Load Data

In [134]:
```
import os
import pandas as pd
import numpy as np
import logging
```

Setup

In [135]:
```
folder = r'/home/ubuntu/path_to_folder
input_file = 'yellow_yelp_all_pypostal2.csv'
output_file = 'yellow_yelp_all_pypostal2_output1.csv'
settings_file = 'yellow_yelp_all_pypostal2_learned_settings1'
training_file = 'yellow_yelp_all_pypostal2_training1.json'
```

In [136]:
```
fp = os.path.join(folder, input_file)
```

In [137]:
```
matches_file = os.path.join(folder, 'yellow_yelp_label2.csv')
```

In [138]:
```
log_level = logging.INFO
log_level = logging.DEBUG
logging.getLogger().setLevel(log_level)
```

Dataframe view

In [139]:
```
input_df = pd.read_csv(fp, sep=',', quotechar='"', dtype={'postalcode':
```

In [140]:
```
def get_clean_postalcode(x):

    if x is not None:
        subparts = str(x).split('.')
        return subparts[0]
    else:
        return None
```

In [141]:
```
input_df['postalcode'] = input_df['postalcode'].apply(lambda x: get_cle
```

In [42]:

```
input_df.head()
```

Out[42]:

	id	source	name	category	phone	city	postalcode	address	street
0	1	yellow_pages	full shilling	Sandwich Shops;Take Out Restaurants;Hamburgers...	(212) 422-3855	New York	10005	160 Pearl St	pearl st
1	2	yellow_pages	dovetail	American Restaurants;French Restaurants;Ice Cr...	(212) 362-3800	New York	10024	103 W 77th St	w 77th st
2	3	yellow_pages	patron mexican grill	Mexican Restaurants;Latin American Restaurants...	(212) 957-9050	New York	10036	608 9th Ave	9th ave
3	4	yellow_pages	ko sushi	Sushi Bars;Japanese Restaurants;Asian Restaura...	(212) 439-1678	New York	10021	1329 2nd Ave	2nd ave
4	5	yellow_pages	famous famiglia pizzeria	Pizza;Restaurants;Italian Restaurants	(212) 564-4144	New York	10001	488 8th Ave	8th ave

In [143]:

```
categories = list(input_df['category'].unique())
categories = [x for x in categories if str(x) != 'nan']
```

In [144]:

```
#category_corpus = input_df[['name', 'category']].drop_duplicates().to_
category_corpus = input_df.drop_duplicates().to_dict(orient='records')
```

In [145]:

```
category_corpus[0]
```

Out[145]:

```
{'id': 1,
 'source': 'yellow_pages',
 'name': 'full shilling',
 'category': 'Sandwich Shops;Take Out Restaurants;Hamburgers & Hot Dog
s;Health Food Restaurants;Bar & Grills;Bars;Restaurants',
 'phone': '(212) 422-3855',
 'city': 'New York',
 'postalcode': '10005',
 'address': '160 Pearl St',
 'street': 'pearl st',
 'house_number': '160',
 'house': nan}
```

In [13]:

```
type(category_corpus)
```

Out[13]:

```
list
```

Dedupe

Import modules

In [146]:
```
import os
import csv
import re
import logging
import optparse

import dedupe
from unidecode import unidecode

from itertools import product
```

In [147]:
```
def pre_process(val):
    """
    Do a little bit of data cleaning with the help of Unidecode and Rege
    Things like casing, extra spaces, quotes and new lines can be ignore
    """
    try:
        val = re.sub('  +', ' ', val)
        val = re.sub('\n', ' ', val)
        val = val.strip().strip('"').strip("'").lower().strip()
        # If data is missing, indicate that by setting the value to `Non
        if not val:
            val = None
    except Exception as e:
        print(e)
    return val
```

In [148]:
```
def get_clean_data_dict(file_path):
    data_d = {}
    with open(fp) as f:
        reader = csv.DictReader(f)
        for row in reader:
            clean_row = [(k, pre_process(v)) for (k, v) in row.items()]
            row_id = int(row['id'])
            data_d[row_id] = dict(clean_row)

    return data_d
```

Get Data in needed format

In [149]:
```
data_dict = get_clean_data_dict(fp)
```

In [151]:

```
data_dict[1]
```

Out[151]:

```
{'id': '1',
 'source': 'yelp',
 'name': 'sunshine co.',
 'category': 'american (new);cocktail bars',
 'phone': '(347) 750-5275',
 'city': 'new york',
 'postalcode': '11238',
 'address': '780 washington avenue',
 'street': 'washington avenue',
 'house_number': '780',
 'house': None}
```

Define the Fields for dedupe

In [152]:

```
fields = [
    {'field' : 'name', 'type': 'Name'},
    {'field' : 'category',
     'type': 'FuzzyCategorical',
     'categories': categories,
     'corpus': category_corpus,
     'has missing' : True},
    {'field' : 'name', 'type': 'String'},
    {'field': 'postalcode', 'variable name': 'postalcode', 'type': 'Exa
    {'field' : 'address', 'type': 'Address'},
    {'field' : 'city', 'type': 'ShortString'},
    {'field' : 'phone', 'type': 'String'},
    {'field' : 'street', 'type': 'String', 'has missing' : True},
    {'field' : 'house_number', 'type': 'Exists', 'has missing' : True},
    {'field' : 'house', 'type': 'String', 'has missing' : True},

]
```

Instantiate Dedupe

In [153]:

```
deduper = dedupe.Dedupe(fields)
```

```
In [54]: deduper.prepare_training(data_dict, blocked_proportion=0.7)
```

```
INFO:dedupe.canopy_index:Removing stop word  s
INFO:dedupe.canopy_index:Removing stop word 9
INFO:dedupe.canopy_index:Removing stop word st
INFO:dedupe.canopy_index:Removing stop word  d
INFO:dedupe.canopy_index:Removing stop word 10
INFO:dedupe.canopy_index:Removing stop word d
INFO:dedupe.canopy_index:Removing stop word er
INFO:dedupe.canopy_index:Removing stop word rd
INFO:dedupe.canopy_index:Removing stop word  m
INFO:dedupe.canopy_index:Removing stop word 30
INFO:dedupe.canopy_index:Removing stop word ar
INFO:dedupe.canopy_index:Removing stop word h
INFO:dedupe.canopy_index:Removing stop word ma
INFO:dedupe.canopy_index:Removing stop word re
INFO:dedupe.canopy_index:Removing stop word t
INFO:dedupe.canopy_index:Removing stop word th
INFO:dedupe.canopy_index:Removing stop word  a
INFO:dedupe.canopy_index:Removing stop word 20
INFO:dedupe.canopy_index:Removing stop word av
INFO:dedupe.canopy_index:Removing stop word
```

