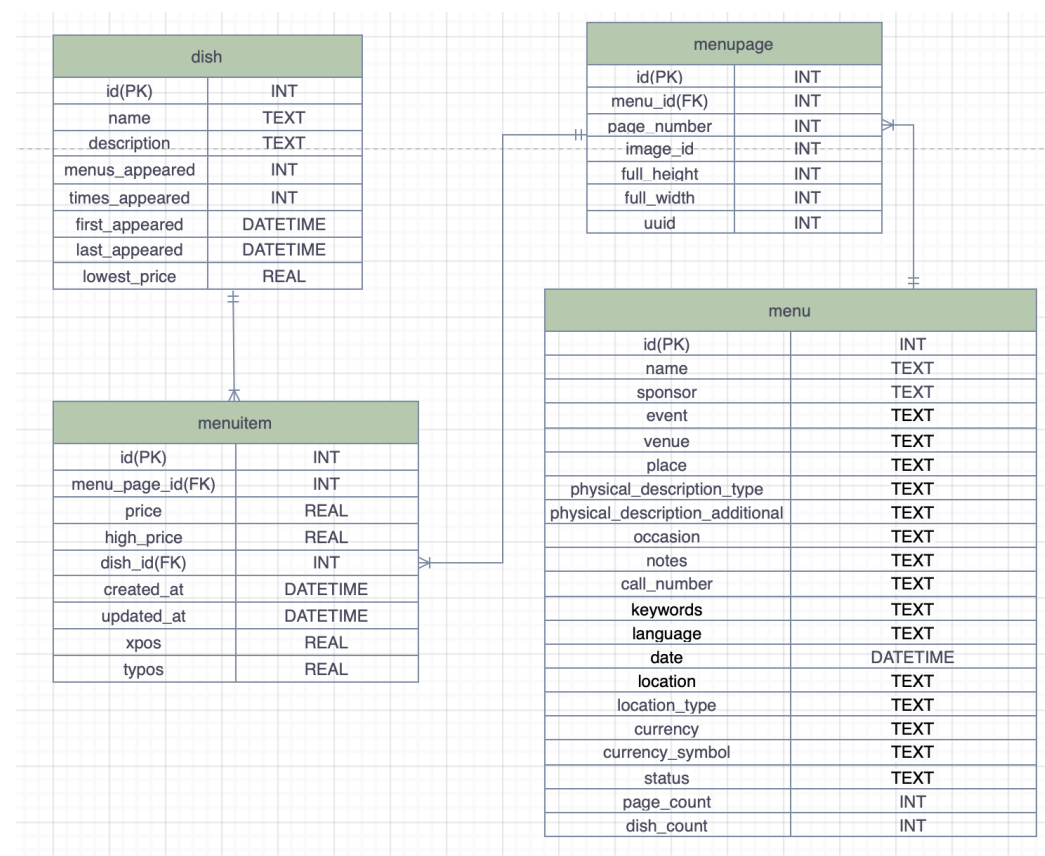


Table creation and Data load

SQL to create Database as per the Schema defined above.

Comparison with Phase I

- i. One dish can be mapped to several IDs

515676 Claret: Chateau Larose, Cruse et
Fils Freres

515677 Claret: Chateau Lafite, Cruse et
Fils Freres

On a closer look, although these two dishes bear striking resemblance, on a closer look they can be two distinct dishes, 'Larose' and 'Lafite'. We should not attempt to merge them without 100% confidence that they are exactly the same dish.

- ii. A dish has more 'menus_appeared' than 'times_appeared'

id	name	description	menus_appeared	times_appeared	first_appeared	last_appeared	lowest_price	highest_price
208	Luncheon	NaN	19	18	1900	1993	0.65	0.65
825	Rice, Semolina	NaN	1	0	1900	1900	0.00	0.00
1082	Caviare	NaN	86	85	1888	1906	0.40	0.50
1136	Carta blanca	NaN	8	7	1900	1981	0.00	0.00
1346	Hackley's Sour Mash	NaN	5	4	1900	1900	0.15	0.15
...
i15598	Apricot C�mpote	NaN	1	0	0	0	0.00	0.00
i15599	Pear C�mpote	NaN	1	0	0	0	0.00	0.00
i15600	Guava C�mpote	NaN	1	0	0	0	0.00	0.00
i15601	Bilberies C�mpote	NaN	1	0	0	0	0.00	0.00
i15602	Fig C�mpote	NaN	1	0	0	0	0.00	0.00

