Load Data

In [2]:
```
import os
import pandas as pd
import numpy as np
import logging
import csv
import re
import logging
import optparse
import dedupe
from unidecode import unidecode
from itertools import product
```

Setup

In [18]:
```
input_file = 'fodors_zagats_pypostal.csv'
output_file = 'fodors_zagats_pypostal_output.csv'
settings_file = 'fodors_zagats_learned_settings'
training_file = 'fodors_zagats_training.json'
```

In [19]:
```
fp = os.path.join(input_file)
```

Dataframe view

In [21]:
```
input_df = pd.read_csv(fp, sep=',', quotechar='"')
```

In [8]:
```
categories = list(input_df['category'].unique())
categories = [x for x in categories if str(x) != 'nan']
```

In [9]:
```
#category_corpus = input_df[['name', 'category']].drop_duplicates().to_
category_corpus = input_df.drop_duplicates().to_dict(orient='records')
```

Dedupe

Import modules

In [12]:
```
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 `No
        if not val:
            val = None
    except Exception as e:
        print(e)
    return val
```

In [13]:
```
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 [14]:
```
data_dict = get_clean_data_dict(fp)
```

Define the Fields for dedupe

In [16]:
```
fields = [
    {'field' : 'name', 'type': 'Name'},
    {'field' : 'category',
     'type': 'FuzzyCategorical',
     'categories': categories,
     'corpus': category_corpus,
     'has missing' : True},
    {'field' : 'name', 'type': 'String'},
    {'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 [17]:
```
deduper = dedupe.Dedupe(fields)
```

In [18]:
```
deduper.prepare_training(data_dict)
```

```
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinCanopyPredicate: (2, house), SimplePr
edicate: (sameSevenCharStartPredicate, street))
```