In [55]:

```
dedupe.consoleLabel(deduper)
```

```
name : the ultimate barbeque
category : barbecue restaurants;restaurants
postalcode : 95123
address : 446 colony knoll dr
city : san jose
phone : (408) 227-7427
street : colony knoll dr
house_number : 446
house : None

name : the ultimate barbeque
category : barbecue restaurants;restaurants
postalcode : 95113
address : None
city : san jose
phone : (408) 227-7427
street : None
house_number : None
house : None

0/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished

y

name : 5411 empanadas
category : argentine;food trucks
postalcode : 60602
address : None
city : chicago
phone : (773) 755-5411
street : None
house_number : None
house : None

name : 5411 empanadas
category : argentine
postalcode : 60657
address : 2850 n clark st
city : chicago
phone : (773) 755-5411
street : n clark st
house_number : 2850
house : None

1/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : cafe rosalena
```

```
category : breakfast & brunch;american (traditional)
postalcode : 95126
address : 1077 the alameda
city : san jose
phone : (408) 287-2400
street : the alameda
house_number : 1077
house : None

name : cafe rosalena
category : coffee shops;restaurants
postalcode : 95126
address : 1077 the alameda
city : san jose
phone : (408) 287-2400
street : the alameda
house_number : 1077
house : None

2/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
name : michi
category : sushi bars;japanese;asian fusion
postalcode : 95110
address : at m8trix 1887 matrix blvd
city : san jose
phone : (408) 378-8000
street : matrix blvd
house_number : 1887
house : at m8trix

name : sushi boat
category : japanese;sushi bars
postalcode : 95110
address : mineta airport
city : san jose
phone : None
street : None
house_number : None
house : mineta airport

3/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), PartialPredicate: (wholeFieldPredicate, name, Surname))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
name : las sabrosas de guanajuato
category : mexican
postalcode : 78216
address : 6825 san pedro
city : san antonio
phone : (210) 785-9211
street : san pedro
house_number : 6825
house : None

name : las sabrosas
category : restaurants
postalcode : 78216
address : 6825 san pedro avenue
city : san antonio
phone : (210) 785-9211
street : san pedro avenue
house_number : 6825
house : None

3/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : taco stop
category : mexican
postalcode : 75207
address : 1900 irving boulevard
city : dallas
phone : (972) 971-4859
street : irving boulevard
house_number : 1900
house : None

name : taco stop
category : mexican restaurants;take out restaurants;latin american res
taurants;restaurants
postalcode : 75207
address : 1900 irving boulevard
city : dallas
phone : (972) 971-4859
street : irving boulevard
house_number : 1900
house : None

4/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, na
me, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
name : myung in dumplings
category : korean
postalcode : 92111
address : 4344 convoy st
city : san diego
phone : (858) 565-2688
street : convoy st
house_number : 4344
house : None

name : myung in dumplings
category : chinese restaurants;restaurants
postalcode : 92111
address : 4344 convoy st
city : san diego
phone : (858) 565-2688
street : convoy st
house_number : 4344
house : None

5/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), TfidfNGramCanopyPredicate: (0.6, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
name : golden wok
category : chinese
postalcode : 77018
address : 3450 ella boulevard
city : houston
phone : (713) 957-1551
street : ella boulevard
house_number : 3450
house : None

name : golden wok
category : chinese restaurants;japanese restaurants;sushi bars;asian r
estaurants;take out restaurants;restaurants
postalcode : 77018
address : 3450 ella boulevard
city : houston
```

```
phone : (713) 957-1551
street : ella boulevard
house_number : 3450
house : None

6/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : kyoto sushi
category : sushi bars
postalcode : 92110
address : 3166 midway dr ste 108
city : san diego
phone : (619) 223-7798
street : midway dr
house_number : 3166
house : None

name : kyoto sushi
category : sushi bars;japanese restaurants;asian restaurants;seafood r
estaurants;restaurants
postalcode : 92110
address : 3166 midway dr ste 108
city : san diego
phone : (619) 223-7798
street : midway dr
house_number : 3166
house : None

7/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, Surname), TfidfNGramCanopyPredicate: (0.6, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, na
me, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, na
me, Surname), SimplePredicate: (firstTokenPredicate, address))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), TfidfTextCanopyPredicate: (0.8, phone))
name : amada
category : tapas bars;breakfast & brunch
postalcode : 19106
address : 217 chestnut st
city : philadelphia
phone : (215) 625-2450
street : chestnut st
house_number : 217
house : None
```

```
name : amada
category : american restaurants;bar & grills;family style restaurant
s;take out restaurants;caterers;restaurants
postalcode : 19106
address : 217 chestnut st # 219
city : philadelphia
phone : (215) 625-2450
street : chestnut st
house_number : 217
house : None

8/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), TfidfNGramCanopyPredicate: (0.6, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (firstTokenPredicate, address))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, Su
rname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
name : sun asian kitchen
category : chinese;sushi bars;asian fusion
postalcode : 85042
address : 2070 e baseline rd ste d112
city : phoenix
phone : (602) 268-7708
street : e baseline rd
house_number : 2070
house : None

name : sun asian kitchen
category : asian restaurants;take out restaurants;japanese restaurant
s;chinese restaurants;restaurants
postalcode : 85042
address : 2070 e baseline rd ste 112
city : phoenix
phone : (602) 268-7708
street : e baseline rd
house_number : 2070
house : None

9/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGra
mCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.2, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
name : taco riendo
category : mexican
postalcode : 78212
address : 5506 san pedro avenue
city : san antonio
phone : (210) 824-2463
street : san pedro avenue
house_number : 5506
house : None

name : taco riendo
category : mexican restaurants;restaurants
postalcode : 78212
address : 5506 san pedro avenue
city : san antonio
phone : (210) 824-2463
street : san pedro avenue
house_number : 5506
house : None

10/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGr
amCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.2, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (sortedAcronym, name, Surnam
e), SimplePredicate: (commonThreeTokens, name))
name : corner table
category : american (new);gluten-free;vegetarian
postalcode : 77098
address : 2736 virginia st
city : houston
phone : (713) 568-9196
street : virginia st
house_number : 2736
house : None

name : corner table
category : american restaurants;take out restaurants;caterers;contine
```

```
ntal restaurants;restaurants
postalcode : 77098
address : 2736 virginia st
city : houston
phone : (713) 568-9196
street : virginia st
house_number : 2736
house : None

11/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGra
mCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.2, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, Su
rname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
name : yellowfish sushi
category : sushi bars;japanese;mexican
postalcode : 78240
address : 9102 wurzbach road
city : san antonio
phone : (210) 614-3474
street : wurzbach road
house_number : 9102
house : None

name : yellowfish sushi
category : sushi bars;japanese restaurants;asian restaurants;take out
restaurants;caterers;restaurants
postalcode : 78240
address : 9102 wurzbach road
city : san antonio
phone : (210) 614-3474
street : wurzbach road
house_number : 9102
house : None

12/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGr
amCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.2, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
```

```
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, S
urname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, Surname), TfidfNGramCanopyPredicate: (0.6, address))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), TfidfTextCanopyPredicate: (0.8, phone))
name : frankford hall
category : pubs;german
postalcode : 19125
address : 1210 frankford avenue
city : philadelphia
phone : (215) 634-3338
street : frankford avenue
house_number : 1210
house : None

name : frankford hall
category : restaurants
postalcode : 19125
address : 1210 frankford avenue
city : philadelphia
phone : (215) 634-3338
street : frankford avenue
house_number : 1210
house : None

13/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
name : bap
category : korean
postalcode : 19147
address : 1224 south st
city : philadelphia
phone : (215) 735-0553
street : south st
house_number : 1224
house : None

name : bap
category : korean restaurants;restaurants
postalcode : 19147
address : 1224 south st
city : philadelphia
phone : (215) 735-0553
street : south st
house_number : 1224
house : None

14/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(PartialPredicate: (doubleMetaphone, name, Surnam
e), SimplePredicate: (sameSevenCharStartPredicate, phone))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), Tf
idfNGramCanopyPredicate: (0.8, name))
name : pho van
category : vietnamese
postalcode : 92105
address : 4233 el cajon boulevard
city : san diego
phone : (619) 281-9420
street : el cajon boulevard
house_number : 4233
house : None

name : pho van restaurant
category : vietnamese restaurants;take out restaurants;asian restauran
ts;family style restaurants;restaurants
postalcode : 92105
address : 4233 el cajon boulevard
city : san diego
phone : (619) 281-9420
street : el cajon boulevard
house_number : 4233
house : None

15/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : rotisseur
category : american (new);vietnamese;sandwiches
postalcode : 19103
address : 102 s 21st st
city : philadelphia
phone : (215) 496-9494
street : s 21st st
house_number : 102
house : None

name : rotisseur
category : vietnamese restaurants;asian restaurants;restaurants
postalcode : 19103
address : 102 s 21st st
city : philadelphia
phone : (215) 496-9494
street : s 21st st
house_number : 102
```

```
house : None

16/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(PartialPredicate: (metaphoneToken, name, Surnam
e), TfidfTextCanopyPredicate: (0.4, address))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), Tf
idfNGramCanopyPredicate: (0.8, name))
name : underbelly
category : pubs;ramen
postalcode : 92101
address : 750 w fir st ste 101
city : san diego
phone : (619) 269-4626
street : w fir st
house_number : 750
house : None

name : underbelly
category : restaurants
postalcode : 92101
address : 750 w fir st
city : san diego
phone : (619) 269-4626
street : w fir st
house_number : 750
house : None

17/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(PartialPredicate: (metaphoneToken, name, Surnam
e), TfidfTextCanopyPredicate: (0.4, address))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, name), PartialPr
edicate: (firstIntegerPredicate, address, StreetName))
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGra
mCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfTextCanopyPredicate: (0.8, phone))
name : hanna garden
category : mediterranean
postalcode : 92121
address : 7160 miramar road
```

```
city : san diego
phone : (858) 537-0888
street : miramar road
house_number : 7160
house : None

name : hanna gardens
category : take out restaurants;caterers;restaurants
postalcode : 92121
address : 7160 miramar road
city : san diego
phone : (858) 537-0888
street : miramar road
house_number : 7160
house : None

18/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : top nosh café
category : breakfast & brunch;coffee & tea;caterers
postalcode : 95125
address : 1167 lincoln avenue
city : san jose
phone : (408) 995-6674
street : lincoln avenue