While this is definitely some data quality problem, we decide to shelve it because we are unable to deduce a dish's a) menus appeared b) times appeared from the dataset.

iii. A same occasion can have different names

'DAILY;', 'DAILY'
'ANNIVERSARY(?);', 'ANNUAL','OTHER (ANNUAL DINNER);'

This has been taken care of in Python (by getting rid of characters such as (?);[], and converting cases) the OpenRefine part (using ABC distance metric)

iv. (Not directly related to the use case) Some menus do not have continuous pages

2. Narrative that ties all steps together

For the main use case U1, we want

- What are the most popular (in terms of time of appearances on all menus in the dataset) dish under each occasion (e.g. Easter, Thanksgiving, Christmas etc.)?
- What are the most popular breakfast/lunch/dinner dishes?

The data cleaning steps described above can help us

- Combine all variations of a same occasion under the same category
e.g. Anniversaries can have variations such as 'ANNIVERSARY(?);',
'ANNUAL','OTHER (ANNUAL DINNER);'
- Make the occasions more human readable (getting rid of special characters etc.)

3.Documentation

a. Python

	id	name	description	menus_appeared	times_appeared	first_appeared	last_appeared	lowest_price	highest_price
0	1	Consomme Printaniere Royal	NaN	8	8	1897	1927	0.20	0.4
1	2	Chicken Gumbo	NaN	111	117	1895	1960	0.10	0.8
2	3	Tomato Aux Croutons	NaN	13	13	1893	1917	0.25	0.4
3	4	Onion Au Gratin	NaN	41	41	1900	1971	0.25	1.0
4	5	St. Emilion	NaN	66	68	1881	1981	0.00	18.0
5	7	Radishes	NaN	3262	3346	1854	2928	0.00	25.0
6	8	Chicken Soup With Rice	NaN	48	49	1897	1961	0.10	0.6
7	9	Clam Broth Cup	NaN	14	16	1899	1962	0.15	0.4
8	10	Cream Of New Asparagus, Croutons	NaN	2	2	1900	1900	0.00	0.0
9	11	Clear Green Turtle	NaN	157	157	1893	1937	0.25	60.0
10	12	Striped Bass Saute, Meuniere	NaN	2	2	1900	1900	0.00	0.0
11	13	Anchovies	NaN	453	484	1858	1987	0.00	30.0
12	14	Fresh Lobsters In Every Style	NaN	4	4	1899	1900	0.00	0.0
13	15	Celery	NaN	4246	4690	1	2928	0.00	50.0
14	16	Pim-Olas	NaN	145	148	1897	1918	0.15	35.0
15	17	Caviar	NaN	505	534	1880	1987	0.00	75.0
16	18	Sardines	NaN	1425	1484	1856	2928	0.00	50.0
17	19	India Chutney	NaN	16	16	1865	1901	0.10	0.2
18	20	Pickles	NaN	453	472	1852	1987	0.00	10.0
19	21	English Walnuts	NaN	83	86	1851	1948	0.10	0.3

Data: Dish (before Python cleaning)

	id	name	description	menus_appeared	times_appeared	first_appeared	last_appeared	lowest_price	highest_price
0	1	Consomme Printaniere Royal	NaN	8	8	1897	1927	0.20	0.4
1	2	Chicken Gumbo	NaN	111	117	1895	1960	0.10	0.8
2	3	Tomato Aux Croutons	NaN	13	13	1893	1917	0.25	0.4
3	4	Onion Au Gratin	NaN	41	41	1900	1971	0.25	1.0
4	5	St. Emilion	NaN	66	68	1881	1981	0.00	18.0
5	7	Radishes	NaN	3262	3346	1854	2928	0.00	25.0
6	8	Chicken Soup With Rice	NaN	48	49	1897	1961	0.10	0.6
7	9	Clam Broth Cup	NaN	14	16	1899	1962	0.15	0.4
8	10	Cream Of New Asparagus, Croutons	NaN	2	2	1900	1900	0.00	0.0
9	11	Clear Green Turtle	NaN	157	157	1893	1937	0.25	60.0
10	12	Striped Bass Saute, Meuniere	NaN	2	2	1900	1900	0.00	0.0
11	13	Anchovies	NaN	453	484	1858	1987	0.00	30.0
12	14	Fresh Lobsters In Every Style	NaN	4	4	1899	1900	0.00	0.0
13	15	Celery	NaN	4246	4690	1	2928	0.00	50.0
14	16	Pim-Olas	NaN	145	148	1897	1918	0.15	35.0
15	17	Caviar	NaN	505	534	1880	1987	0.00	75.0
16	18	Sardines	NaN	1425	1484	1856	2928	0.00	50.0
17	19	India Chutney	NaN	16	16	1865	1901	0.10	0.2
18	20	Pickles	NaN	453	472	1852	1987	0.00	10.0
19	21	English Walnuts	NaN	83	86	1851	1948	0.10	0.3