In [19]:
```
dedupe.consoleLabel(deduper)
```

```
name : rainbow room
category : or 212/632-5100 american
address : 30 rockefeller plaza
city : new york
phone : 212/632-5000
street : None
house_number : None
house : 30 rockefeller plaza

name : rainbow room
category : american (new)
address : 30 rockefeller plaza
city : new york city
phone : 212-632-5000
street : None
house_number : None
house : 30 rockefeller plaza

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

y

name : rose pistola
category : italian
address : 532 columbus avenue
city : san francisco
phone : 415/399-0499
street : columbus avenue
house_number : 532
house : None

name : rose pistola
category : italian
address : 532 columbus avenue
city : san francisco
phone : 415-399-0499
street : columbus avenue
house_number : 532
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:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
name : georgia grille
category : american
address : 2290 peachtree road peachtree square shopping center
city : atlanta
phone : 404/352-3517
```

```
street : peachtree road
house_number : 2290
house : square shopping center

name : georgia grille
category : southwestern
address : 2290 peachtree road
city : atlanta
phone : 404-352-3517
street : peachtree road
house_number : 2290
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:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
name : jo jo
category : american
address : 160 e. 64th st.
city : new york
phone : 212/223-5656
street : e. 64th st.
house_number : 160
house : None

name : jo jo
category : french bistro
address : 160 e. 64th st.
city : new york city
phone : 212-223-5656
street : e. 64th st.
house_number : 160
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

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
name : le bernardin
category : french
address : 155 w. 51st st.
```

```
city : new york
phone : 212/489-1515
street : w. 51st st.
house_number : 155
house : None

name : le bernardin
category : seafood
address : 155 w. 51st st.
city : new york city
phone : 212-489-1515
street : w. 51st st.
house_number : 155
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:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (twoGramFingerprint, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : ritz-carlton restaurant
category : french (classic)
address : 181 peachtree st.
city : atlanta
phone : 404-659-0400
street : peachtree st.
house_number : 181
house : None

name : ritz-carlton cafe (atlanta)
category : american (new)
address : 181 peachtree st.
city : atlanta
phone : 404-659-0400
street : peachtree st.
house_number : 181
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

n

name : plumpjack cafe
category : mediterranean
address : 3201 fillmore st.
city : san francisco
phone : 415/563-4755
street : fillmore st.
house_number : 3201
house : None
```

```
name : plumpjack cafe
category : american (new)
address : 3127 fillmore st.
city : san francisco
phone : 415-563-4755
street : fillmore st.
house_number : 3127
house : None

5/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : lulu
category : mediterranean
address : 816 folsom st.
city : san francisco
phone : 415/495-5775
street : folsom st.
house_number : 816
house : None

name : lulu restaurant-bis-cafe
category : mediterranean
address : 816 folsom st.
city : san francisco
phone : 415-495-5775
street : folsom st.
house_number : 816
house : None

6/10 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:(PartialPredicate: (sameFiveCharStartPredicate,
name, Surname), SimplePredicate: (twoGramFingerprint, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, na
me, Surname), SimplePredicate: (twoGramFingerprint, street))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : rex il ristorante
category : italian
address : 617 s. olive st.
city : los angeles
phone : 213/627-2300
street : s. olive st.
house_number : 617
house : None

name : rex il ristorante
category : nuova cucina italian
```

```
address : 617 s. olive st.
city : los angeles
phone : 213-627-2300
street : s. olive st.
house_number : 617
house : None

7/10 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:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : matsuhisa
category : asian
address : 129 n. la cienega blvd.
city : beverly hills
phone : 310/659-9639
street : n. la cienega blvd.
house_number : 129
house : None

name : matsuhisa
category : seafood
address : 129 n. la cienega blvd.
city : beverly hills
phone : 310-659-9639
street : n. la cienega blvd.
house_number : 129
house : None

8/10 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:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(PartialPredicate: (wholeFieldPredicate, name, C
orporationName), TfidfNGramCanopyPredicate: (0.8, city))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, addres
s), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, na
```

```
me, Surname), SimplePredicate: (twoGramFingerprint, street))
name : island spice
category : tel caribbean
address : 402 w. 44th st.
city : new york
phone : 212/765-1737
street : w. 44th st.
house_number : 402
house : None

name : island spice
category : caribbean
address : 402 w. 44th st.
city : new york city
phone : 212-765-1737
street : w. 44th st.
house_number : 402
house : None

9/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : 21 club
category : american
address : 21 w. 52nd st.
city : new york
phone : 212/582-7200
street : w. 52nd st.
house_number : 21
house : None

name : 21 club
category : american (new)
address : 21 w. 52nd st.
city : new york city
phone : 212-582-7200
street : w. 52nd st.
house_number : 21
house : None

10/10 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:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(PartialPredicate: (fingerprint, name, Corporatio
nName), TfidfTextCanopyPredicate: (0.8, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
```

```