house_number : 1167
house : None

name : top nosh
category : coffee shops;restaurants
postalcode : 95125
address : 1167 lincoln avenue
city : san jose
phone : (408) 995-6674
street : lincoln avenue
house_number : 1167
house : None

19/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (firstTokenPredicate, name), Tf
idfTextCanopyPredicate: (0.8, address))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, Su
rname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, Su
rname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, name), PartialPr
edicate: (firstIntegerPredicate, address, StreetName))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), TfidfNGramCanopyPredicate: (0.6, address))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), Tf
idfNGramCanopyPredicate: (0.8, name))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : frederick’s
category : french;tapas/small plates
postalcode : 78209
address : 7701 broadway suite 20
city : san antonio
phone : (210) 828-9050
street : broadway
house_number : 7701
house : None

name : frederick's
category : french restaurants;family style restaurants;restaurants
postalcode : 78209
address : 7701 broadway st
city : san antonio
phone : (210) 828-9050
street : broadway st
house_number : 7701
house : None

20/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, n
ame), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, Su
rname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, Su
rname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, name), PartialPr
edicate: (firstIntegerPredicate, address, StreetName))
name : ly heng
category : asian fusion
postalcode : 92105
address : 4451 university avenue
city : san diego
phone : (619) 280-8688
street : university avenue
house_number : 4451
house : None

name : ly heng asian restaurant
category : asian restaurants;restaurants
postalcode : 92105
address : 4451 university avenue
city : san diego
phone : (619) 280-8688
```

```
street : university avenue
house_number : 4451
house : None

21/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : villa di roma
category : italian
postalcode : 19147
address : 936 s 9th st
city : philadelphia
phone : (215) 592-1295
street : s 9th st
house_number : 936
house : None

name : villa di roma restaurant
category : breakfast brunch & lunch restaurants;italian restaurants;ba
r & grills;family style restaurants;caterers;mediterranean restaurant
s;restaurants
postalcode : 19147
address : 936 s 9th st
city : philadelphia
phone : (215) 592-1295
street : s 9th st
house_number : 936
house : None

22/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), PartialP
redicate: (firstTokenPredicate, name, Surname))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), TfidfT
extCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, name), PartialPr
edicate: (firstIntegerPredicate, address, StreetName))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), Tf
idfNGramCanopyPredicate: (0.8, name))
name : cibodivino marketplace
category : italian;cafes
postalcode : 75208
address : 1868 sylvan ave ste d100
city : dallas
phone : (214) 653-2426
street : sylvan ave
house_number : 1868
house : None
```

```
name : cibodivino
category : italian restaurants;mediterranean restaurants;restaurants
postalcode : 75208
address : 1868 sylvan ave # d100
city : dallas
phone : (214) 653-2426
street : sylvan ave # d100
house_number : 1868
house : None

23/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), PartialP
redicate: (firstTokenPredicate, name, Surname))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), TfidfT
extCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addres
s, StreetName), SimplePredicate: (sameSevenCharStartPredicate, name))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), Tf
idfNGramCanopyPredicate: (0.8, name))
name : sushiya
category : sushi bars;japanese
postalcode : 75202
address : 1306 elm st
city : dallas
phone : (214) 744-9600
street : elm st
house_number : 1306
house : None

name : sushiya japanese restaurant
category : japanese restaurants;asian restaurants;sushi bars;bar & gri
lls;family style restaurants;take out restaurants;restaurants
postalcode : 75202
address : 1306 elm st
city : dallas
phone : (214) 744-9600
street : elm st
house_number : 1306
house : None

24/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), Partial
Predicate: (firstTokenPredicate, name, Surname))
```

```
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Tfidf
TextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addre
ss, StreetName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, na
me, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (oneGramFingerprint, phone), T
fidfNGramCanopyPredicate: (0.8, name))
INFO:dedupe.training:(TfidfNGramCanopyPredicate: (0.4, phone), TfidfN
GramCanopyPredicate: (0.6, name))
name : shogun
category : japanese;sushi bars
postalcode : 85032
address : 12615 n tatum blvd
city : phoenix
phone : (602) 953-3264
street : n tatum blvd
house_number : 12615
house : None

name : shogun restaurant
category : sushi bars;japanese restaurants;asian restaurants;caterer
s;restaurants
postalcode : 85032
address : 12615 n tatum blvd
city : phoenix
phone : (602) 953-3264
street : n tatum blvd
house_number : 12615
house : None

25/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameFiveCharStartPredicate, n
ame), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, S
urname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, S
urname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addre
ss, StreetName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
name : dairy queen
category : fast food restaurants;ice cream & frozen desserts;hamburge
rs & hot dogs;restaurants;dessert restaurants
postalcode : 85033
address : desert sky mall
city : phoenix
phone : (623) 849-2440
street : None
house_number : None
house : desert sky mall
```

```
name : dairy queen
category : dessert restaurants;fast food restaurants;restaurants;ice
cream & frozen desserts;hamburgers & hot dogs
postalcode : 85029
address : 12456 n 28th dr
city : phoenix
phone : (602) 942-1496
street : n 28th dr
house_number : 12456
house : None

26/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
n
```

```
name : z’tejas southwestern grill
category : tex-mex;mexican;cafes
postalcode : 85014
address : 1525 e bethany home rd
city : phoenix
phone : (602) 680-2806
street : e bethany home rd
house_number : 1525
house : None

name : z'tejas
category : mexican restaurants;latin american restaurants;restaurants
postalcode : 85014
address : 1525 e bethany home rd
city : phoenix
phone : (602) 680-2806
street : e bethany home rd
house_number : 1525
house : None

26/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
name : mejico cocina mejicana
category : mexican
postalcode : 85016
address : 2333 e osborn rd
city : phoenix
phone : (602) 956-4420
street : e osborn rd
house_number : 2333
house : None

name : mijico
category : restaurants
postalcode : 85016
address : 2333 e osborn rd
```

```
city : phoenix
phone : (602) 956-4420
street : e osborn rd
house_number : 2333
house : None

27/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameFiveCharStartPredicate, na
me), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, Su
rname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, Su
rname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addres
s, StreetName), SimplePredicate: (sameSevenCharStartPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), SimplePredicate: (wholeFieldPredicate, phone))
name : gossip grill
category : american (new);gay bars;cocktail bars
postalcode : 92103
address : 1220 university avenue
city : san diego
phone : (619) 260-8023
street : university avenue
house_number : 1220
house : None

name : gossip girl
category : american restaurants;bars;taverns;restaurants
postalcode : 92103
address : 1220 university avenue
city : san diego
phone : (619) 260-8023
street : university avenue
house_number : 1220
house : None

28/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(TfidfTextCanopyPredicate: (0.8, phone), TfidfTe
xtCanopyPredicate: (0.8, street))
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGr
amCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, S
urname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, S
urname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addre
ss, StreetName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), SimplePredicate: (wholeFieldPredicate, phone))
name : rouge
category : breakfast & brunch;burgers;lounges
postalcode : 19103
address : 205 s 18th st
city : philadelphia
phone : (215) 732-6622
street : s 18th st
house_number : 205
house : None

name : rouge ninety eight inc
category : american restaurants;bars;french restaurants;hamburgers &
hot dogs;bar & grills;breakfast brunch & lunch restaurants;asian rest
aurants;restaurants
postalcode : 19103
address : 205 s 18th st
city : philadelphia
phone : (215) 732-6622
street : s 18th st
house_number : 205
house : None

29/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
name : mugshot diner
category : diners;american (new)
postalcode : 19125
address : 2424 e york st
city : philadelphia
phone : (215) 426-2424
street : e york st
house_number : 2424
house : None

name : mugshots diner philadelphia
category : american restaurants;coffee shops;caterers;restaurants
postalcode : 19125
address : 2424 e york st
city : philadelphia
phone : (215) 426-2424
street : e york st
house_number : 2424
house : None

30/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(TfidfTextCanopyPredicate: (0.8, phone), TfidfTe
xtCanopyPredicate: (0.8, street))
INFO:dedupe.training:(SimplePredicate: (fingerprint, phone), TfidfNGr
amCanopyPredicate: (0.6, name))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, S
urname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, S
urname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonIntegerPredicate, addr
ess, StreetName), SimplePredicate: (sameFiveCharStartPredicate, nam
e))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), SimplePredicate: (wholeFieldPredicate, phone))
name : saaghi
category : persian/iranian
postalcode : 95128
address : 1392 s bascom ave
city : san jose
phone : (408) 998-0122
street : s bascom ave
house_number : 1392
house : None

name : saaghi resturant
category : restaurants
postalcode : 95128
address : 1392 s bascome ave
city : san jose
phone : (408) 998-0122
street : s bascome ave
house_number : 1392
house : None

31/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameFiveCharStartPredicate, n
ame), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (firstTokenPredicate, name, S
urname), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, S
urname), TfidfNGramCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (firstIntegerPredicate, addre
ss, StreetName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGr
amCanopyPredicate: (0.6, address))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), SimplePredicate: (wholeFieldPredicate, phone))
INFO:dedupe.training:(TfidfNGramCanopyPredicate: (0.2, phone), TfidfN
GramCanopyPredicate: (0.6, address))
```

```
name : bingsoo pong
category : desserts;korean
postalcode : 75229
address : 2240 royal ln
city : dallas
phone : None
street : royal ln
house_number : 2240
house : None

name : sck restaurant group dba pho que huong
category : restaurants
postalcode : 75229
address : 2240 royal ln
city : dallas
phone : (972) 290-0743
street : royal ln
house_number : 2240
house : None

32/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
u
```

```
name : eggslut
category : breakfast & brunch
postalcode : 90013
address : grand central market 317 s broadway
city : los angeles
phone : None
street : s broadway
house_number : 317
house : grand central market

name : tacos tumbras a tomas
category : mexican
postalcode : 90013
address : grand central market, space a-5 317 s broadway
city : los angeles
phone : (213) 620-1071
street : s broadway
house_number : 317
house : grand central market space

32/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
n
```

```
name : the mediterranean grill
category : mediterranean
postalcode : 77010
address : 1200 mckinney st ste 36
city : houston
phone : None
```