Data: Dish (after Python cleaning)

	id	name	sponsor	event	venue	place	physical_description	occasion	notes	call_number	keywords
0	12463	NaN	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	NaN	1900-2822	NaN
1	12464	NaN	REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, WI];	CARD; ILLU; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY ...	1900-2825	NaN
2	12465	NaN	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;	NaN	MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP A...	1900-2827	NaN
3	12466	NaN	NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;	NaN	MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCEN...	1900-2828	NaN
4	12467	NaN	NORDDEUTSCHER LLOYD BREMEN	DINNER;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	FOLDER; ILLU; COL; 5.5X7.5;	NaN	MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCEN...	1900-2829	NaN

Data: menu (before Python cleaning)

	id	name	sponsor	event	venue	place	physical_description	occasion	notes	call_number	keywords	language	c
0	12463	Unknown	Hotel Eastman	Breakfast	Commercial	Hot Springs, Ar	CARD; 4.75X7.5;	Easter	Unknown	1900-2822	NaN	NaN	1900
1	12464	Unknown	Republican House	Dinner	Commercial	Milwaukee, Wi	CARD; ILLU; COL; 7.0X9.0;	Easter	Wedgewood Blue Card White Embossed Greek Key B...	1900-2825	NaN	NaN	1900
2	12465	Unknown	Norddeutscher Lloyd Bremen	Frühstück/Breakfast	Commercial	Dampfer Kaiser Wilhelm Der Grosse	CARD; ILLU; COL; 5.5X8.0;	Unknown	Menu In German And English Illus, Steamship An...	1900-2827	NaN	NaN	1900
3	12466	Unknown	Norddeutscher Lloyd Bremen	Lunch	Commercial	Dampfer Kaiser Wilhelm Der Grosse	CARD; ILLU; COL; 5.5X8.0;	Unknown	Menu In German And English Illus, Harbor Scene...	1900-2828	NaN	NaN	1900
4	12467	Unknown	Norddeutscher Lloyd Bremen	Dinner	Commercial	Dampfer Kaiser Wilhelm Der Grosse	FOLDER; ILLU; COL; 5.5X7.5;	Unknown	Menu In German And English Illus, Harbor Scene...	1900-2829	NaN	NaN	1900

Data: menu (after Python cleaning)


```
temp = menupage.groupby('menu_id').agg({'page_number':set}).reset_index()
temp['missingPage'] = temp['page_number'].apply(lambda x:missingPage(x))
```

```
temp[temp['missingPage'] == True]
```

	menu_id	page_number	missingPage
5810	21467	{1.0, 3.0}	True
6068	21725	{32.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,...	True
6250	21907	{1.0, 2.0, 4.0, 5.0, 6.0}	True
6605	22265	{2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0,...	True
8267	23936	{1.0, 2.0, 4.0, 5.0}	True
...
18243	33953	{1.0, 4.0}	True
18311	34021	{2.0, 3.0, 4.0, 5.0, 6.0}	True
19186	34897	{1.0, 7.0}	True
19256	34967	{1.0, 2.0, 3.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0,...	True
19761	35472	{1.0, 3.0}	True

1070 rows x 3 columns

Certain pages do not exist in certain menus (such as page 2 in menu_id 21467).

OpenRefine

4.A summary of data changes

5.A summary of findings, problems encountered and lessons learned

- data can be very messy (there might be 100+ phrases that is related to different sort of anniversaries)
- sometimes we need to make a judgement call. It's ambiguous that whether 'anniversary celebration' and 'anniversary reunion' should be merged to the same category
- Sometimes Integrity Constraints are not straightforward to fix (such as missing pages in a menu)

Workflow

Figure 1. In our high-level workflow document, we start with the 4 csv files. Then, we clean the data twice with Python and OpenRefine. We stood up a SQLite db with the schema files described in Phase I. Finally, we import the cleaned data into the database.