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : hedgerose heights inn
category : international
address : 490 e. paces ferry rd.
city : atlanta
phone : 404/233-7673
street : e. paces ferry rd.
house_number : 490
house : None

name : hedgerose heights inn the
category : continental
address : 490 e. paces ferry rd. ne
city : atlanta
phone : 404-233-7673
street : e. paces ferry rd. ne
house_number : 490
house : None

11/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : steak house
category : steak houses
address : 2880 las vegas blvd. s
city : las vegas
phone : 702/734-0410
street : las vegas blvd. s
house_number : 2880
house : None

name : steak house the
category : steakhouses
address : 2880 las vegas blvd. s.
city : las vegas
phone : 702-734-0410
street : las vegas blvd. s.
house_number : 2880
house : None

12/10 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:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(PartialPredicate: (oneGramFingerprint, name, Cor
porationName), TfidfTextCanopyPredicate: (0.8, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
```

```
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : tillerman
category : seafood
address : 2245 e. flamingo rd.
city : las vegas
phone : 702/731-4036
street : e. flamingo rd.
house_number : 2245
house : None

name : tillerman the
category : steakhouses
address : 2245 e. flamingo rd.
city : las vegas
phone : 702-731-4036
street : e. flamingo rd.
house_number : 2245
house : None

13/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : gramercy tavern
category : american
address : 42 e. 20th st. between park ave. s and broadway
city : new york
phone : 212/477-0777
street : e. 20th st. between park ave. s and broadway
house_number : 42
house : None

name : gramercy tavern
category : american (new)
address : 42 e. 20th st.
city : new york city
phone : 212-477-0777
street : e. 20th st.
house_number : 42
house : None

14/10 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:(LevenshteinCanopyPredicate: (4, address), Partia
lPredicate: (sameSevenCharStartPredicate, name, CorporationName))
INFO:dedupe.training:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
```

```
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : heera of india
category : asian
address : 595 piedmont avenue rio shopping center mall
city : atlanta
phone : 404/876-4408
street : piedmont avenue
house_number : 595
house : rio shopping center mall

name : heera of india
category : indian
address : 595 piedmont avenue
city : atlanta
phone : 404-876-4408
street : piedmont avenue
house_number : 595
house : None

15/10 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:(PartialPredicate: (sameSevenCharStartPredicate,
name, CorporationName), SimplePredicate: (commonThreeTokens, street))
INFO:dedupe.training:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : boulevard
category : american
address : 1 mission st.
city : san francisco
phone : 415/543-6084
street : mission st.
house_number : 1
house : None

name : boulevard
category : american (new)
address : 1 mission st.
city : san francisco
phone : 415-543-6084
street : mission st.
house_number : 1
house : None
```

```
16/10 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:(PartialIndexTfidfNGramCanopyPredicate: (0.8, add
ress, StreetName), PartialPredicate: (sameSevenCharStartPredicate, nam
e, CorporationName))
INFO:dedupe.training:(PartialPredicate: (fingerprint, name, Surname),
SimplePredicate: (commonIntegerPredicate, street))
INFO:dedupe.training:(SimplePredicate: (wholeFieldPredicate, address),
SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : aqua
category : seafood
address : 252 california st.
city : san francisco
phone : 415/956-9662
street : california st.
house_number : 252
house : None

name : aqua
category : american (new)
address : 252 california st.
city : san francisco
phone : 415-956-9662
street : california st.
house_number : 252
house : None

17/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : montrachet
category : french
address : 239 w. broadway between walker and white sts.
city : new york
phone : 212/ 219-2777
street : w. broadway between walker and white sts.
house_number : 239
house : None

name : montrachet
category : french bistro
address : 239 w. broadway
city : new york city
phone : 212-219-2777
street : w. broadway
```

```
house_number : 239
house : None

18/10 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:(PartialIndexTfidfNGramCanopyPredicate: (0.8, add
ress, StreetName), PartialPredicate: (sameSevenCharStartPredicate, nam
e, CorporationName))
INFO:dedupe.training:(PartialIndexLevenshteinCanopyPredicate: (1, nam
e, CorporationName), SimplePredicate: (wholeFieldPredicate, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : coyote cafe
category : southwestern
address : 3799 las vegas blvd. s
city : las vegas
phone : 702/891-7349
street : las vegas blvd. s
house_number : 3799
house : None

name : coyote cafe (las vegas)
category : southwestern
address : 3799 las vegas blvd. s.
city : las vegas
phone : 702-891-7349
street : las vegas blvd. s.
house_number : 3799
house : None

19/10 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:(PartialPredicate: (firstTokenPredicate, name, Co
rporationName), PartialPredicate: (sameFiveCharStartPredicate, addres
s, StreetName))
INFO:dedupe.training:(PartialIndexLevenshteinCanopyPredicate: (1, nam
e, CorporationName), SimplePredicate: (wholeFieldPredicate, address))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, nam
e, Surname), SimplePredicate: (twoGramFingerprint, street))
name : lespinasse
category : american
address : 2 e. 55th st.
city : new york
phone : 212/339-6719
street : e. 55th st.
house_number : 2
house : None
```

```
name : lespinasse (new york city)
category : asian
address : 2 e. 55th st.
city : new york city
phone : 212-339-6719
street : e. 55th st.
house_number : 2
house : None