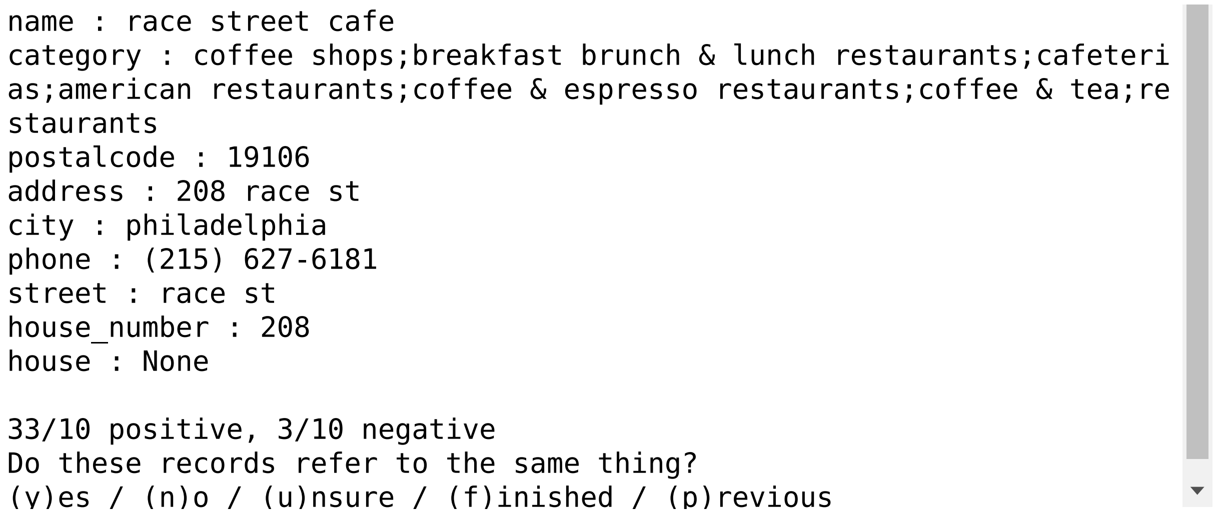
```
street : mckinney st
house_number : 1200
house : None

name : longhorn cafe
category : coffee shops;restaurants
postalcode : 77010
address : 1200 mckinney st ste 439
city : houston
phone : (713) 751-0021
street : mckinney st
house_number : 1200
house : None

32/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
n
```

```
name : happy guy chinese cuisine
category : chinese
postalcode : 78251
address : 8373 culebra road
city : san antonio
phone : (210) 520-1728
street : culebra road
house_number : 8373
house : None

name : happy guy chinese cuisine
category : chinese restaurants;asian restaurants;take out restaurants;
restaurants
postalcode : 78251
address : 8373 culebra rd ste 205
city : san antonio
phone : (210) 520-1728
street : culebra rd
house_number : 8373
house : None

32/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

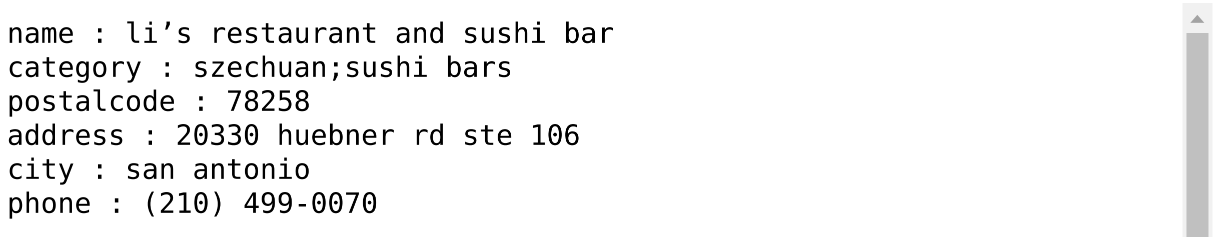
```
y
```

