In [1]:

```
import os
import pandas as pd
pd.options.display.float_format = '{:20,.2f}'.format
pd.set_option('display.max_rows', 5000)
pd.set_option('display.max_columns', 5000)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', -1)
```

In [2]:

```
yellow_pages_yelp_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Dat
                          r'/Unprocessed Data/customer_samples/yelp_yel
```

Yellow Pages data

In [54]:

```
yp_fields = ['id', 'name', 'streetAddress', 'city', 'zipCode', 'phone',
```

In [55]:

```
yellow_pages_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'yellow_pages.csv'),
    sep = ',',
    quotechar = '"',
    usecols  = yp_fields,
    dtype = {'zipCode': 'str'}
)[yp_fields]
```

In [56]:

```
yellow_pages_data.columns
```

Out[56]:

```
Index(['id', 'name', 'streetAddress', 'city', 'zipCode', 'phone', 'cat
egories'], dtype='object')
```

In [57]:

```
yellow_pages_data.head()
```

Out[57]:

	id	name	streetAddress	city	zipCode	phone	categories
0	1	full shilling	160 Pearl St	New York	10005	(212) 422-3855	Sandwich Shops;Take Out Restaurants;Hamburgers & Hot Dogs;Health Food Restaurants;Bar & Grills;Bars;Restaurants
1	2	dovetail	103 W 77th St	New York	10024	(212) 362-3800	American Restaurants;French Restaurants;Ice Cream & Frozen Desserts;Fine Dining Restaurants;Bar & Grills;Restaurants
2	3	patron mexican grill	608 9th Ave	New York	10036	(212) 957-9050	Mexican Restaurants;Latin American Restaurants;Bar & Grills;Take Out Restaurants;Restaurants
3	4	ko sushi	1329 2nd Ave	New York	10021	(212) 439-1678	Sushi Bars;Japanese Restaurants;Asian Restaurants;Caterers;Family Style Restaurants;Restaurants
4	5	famous famiglia pizzeria	488 8th Ave	New York	10001	(212) 564-4144	Pizza;Restaurants;Italian Restaurants

In [58]:

```
yellow_pages_data.rename(
    columns = {
        'zipCode': 'postalcode',
        'streetAddress': 'address',
        'categories': 'category'
    },
    inplace=True
)
```

Yelp data

In [66]:

```
yelp_fields = ['id', 'name', 'streetAddress', 'city', 'zipCode', 'telep
```

In [67]:

```
yelp_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'yelp.csv'),
    sep = ',',
    quotechar = '"',
    dtype = {'zipCode': 'str'},
)[yelp_fields]
```

In [71]:

```
yelp_data.columns
```

Out[71]:

```
Index(['id', 'name', 'address', 'city', 'postalcode', 'phone', 'catego
ry'], dtype='object')
```

In [70]:

```
yelp_data.rename(
    columns = {
        'zipCode': 'postalcode',
        'streetAddress': 'address',
        'telephone':'phone'
    },
    inplace=True
)
```

In [72]:

```
yelp_data.columns
```

Out[72]:

```
Index(['id', 'name', 'address', 'city', 'postalcode', 'phone', 'catego
ry'], dtype='object')
```

Labeled data fields

In [73]:

```
lb_fields = ['ltable.id', 'rtable.id', 'gold']
```

In [74]:
```
labeled_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'labeled_data.csv'),
    sep = ',',
    quotechar = '"',
    comment = '#',
    usecols  = lb_fields,
    dtype = {'gold': 'str'}
)[lb_fields]
```

In [34]:
```
labeled_data.head()
```

Out[34]:

	ltable.id	rtable.id	gold
0	8	2437	0
1	74	1175	0
2	109	2339	0
3	154	1568	0
4	215	2445	1

In [75]:
```
labeled_data.rename(
    columns =
    {
        'ltable.id' : 'yelp_id',
        'rtable.id' : 'yellow_pages_id',
        'gold'      : 'duplicate'
    },
    inplace = True
)
```

In [36]:
```
labeled_data.columns
```

Out[36]:
```
Index(['yelp_id', 'yellow_pages_id', 'duplicate'], dtype='object')
```

Cands set

In [79]:
```
cands_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'candset.csv'),
    sep = ',',
    quotechar = '"',
    comment = '#',
    #usecols  = lb_fields,
    #dtype = {'gold': 'str'}
)#[lb_fields]
```

Remove invalid nan entries from all data: yelp, yellow pages, cand sets and labeled data

```
In [ ]: labeled_data
```

```
In [93]: len(labeled_data[labeled_data['duplicate'] == '1'])
```

```
Out[93]: 84
```

```
In [92]: labeled_data.head()
```