20/10 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: (firstTokenPredicate, name), Tf
idfTextCanopyPredicate: (0.8, street))
INFO:dedupe.training:(PartialIndexLevenshteinCanopyPredicate: (2, nam
e, CorporationName), PartialIndexTfidfTextCanopyPredicate: (0.2, name,
CorporationName))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
name : arnie morton\s of chicago
category : american
address : 435 s. la cienega blv.
city : los angeles
phone : 310/246-1501
street : s. la cienega blv.
house_number : 435
house : None

name : arnie morton\s of chicago
category : steakhouses
address : 435 s. la cienega blvd.
city : los angeles
phone : 310-246-1501
street : s. la cienega blvd.
house_number : 435
house : None

21/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : brasserie le coze
category : french
address : 3393 peachtree road lenox square mall near neiman marcus
city : atlanta
phone : 404/266-1440
street : peachtree road lenox square mall near neiman marcus
house_number : 3393
house : None
```

```
name : brasserie le coze
category : french bistro
address : 3393 peachtree road
city : atlanta
phone : 404-266-1440
street : peachtree road
house_number : 3393
house : None

22/10 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: (firstTokenPredicate, name), Tf
idfTextCanopyPredicate: (0.6, street))
INFO:dedupe.training:(PartialIndexLevenshteinCanopyPredicate: (2, nam
e, CorporationName), PartialIndexTfidfTextCanopyPredicate: (0.2, name,
CorporationName))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, Surname), SimplePredicate: (sameFiveCharStartPredicate, house))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), PartialPredicate: (alphaNumericPredicate, addres
s, StreetName))
name : la cote basque
category : french
address : 60 w. 55th st. between 5th and 6th ave.
city : new york
phone : 212/688-6525
street : w. 55th st. between 5th and 6th ave.
house_number : 60
house : None

name : la cote basque
category : french (classic)
address : 60 w. 55th st.
city : new york city
phone : 212-688-6525
street : w. 55th st.
house_number : 60
house : None

23/10 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: (firstTokenPredicate, name), S
implePredicate: (sameSevenCharStartPredicate, address))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : les celebrites
category : french
```

```
address : 160 central park s
city : new york
phone : 212/484-5113
street : central park s
house_number : 160
house : None

name : les celebrites
category : french (classic)
address : 155 w. 58th st.
city : new york city
phone : 212-484-5113
street : w. 58th st.
house_number : 155
house : None

24/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
n

name : tavern on the green
category : american
address : in central park at 67th st.
city : new york
phone : 212/873-3200
street : None
house_number : None
house : in central park at 67th st.

name : tavern on the green
category : american (new)
address : central park west
city : new york city
phone : 212-873-3200
street : None
house_number : None
house : central park west

24/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
y

name : dante\s down the hatch buckhead
category : continental
address : 3380 peachtree road
city : atlanta
phone : 404/266-1600
street : peachtree road
house_number : 3380
house : None

name : dante\s down the hatch
category : continental
address : underground underground mall underground atlanta
```

```
city : atlanta
phone : 404/577-1800
street : None
house_number : None
house : underground underground mall underground

25/10 positive, 2/10 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), Si
mplePredicate: (sameSevenCharStartPredicate, address))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : cafe ritz-carlton buckhead
category : ext 6108 international
address : 3434 peachtree road
city : atlanta
phone : 404/237-2700
street : peachtree road
house_number : 3434
house : None

name : dining room ritz-carlton buckhead
category : international
address : 3434 peachtree road
city : atlanta
phone : 404/237-2700
street : peachtree road
house_number : 3434
house : None

26/10 positive, 2/10 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: (commonTwoTokens, city), Simpl
ePredicate: (wholeFieldPredicate, phone))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Cor
porationName), SimplePredicate: (oneGramFingerprint, city))
INFO:dedupe.training:(PartialIndexLevenshteinCanopyPredicate: (2, nam
e, CorporationName), PartialIndexTfidfTextCanopyPredicate: (0.2, nam
e, CorporationName))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, na
me, Surname), SimplePredicate: (twoGramFingerprint, street))
name : manhattan ocean club
category : seafood
address : 57 w. 58th st.
city : new york
phone : 212/ 371-7777
```

```
street : w. 58th st.
house_number : 57
house : None

name : manhattan ocean club
category : seafood
address : 57 w. 58th st.
city : new york city
phone : 212-371-7777
street : w. 58th st.
house_number : 57
house : None

26/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : aureole
category : american
address : 34 e. 61st st.
city : new york
phone : 212/ 319-1660
street : e. 61st st.
house_number : 34
house : None

name : aureole
category : american (new)
address : 34 e. 61st st.
city : new york city
phone : 212-319-1660
street : e. 61st st.
house_number : 34
house : None

27/10 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: (fingerprint, address), Simple
Predicate: (firstTokenPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Cor
porationName), SimplePredicate: (oneGramFingerprint, city))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (sameFiveCharStartPredicate, st
reet))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.6, na
me, Surname), SimplePredicate: (twoGramFingerprint, street))
name : mi cocina
category : mexican
address : 57 jane st. off hudson st.
```

```
city : new york
phone : 212/627-8273
street : jane st. off hudson st.