```
name : race street cafe
category : american (new);cafes
postalcode : 19106
address : 208 race st
city : philadelphia
phone : (215) 627-6181
street : race st
house_number : 208
house : None
```

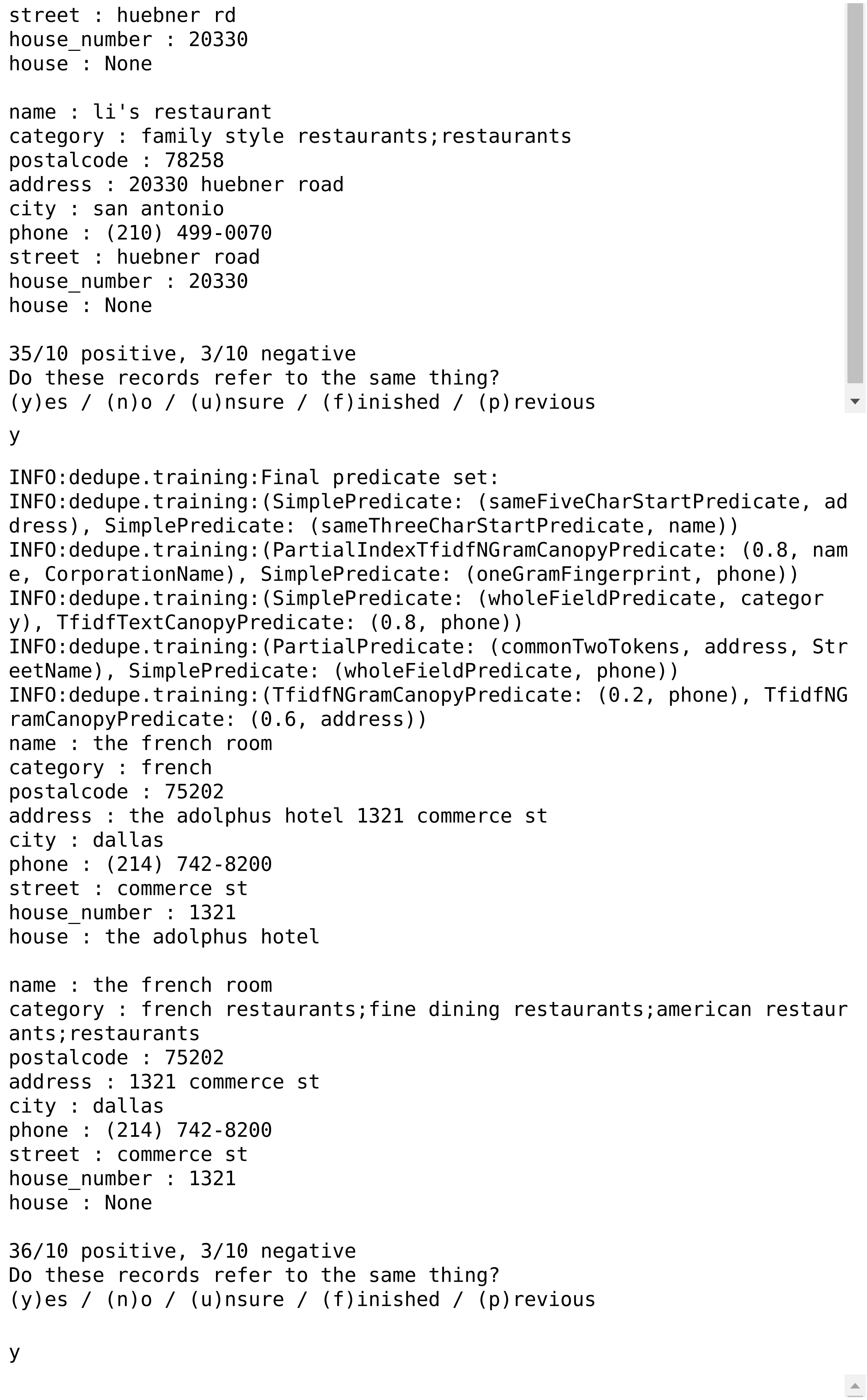