Out[92]:

	yelp_id	yellow_pages_id	duplicate
0	8	2437	0
1	74	1175	0
2	109	2339	0
3	154	1568	0
4	215	2445	1

```
In [77]: yellow_pages_to_remove = labeled_data[labeled_data['duplicate'].isnull(
         yelp_to_remove = labeled_data[labeled_data['duplicate'].isnull()]['yelp_
```

```
In [78]: yellow_pages_data = yellow_pages_data[~(yellow_pages_data['id'].isin(ye
         yelp_data = yelp_data[~(yelp_data['id'].isin(yelp_to_remove))]
```

```
In [80]: labeled_data = labeled_data[ ~(labeled_data['duplicate'].isnull())]
```

```
In [40]: labeled_data['duplicate'] = labeled_data['duplicate'].astype('int64')
```

```
In [41]: labeled_data.head()
```

Out[41]:

	yelp_id	yellow_pages_id	duplicate
0	8	2437	0
1	74	1175	0
2	109	2339	0
3	154	1568	0
4	215	2445	1

```
In [ ]: labeled_data['yellow_name'] = labeled_data[labeled_data['yellow_pages]]
```

In [98]:

```
yellow_pages_data[yellow_pages_data['id'] == 1175]
```

Out[98]:

	id	name	address	city	postalcode	phone	category	source
1174	1175	morning star restaurant	879 9th Ave	New York	10019	(212) 246-1593	Health Food Restaurants;Family Style Restaurants;Sandwich Shops;Greek Restaurants;Breakfast Brunch & Lunch Restaurants;Restaurants	yellow_pages

In [99]:

```
yelp_data[yelp_data['id'] == 74]
```

Out[99]:

	id	name	address	city	postalcode	phone	category	source
73	74	monsignor's restaurant	679 5th Ave	New York	11215	(718) 369-2575	Italian;Mexican	yelp

In [82]:

```
customer_all_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Data Int
                     r'/Processed Data/customer_samples/')
```

In [83]:

```
yellow_pages_data['source'] = 'yellow_pages'
yelp_data['source'] = 'yelp'
```

In [100]:

```
yellow_pages_data.head()
```

Out[100]:

	id	name	address	city	postalcode	phone	category	source
0	1	full shilling	160 Pearl St	New York	10005	(212) 422-3855	Sandwich Shops;Take Out Restaurants;Hamburgers & Hot Dogs;Health Food Restaurants;Bar & Grills;Bars;Restaurants	yellow_pages
1	2	dovetail	103 W 77th St	New York	10024	(212) 362-3800	American Restaurants;French Restaurants;Ice Cream & Frozen Desserts;Fine Dining Restaurants;Bar & Grills;Restaurants	yellow_pages
2	3	patron mexican grill	608 9th Ave	New York	10036	(212) 957-9050	Mexican Restaurants;Latin American Restaurants;Bar & Grills;Take Out Restaurants;Restaurants	yellow_pages
3	4	ko sushi	1329 2nd Ave	New York	10021	(212) 439-1678	Sushi Bars;Japanese Restaurants;Asian Restaurants;Caterers;Family Style Restaurants;Restaurants	yellow_pages
4	5	famous famiglia pizzeria	488 8th Ave	New York	10001	(212) 564-4144	Pizza;Restaurants;Italian Restaurants	yellow_pages

```
In [101]: yelp_data.head()
```

Out[101]:

	id	name	address	city	postalcode	phone	category	source
0	1	sunshine co.	780 Washington Ave	New York	11238	(347) 750-5275	American (New);Cocktail Bars	yelp
1	2	adella	410 W 43rd St	New York	10036	(212) 273-0737	Tapas Bars;American (New);Wine Bars	yelp
2	3	rex	864 10th Ave	New York	10019	(929) 900-5784	Coffee & Tea;Sandwiches	yelp
3	4	bistro petit	170 S 3rd St	New York	11211	(718) 782-2582	French	yelp
4	5	jora restaurant & bar	47-46 11th St	New York	11101	(718) 392-2033	Peruvian	yelp

```
In [84]: yellow_yelp_all2 = pd.concat([yellow_pages_data, yelp_data])
         #yellow_yelp_all2.rename(columns={'record_id': 'Id'}, inplace=True)
```

```
In [85]: yellow_yelp_all2.columns
```

```
Out[85]: Index(['id', 'name', 'address', 'city', 'postalcode', 'phone', 'catego
         ry', 'source'], dtype='object')
```

```
In [86]: yellow_yelp_all2.to_csv(customer_all_path + 'yellow_yelp_all2.csv', sep
```

```
In [49]: labeled_data.to_csv(customer_all_path + 'yellow_yelp_label2.csv', sep='
```