house_number : 57
house : None

name : mi cocina
category : mexican
address : 57 jane st.
city : new york city
phone : 212-627-8273
street : jane st.
house_number : 57
house : None

28/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : la grotta at ravinia dunwoody rd.
category : italian
address : holiday inn crowne plaza at ravinia dunwoody
city : atlanta
phone : 770/395-9925
street : None
house_number : None
house : holiday inn crowne plaza at ravinia dunwoody

name : la grotta
category : italian
address : 2637 peachtree rd. ne
city : atlanta
phone : 404-231-1368
street : peachtree rd. ne
house_number : 2637
house : None

29/10 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: (firstTokenPredicate, name), S
implePredicate: (sameSevenCharStartPredicate, address))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Cor
porationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : la grotta
category : italian
address : 2637 peachtree road peachtree house condominium
city : atlanta
```

```
phone : 404/231-1368
street : peachtree road
house_number : 2637
house : peachtree house condominium

name : la grotta at ravinia dunwoody rd.
category : italian
address : holiday inn crowne plaza at ravinia dunwoody
city : atlanta
phone : 770/395-9925
street : None
house_number : None
house : holiday inn crowne plaza at ravinia dunwoody

29/10 positive, 4/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious


n

name : pisces
category : seafood
address : 95 ave. a at 6th st.
city : new york
phone : 212/260-6660
street : ave. a at 6th st.
house_number : 95
house : None

name : pisces
category : seafood
address : 95 avenue a
city : new york city
phone : 212-260-6660
street : avenue a
house_number : 95
house : None

29/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : philippe\s the original
category : american
address : 1001 n. alameda st.
city : los angeles
phone : 213/628-3781
street : n. alameda st.
house_number : 1001
house : None

name : philippe the original
category : cafeterias
address : 1001 n. alameda st.
city : chinatown
```

```
phone : 213-628-3781
street : n. alameda st.
house_number : 1001
house : None

30/10 positive, 5/10 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), Si
mplePredicate: (sameFiveCharStartPredicate, address))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : le montrachet
category : continental
address : 3000 w. paradise rd.
city : las vegas
phone : 702/732-5111
street : w. paradise rd.
house_number : 3000
house : None

name : le montrachet bistro
category : french bistro
address : 3000 paradise road
city : las vegas
phone : 702-732-5651
street : paradise road
house_number : 3000
house : None

31/10 positive, 5/10 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:(PartialPredicate: (sameFiveCharStartPredicate, n
ame, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : mesa grill
category : american
address : 102 5th ave. between 15th and 16th sts.
city : new york
phone : 212/807-7400
```

```
street : 5th ave. between 15th and 16th sts.
house_number : 102
house : None

name : mesa grill
category : southwestern
address : 102 fifth ave.
city : new york city
phone : 212-807-7400
street : fifth ave.
house_number : 102
house : None

32/10 positive, 5/10 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: (2, phone), SimplePr
edicate: (sameThreeCharStartPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
name : second street grille
category : seafood
address : 200 e. fremont st.
city : las vegas
phone : 702/385-3232
street : e. fremont st.
house_number : 200
house : None

name : second street grill
category : pacific rim
address : 200 e. fremont st.
city : las vegas
phone : 702-385-6277
street : e. fremont st.
house_number : 200
house : None

33/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : palace court
category : continental
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7547
street : las vegas blvd. s
house_number : 3570
house : None

name : palace court
category : french (new)
```

```
address : 3570 las vegas blvd. s.
city : las vegas
phone : 702-731-7110
street : las vegas blvd. s.
house_number : 3570
house : None

34/10 positive, 5/10 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: (3, phone), SimplePr
edicate: (sameThreeCharStartPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
name : virgil\s
category : american
address : 152 w. 44th st.
city : new york
phone : 212/ 921-9494
street : w. 44th st.
house_number : 152
house : None

name : virgil\s real bbq
category : bbq
address : 152 w. 44th st.
city : new york city
phone : 212-921-9494
street : w. 44th st.
house_number : 152
house : None

35/10 positive, 5/10 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: (sameThreeCharStartPredicate,
address), SimplePredicate: (sameThreeCharStartPredicate, name))
INFO:dedupe.training:(PartialPredicate: (sameFiveCharStartPredicate,
name, CorporationName), TfidfNGramCanopyPredicate: (0.4, house))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Cor
porationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialIndexTfidfTextCanopyPredicate: (0.8, nam
e, CorporationName), SimplePredicate: (twoGramFingerprint, city))
name : toulouse
category : french
address : b peachtree road
city : atlanta
phone : 404/351-9533
street : b peachtree road
```

```
house_number : None
house : None

name : toulouse
category : french (new)
address : 293-b peachtree rd.
city : atlanta
phone : 404-351-9533
street : peachtree rd.
house_number : 293-b
house : None

36/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : the palm
category : american
address : 9001 santa monica boulevard
city : los angeles
phone : 310/550-8811
street : santa monica boulevard
house_number : 9001
house : None

name : palm the (los angeles)
category : steakhouses
address : 9001 santa monica boulevard
city : w. hollywood
phone : 310-550-8811
street : santa monica boulevard
house_number : 9001