```
name : race street cafe
category : coffee shops;breakfast brunch & lunch restaurants;cafeteri
as;american restaurants;coffee & espresso restaurants;coffee & tea;re
staurants
postalcode : 19106
address : 208 race st
city : philadelphia
phone : (215) 627-6181
street : race st
house_number : 208
house : None

33/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameFiveCharStartPredicate, na
me), SimplePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), SimplePredicate: (wholeFieldPredicate, phone))
INFO:dedupe.training:(TfidfNGramCanopyPredicate: (0.2, phone), TfidfNG
ramCanopyPredicate: (0.6, address))
name : saad’s halal restaurant
category : halal;middle eastern
postalcode : 19139
address : 4500 walnut st
city : philadelphia
phone : (215) 222-7223
street : walnut st
house_number : 4500
house : None

name : saad's halal restaurant
category : american restaurants;sandwich shops;vegetarian restaurants;
health food restaurants;middle eastern restaurants;restaurants
postalcode : 19139
address : 4500 walnut st
city : philadelphia
phone : (215) 222-7223
street : walnut st
house_number : 4500
house : None

34/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious


y
```



```
name : li’s restaurant and sushi bar
category : szechuan;sushi bars
postalcode : 78258
address : 20330 huebner rd ste 106
city : san antonio
phone : (210) 499-0070
```



```
street : huebner rd
house_number : 20330
house : None

name : li's restaurant
category : family style restaurants;restaurants
postalcode : 78258
address : 20330 huebner road
city : san antonio
phone : (210) 499-0070
street : huebner road
house_number : 20330
house : None

35/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameFiveCharStartPredicate, ad
dress), SimplePredicate: (sameThreeCharStartPredicate, name))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), SimplePredicate: (wholeFieldPredicate, phone))
INFO:dedupe.training:(TfidfNGramCanopyPredicate: (0.2, phone), TfidfNG
ramCanopyPredicate: (0.6, address))
name : the french room
category : french
postalcode : 75202
address : the adolphus hotel 1321 commerce st
city : dallas
phone : (214) 742-8200
street : commerce st
house_number : 1321
house : the adolphus hotel

name : the french room
category : french restaurants;fine dining restaurants;american restaur
ants;restaurants
postalcode : 75202
address : 1321 commerce st
city : dallas
phone : (214) 742-8200
street : commerce st
house_number : 1321
house : None

36/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, phone), Simp
lePredicate: (firstTokenPredicate, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.8, na
me, CorporationName), SimplePredicate: (oneGramFingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : 45 mint vietnamese bistro
category : vietnamese
postalcode : 75254
address : 14902 preston rd ste 512b
city : dallas
phone : (972) 387-4545
street : preston rd
house_number : 14902
house : None

name : 45 mint vietnamese bistro
category : restaurants
postalcode : 75254
address : 14092 preston road
city : dallas
phone : (972) 387-4545
street : preston road
house_number : 14092
house : None

37/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, p
hone), SimplePredicate: (twoGramFingerprint, phone))
name : pier 22 seafood
category : family style restaurants;restaurants
postalcode : 19146
address : 2204 washington avenue
city : philadelphia
phone : (215) 732-3107
street : washington avenue
house_number : 2204
house : None

name : pier 22 seafood
category : seafood restaurants;restaurants
postalcode : 19146
address : 2206 washington avenue
city : philadelphia
phone : (215) 732-2037
street : washington avenue
house_number : 2206
house : None

38/10 positive, 3/10 negative
```

```
Do these records refer to the same thing?
```

```
y
```

```
name : five guys burgers & fries
category : fast food restaurants;restaurants;hamburgers & hot dogs
postalcode : 95123
address : 5550 cottle road
city : san jose
phone : (408) 366-1006
street : cottle road
house_number : 5550
house : None

name : five guys burgers & fries
category : hamburgers & hot dogs;restaurants;fast food restaurants
postalcode : 95123
address : 5660 cottle road
city : san jose
phone : (408) 363-8200
street : cottle road
house_number : 5660
house : None

39/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
n
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, p
hone), SimplePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
name : rockfish seafood grill
category : seafood
postalcode : 75206
address : 5331 e mockingbird ln ste 160
city : dallas
phone : (214) 823-8444
street : e mockingbird ln
house_number : 5331
house : None

name : rockfish seafood grill
category : seafood restaurants;restaurants
postalcode : 75206
address : 5531 e mockingbird ln
city : dallas
phone : (214) 826-6342
street : e mockingbird ln
house_number : 5531
house : None

39/10 positive, 4/10 negative
```

```
Do these records refer to the same thing?
y

name : tiffin
category : middle eastern restaurants;indian restaurants;caterers;asi
an restaurants;restaurants
postalcode : 19119
address : 7105 emlen st
city : philadelphia
phone : (215) 242-3656
street : emlen st
house_number : 7105
house : None

name : tiffin
category : indian restaurants;restaurant delivery service;asian resta
urants;caterers;take out restaurants;middle eastern restaurants;resta
urants
postalcode : 19123
address : 710 w girard ave
city : philadelphia
phone : (215) 922-1297
street : w girard ave
house_number : 710
house : None

40/10 positive, 4/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, p
hone), SimplePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGra
mCanopyPredicate: (0.6, address))
name : carl's jr.
category : restaurants;fast food restaurants;hamburgers & hot dogs;ame
rican restaurants;mexican restaurants
postalcode : 95148
address : 2802 s white rd
city : san jose
phone : (408) 238-3821
street : s white rd
house_number : 2802
house : None

name : carl's jr.
category : mexican restaurants;restaurants;fast food restaurants;ameri
can restaurants;hamburgers & hot dogs
postalcode : 95116
address : 2 n jackson ave
city : san jose
phone : (408) 258-3680
```

```
street : n jackson ave
house_number : 2
house : None

40/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : jc’s famous bbq
category : barbeque
postalcode : 95129
address : 1080 saratoga ave ste 8
city : san jose
phone : (408) 246-2146
street : saratoga ave
house_number : 1080
house : None

name : jc's famous barbeque
category : barbecue restaurants;take out restaurants;caterers;restaura
nts
postalcode : 95129
address : 1080 saratoga avenue number 8
city : san jose
phone : (408) 246-2147
street : saratoga avenue
house_number : 1080
house : None

40/10 positive, 6/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : nola’s creole 2 geaux
category : street vendors;food delivery services;cajun/creole
postalcode : 77002
address : None
city : houston
phone : None
street : None
house_number : None
house : None

name : bcr2
category : restaurants
postalcode : 95123
address : 925 blossom hill road
city : san jose
phone : (408) 225-2016
street : blossom hill road
house_number : 925
house : None

41/10 positive, 6/10 negative
```

```
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), PartialI
ndexLevenshteinCanopyPredicate: (2, address, StreetName))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGra
mCanopyPredicate: (0.6, address))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : mama margies mexican restaurant
category : mexican;breakfast & brunch
postalcode : 78230
address : 9950 interstate 10 frontage road
city : san antonio
phone : (210) 561-0400
street : interstate 10 frontage road
house_number : 9950
house : None

name : mama margie's
category : mexican restaurants;breakfast brunch & lunch restaurants;fa
mily style restaurants;restaurants
postalcode : 78230
address : 9950 w interstate 10
city : san antonio
phone : (210) 561-0400
street : w interstate 10
house_number : 9950
house : None

41/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious



y

name : chili's grill & bar
category : bars;take out restaurants;american restaurants;bar & grill
s;restaurants
postalcode : 77032
address : ca1 services george bush- houston intercontinental airpo 31
00 terminal way
city : houston
phone : (281) 233-3191
street : terminal way
house_number : 3100
house : ca1 services george bush houston intercontinental

name : chili's grill & bar
category : bars;take out restaurants;american restaurants;bar & grill
```

```
s;restaurants
postalcode : 77032
address : ca1 services george bush- houston intercontiential airpor 31
00 terminal way
city : houston
phone : (281) 230-3485
street : terminal way
house_number : 3100
house : ca1 services george bush houston intercontiential

42/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), TfidfTex
tCanopyPredicate: (0.4, street))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGra
mCanopyPredicate: (0.6, address))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : sushi tadokoro
category : sushi bars;japanese
postalcode : 92110
address : 2244 san diego ave ste c
city : san diego
phone : (619) 297-0298
street : san diego ave
house_number : 2244
house : None

name : sushi tadokoro
category : sushi bars;restaurants
postalcode : 92110
address : 2244 san diego avenue
city : san diego
phone : (619) 297-0298
street : san diego avenue
house_number : 2244
house : None

43/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), TfidfTe
xtCanopyPredicate: (0.4, street))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, house), Simp
lePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGr
amCanopyPredicate: (0.6, address))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
name : la madeleine
category : french restaurants;restaurants;breakfast brunch & lunch re
staurants;bakeries;caterers;sandwich shops;coffee & tea;beverages
postalcode : 75205
address : 3072 mockingbird ln
city : dallas
phone : (214) 696-0800
street : mockingbird ln
house_number : 3072
house : None

name : la madeleine
category : breakfast brunch & lunch restaurants;bakeries;restaurants;
caterers;sandwich shops;coffee & tea;beverages;french restaurants
postalcode : 75235
address : 3072 w mockingbird ln
city : dallas
phone : (214) 696-0800
street : w mockingbird ln
house_number : 3072
house : None

44/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
f

Finished labeling
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), TfidfTex
tCanopyPredicate: (0.4, street))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfNGramCanopyPredicate: (0.6, name))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, house), Simpl
ePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGra
mCanopyPredicate: (0.6, address))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, categor
y), TfidfTextCanopyPredicate: (0.8, phone))
```

In [56]:

```
deduper.train()
```

```
INFO:rlr.crossvalidation:using cross validation to find optimum alph
a...
INFO:rlr.crossvalidation:optimum alpha: 0.100000, score 0.522347173396
5027
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (1, phone), SimplePr
edicate: (firstTokenPredicate, address))
INFO:dedupe.training:(PartialPredicate: (alphaNumericPredicate, name,
CorporationName), SimplePredicate: (sameSevenCharStartPredicate, nam
e))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, house), Simpl
ePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, name), TfidfNGra
mCanopyPredicate: (0.6, address))
```

In [33]:

```
with open(training_file, 'w') as tf:
    deduper.writeTraining(tf)
```

In [34]:

```
with open(settings_file, 'wb') as sf:
    deduper.writeSettings(sf)
```

run dedupe based on prior settings file

In [154]:

```
deduper = None
with open(settings_file, 'rb') as f:
    deduper = dedupe.StaticDedupe(f)
```

```
----------------------------------------------------------------------
-----
FileNotFoundError                         Traceback (most recent call
 last)
<ipython-input-154-f407f66db873> in <module>
      1 deduper = None
----> 2 with open(settings_file, 'rb') as f:
      3     deduper = dedupe.StaticDedupe(f)

FileNotFoundError: [Errno 2] No such file or directory: 'yellow_yelp_a
ll_pypostal2_learned_settings1'
```

In [57]:

```
threshold = deduper.threshold(data_dict, recall_weight=1)
```

```
INFO:dedupe.canopy_index:Removing stop word  s
INFO:dedupe.canopy_index:Removing stop word 9
INFO:dedupe.canopy_index:Removing stop word st
INFO:dedupe.canopy_index:Removing stop word  d
INFO:dedupe.canopy_index:Removing stop word 10
INFO:dedupe.canopy_index:Removing stop word d
INFO:dedupe.canopy_index:Removing stop word er
INFO:dedupe.canopy_index:Removing stop word rd
INFO:dedupe.canopy_index:Removing stop word  m
INFO:dedupe.canopy_index:Removing stop word 30
INFO:dedupe.canopy_index:Removing stop word ar
INFO:dedupe.canopy_index:Removing stop word h
INFO:dedupe.canopy_index:Removing stop word ma
INFO:dedupe.canopy_index:Removing stop word re
INFO:dedupe.canopy_index:Removing stop word t
INFO:dedupe.canopy_index:Removing stop word th
INFO:dedupe.canopy_index:Removing stop word  a
INFO:dedupe.canopy_index:Removing stop word 20
INFO:dedupe.canopy_index:Removing stop word av
INFO:dedupe.canopy_index:Removing stop word  w
INFO:dedupe.canopy_index:Removing stop word 2
INFO:dedupe.canopy_index:Removing stop word es
INFO:dedupe.canopy_index:Removing stop word  l
INFO:dedupe.canopy_index:Removing stop word 13
INFO:dedupe.canopy_index:Removing stop word e
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word n
INFO:dedupe.canopy_index:Removing stop word te
INFO:dedupe.canopy_index:Removing stop word s
INFO:dedupe.canopy_index:Removing stop word  n
INFO:dedupe.canopy_index:Removing stop word 7
INFO:dedupe.canopy_index:Removing stop word  c
INFO:dedupe.canopy_index:Removing stop word  1
INFO:dedupe.canopy_index:Removing stop word 4
INFO:dedupe.canopy_index:Removing stop word la
INFO:dedupe.canopy_index:Removing stop word r
INFO:dedupe.canopy_index:Removing stop word 60
INFO:dedupe.canopy_index:Removing stop word y
INFO:dedupe.canopy_index:Removing stop word 50
INFO:dedupe.canopy_index:Removing stop word ay
INFO:dedupe.canopy_index:Removing stop word oa
INFO:dedupe.canopy_index:Removing stop word wa
INFO:dedupe.canopy_index:Removing stop word 00
INFO:dedupe.canopy_index:Removing stop word 01
INFO:dedupe.canopy_index:Removing stop word nt
INFO:dedupe.canopy_index:Removing stop word  r
INFO:dedupe.canopy_index:Removing stop word me
INFO:dedupe.canopy_index:Removing stop word 16
INFO:dedupe.canopy_index:Removing stop word bl
INFO:dedupe.canopy_index:Removing stop word en
INFO:dedupe.canopy_index:Removing stop word 25
INFO:dedupe.canopy_index:Removing stop word a
INFO:dedupe.canopy_index:Removing stop word le
INFO:dedupe.canopy_index:Removing stop word or
```