house : None

37/10 positive, 5/10 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: (sameThreeCharStartPredicate, n
ame), TfidfTextCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
name : pacific dining car
category : american
address : 6th st.
city : los angeles
phone : 213/483-6000
street : 6th st.
house_number : None
house : None

name : pacifica
category : asian
```

```
address : 138 lafayette st. between canal and howard sts.
city : new york
phone : 212/941-4168
street : lafayette st. between canal and howard sts.
house_number : 138
house : None

38/10 positive, 5/10 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: (sameThreeCharStartPredicate, n
ame), TfidfTextCanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialIndexTfidfNGramCanopyPredicate: (0.4, nam
e, CorporationName), PartialPredicate: (commonTwoTokens, address, Stre
etName))
name : pacifica
category : asian
address : 138 lafayette st. between canal and howard sts.
city : new york
phone : 212/941-4168
street : lafayette st. between canal and howard sts.
house_number : 138
house : None

name : pacific pan pacific hotel
category : french
address : 500 post st.
city : san francisco
phone : 415/929-2087
street : post st.
house_number : 500
house : None

38/10 positive, 6/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : bistro
category : continental
address : 3400 las vegas blvd. s
city : las vegas
phone : 702/791-7111
street : las vegas blvd. s
house_number : 3400
house : None

name : mikado
category : asian
address : 3400 las vegas blvd. s
city : las vegas
```

```
phone : 702/791-7111
street : las vegas blvd. s
house_number : 3400
house : None

38/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
u

name : ritz-carlton cafe (buckhead)
category : american (new)
address : 3434 peachtree rd. ne
city : atlanta
phone : 404-237-2700
street : peachtree rd. ne
house_number : 3434
house : None

name : ritz-carlton cafe (atlanta)
category : american (new)
address : 181 peachtree st.
city : atlanta
phone : 404-659-0400
street : peachtree st.
house_number : 181
house : None

38/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : restaurant ritz-carlton atlanta
category : continental
address : 181 peachtree st.
city : atlanta
phone : 404/659-0400
street : peachtree st.
house_number : 181
house : None

name : ritz-carlton cafe (atlanta)
category : american (new)
address : 181 peachtree st.
city : atlanta
phone : 404-659-0400
street : peachtree st.
house_number : 181
house : None

38/10 positive, 8/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y
```

```
name : ritz-carlton restaurant and dining room
category : american
address : 600 stockton st.
city : san francisco
phone : 415/296-7465
street : stockton st.
house_number : 600
house : None

name : ritz-carlton dining room (san francisco)
category : french (new)
address : 600 stockton st.
city : san francisco
phone : 415-296-7465
street : stockton st.
house_number : 600
house : None

39/10 positive, 8/10 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: (0.4, name), TfidfText
CanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
name : palace court
category : continental
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7547
street : las vegas blvd. s
house_number : 3570
house : None

name : cafe roma
category : coffee shops/diners
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7547
street : las vegas blvd. s
house_number : 3570
house : None

40/10 positive, 8/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious


y

name : golden nugget hotel
category : buffets
address : 129 e. fremont st.
city : las vegas
```

```
phone : 702/385-7111
street : e. fremont st.
house_number : 129
house : None

name : lillie langtry\s
category : asian
address : 129 e. fremont st.
city : las vegas
phone : 702/385-7111
street : e. fremont st.
house_number : 129
house : None

41/10 positive, 8/10 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:(TfidfTextCanopyPredicate: (0.4, name), TfidfText
CanopyPredicate: (0.4, phone))
INFO:dedupe.training:(PartialPredicate: (commonThreeTokens, name, Corp
orationName), SimplePredicate: (wholeFieldPredicate, category))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), SimplePredicate: (commonIntegerPredicate, phone))
name : coyote cafe
category : southwestern
address : 3799 las vegas blvd. s
city : las vegas
phone : 702/891-7349
street : las vegas blvd. s
house_number : 3799
house : None

name : emeril\s new orleans fish house
category : seafood
address : 3799 las vegas blvd. s.
city : las vegas
phone : 702-891-7374
street : las vegas blvd. s.
house_number : 3799
house : None

41/10 positive, 8/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : golden nugget hotel
category : buffets
address : 129 e. fremont st.
city : las vegas
phone : 702/385-7111
street : e. fremont st.
```

```
house_number : 129
house : None

name : stefano\s
category : italian
address : 129 fremont st.
city : las vegas
phone : 702-385-7111
street : fremont st.
house_number : 129
house : None

41/10 positive, 9/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

name : palace court
category : continental
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7547
street : las vegas blvd. s
house_number : 3570
house : None

name : empress court
category : asian
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7888
street : las vegas blvd. s
house_number : 3570
house : None

41/10 positive, 9/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : empress court
category : asian
address : 3570 las vegas blvd. s
city : las vegas
phone : 702/731-7888
street : las vegas blvd. s
house_number : 3570
house : None

name : palace court
category : french (new)
address : 3570 las vegas blvd. s.
city : las vegas
phone : 702-731-7110
street : las vegas blvd. s.
house_number : 3570
```

```
house : None

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

f

Finished labeling
```