```
INFO:dedupe.canopy_index:Removing stop word nu
INFO:dedupe.canopy_index:Removing stop word ue
INFO:dedupe.canopy_index:Removing stop word  e
INFO:dedupe.canopy_index:Removing stop word  f
INFO:dedupe.canopy_index:Removing stop word ra
INFO:dedupe.canopy_index:Removing stop word 11
INFO:dedupe.canopy_index:Removing stop word ou
INFO:dedupe.canopy_index:Removing stop word  p
INFO:dedupe.canopy_index:Removing stop word el
INFO:dedupe.canopy_index:Removing stop word 15
INFO:dedupe.canopy_index:Removing stop word ll
INFO:dedupe.canopy_index:Removing stop word ul
INFO:dedupe.canopy_index:Removing stop word 8
INFO:dedupe.canopy_index:Removing stop word nd
INFO:dedupe.canopy_index:Removing stop word va
INFO:dedupe.canopy_index:Removing stop word 6
INFO:dedupe.canopy_index:Removing stop word bo
INFO:dedupe.canopy_index:Removing stop word to
INFO:dedupe.canopy_index:Removing stop word ca
INFO:dedupe.blocking:10000, 9.3271432 seconds
INFO:dedupe.blocking:20000, 16.8941182 seconds
INFO:dedupe.api:Maximum expected recall and precision
INFO:dedupe.api:recall: 0.965
INFO:dedupe.api:precision: 0.962
INFO:dedupe.api:With threshold: 0.483
```

In [61]:

```
print('# duplicate sets', len(clustered_dupes))
```

```
# duplicate sets 4493
```

In [62]:

```
cluster_membership = {}
cluster_id = 0
for (cluster_id, cluster) in enumerate(clustered_dupes):
    id_set, scores = cluster
    cluster_d = [data_dict[c] for c in id_set]
    canonical_rep = dedupe.canonicalize(cluster_d)
    for record_id, score in zip(id_set, scores):
        cluster_membership[record_id] = {
            "cluster id" : cluster_id,
            "canonical representation" : canonical_rep,
            "confidence": score
        }
```

In [63]:
```
singleton_id = cluster_id + 1
with open(output_file, 'w') as f_output, open(fp) as f_input:
    writer = csv.writer(f_output)
    reader = csv.reader(f_input)

    heading_row = next(reader)
    heading_row.insert(0, 'confidence_score')
    heading_row.insert(0, 'Cluster ID')
    canonical_keys = canonical_rep.keys()
    for key in canonical_keys:
        heading_row.append('canonical_' + key)

    writer.writerow(heading_row)

    for row in reader:
        row_id = int(row[0])
        if row_id in cluster_membership:
            cluster_id = cluster_membership[row_id]["cluster id"]
            canonical_rep = cluster_membership[row_id]["canonical repre
            row.insert(0, cluster_membership[row_id]['confidence'])
            row.insert(0, cluster_id)
            for key in canonical_keys:
                row.append(canonical_rep[key].encode('utf8'))
        else:
            row.insert(0, None)
            row.insert(0, singleton_id)
            singleton_id += 1
            for key in canonical_keys:
                row.append(None)
        writer.writerow(row)
```

Predictions

In [64]:
```
df = pd.read_csv(output_file)
```

In [55]:
```
df.columns
```

Out[55]:
```
Index(['Cluster ID', 'confidence_score', 'Id', 'source', 'name', 'cate
gory',
       'phone', 'city', 'address', 'street', 'house_number', 'house',
       'canonical_Id', 'canonical_source', 'canonical_name',
       'canonical_category', 'canonical_phone', 'canonical_city',
       'canonical_address', 'canonical_street', 'canonical_house_numbe
r',
       'canonical_house'],
      dtype='object')
```

In [3]:
```
import pandas as pd
```

In [133]:
```
df = pd.read_csv(output_file)
df.sort_values(['Cluster ID'], inplace=True)
relevant_data = df[['Cluster ID', 'confidence_score', 'source', 'id']]
```

In [72]:

```
df = pd.read_csv(output_file)
df.sort_values(['Cluster ID'], inplace=True)
relevant_data = df[['Cluster ID', 'confidence_score', 'source', 'id']]

predictions = []
cluster_ids = relevant_data['Cluster ID'].value_counts()
for cluster_id in cluster_ids[cluster_ids>1].index:

    fodors_ids = relevant_data[
        (relevant_data['Cluster ID'] == cluster_id) &
        (relevant_data['source'] == 'yellow_pages')
    ].id.values
    zagats_ids = relevant_data[
        (relevant_data['Cluster ID'] == cluster_id) &
        (relevant_data['source'] == 'yelp')
    ].id.values

    match_interim = list(product(fodors_ids, zagats_ids))
    predictions.append(match_interim)

m = []
for cluster in predictions:
    for combo in cluster:
        m.append([combo[0], combo[1]])

predictions = pd.DataFrame(m, columns=['yellow_pages_id', 'yelp_id'])

predictions['yp-y'] = predictions.apply(lambda row: f"{row['yellow_page
```

In [83]:

```
results = pd.read_csv(matches_file)
results['yp-y'] = results.apply(lambda row: f"{row['yellow_pages_id']}-
```

In [131]:

```
pred_set = set(preds_comparable_with_duplicates['yp-y'].values.tolist()
```

In [75]:

```
results = results[results['duplicate'] == 1]
```

In [76]:

```
len(results)
```

Out[76]:

```
84
```

In [130]:

```
res_set = set(duplicates['yp-y'].values.tolist())
#pred_set = set(predictions['yp-y'].values.tolist())
```

In [132]:

```
tp = len(res_set & pred_set)
fn = len(res_set-pred_set)
fp = len(pred_set-res_set)

print(f'tp: {tp} fp: {fp} fn: {fn}')
```

```
tp: 57 fp: 74 fn: 27
```