In [20]:
```
deduper.train()
```

```
INFO:rlr.crossvalidation:using cross validation to find optimum alph
a...
INFO:rlr.crossvalidation:optimum alpha: 0.100000, score 0.554502752179
9167
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(TfidfTextCanopyPredicate: (0.4, name), TfidfText
CanopyPredicate: (0.4, phone))
```

In [33]:
```
# use a saved training file
# #with open(training_file, 'w') as tf:
#    deduper.writeTraining(tf)
```

In [34]:
```
# use saved settings file for duplication
#with open(settings_file, 'wb') as sf:
#    deduper.writeSettings(sf)
```

run dedupe based on prior settings file

In [35]:
```
deduper = None
with open(settings_file, 'rb') as f:
    deduper = dedupe.StaticDedupe(f)
```

```
INFO:dedupe.api:((TfidfTextCanopyPredicate: (0.4, name), TfidfTextCano
pyPredicate: (0.4, phone)),)
```

In [36]:
```
threshold = deduper.threshold(data_dict, recall_weight=1)
```

```
INFO:dedupe.api:Maximum expected recall and precision
INFO:dedupe.api:recall: 0.986
INFO:dedupe.api:precision: 0.967
INFO:dedupe.api:With threshold: 0.544
```

In [37]:
```
clustered_dupes = deduper.match(data_dict, 0)
```

In [38]:
```
print('# duplicate sets', len(clustered_dupes))
```

```
# duplicate sets 112
```

In [39]:
```
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 [52]:
```
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 repres
            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 [54]:
```
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 [ ]:
```
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'] == 'fodors')
    ].Id.values
    zagats_ids = relevant_data[
        (relevant_data['Cluster ID'] == cluster_id) &
        (relevant_data['source'] == 'zagats')
    ].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=['fodors_id', 'zagats_id'])

predictions['f-z'] = predictions.apply(lambda row: f"{row['fodors_id']}


predictions.head(10)
```

In [57]:
```
results = pd.read_csv('matches_fodors_zagats.csv')
results['f-z'] = results.apply(lambda row: f"{row['fodors_id']}-{row['za


```

In [58]:
```
res_set = set(results['f-z'].values.tolist())
pred_set = set(predictions['f-z'].values.tolist())
```

In [59]:
```
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: 111 fp: 4 fn: 1
```