Load Data

In [1]:
```
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 [2]:
```
folder = r'/filepath_to_folder/'

input_file = 'yellow_yelp_all_pypostal.csv'
output_file = 'yellow_yelp_all_pypostal_output2.csv'
settings_file = 'yellow_yelp_all_pypostal_learned_settings2'
training_file = 'yellow_yelp_all_pypostal_training2.json'
```

In [3]:
```
fp = os.path.join(folder, input_file)
```

In [4]:
```
matches_file = os.path.join(folder, 'yellow_yelp_label.csv')
```

In [5]:
```
log_level = logging.INFO
log_level = logging.DEBUG
logging.getLogger().setLevel(log_level)
```

Dataframe view

In [6]:
```
input_df = pd.read_csv(fp, sep=',', quotechar='"', dtype={'postalcode':
```

In [7]:
```
input_df.columns
```

Out[7]:
```
Index(['id', 'source', 'name', 'category', 'phone', 'city', 'postalcod
e',
       'address', 'street', 'house_number', 'house'],
      dtype='object')
```

In [8]:
```
def get_clean_postalcode(x):

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

In [9]:
```
input_df['postalcode'] = input_df['postalcode'].apply(lambda x: get_clea
```

In [ ]:
```
input_df.head(5)
```

In [10]:
```
categories = list(input_df['category'].unique())
categories = [x for x in categories if str(x) != 'nan']
```

In [11]:
```
#category_corpus = input_df[['name', 'category']].drop_duplicates().to_
category_corpus = input_df.drop_duplicates().to_dict(orient='records')
```

In [12]:
```
category_corpus[0]
```

Out[12]:
```
{'id': 1,
 'source': 'yellow_pages',
 'name': 'Tao Tao Restaurant',
 'category': nan,
 'phone': '(408) 736-3731',
 'city': 'Sunnyvale ',
 'postalcode': '94086',
 'address': '175 S Murphy Ave',
 'street': 's murphy ave',
 'house_number': '175',
 'house': nan}
```

Dedupe

Import modules

In [15]:
```
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 [16]:
```
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 [20]:
```
data_dict = get_clean_data_dict(fp)
```

In [ ]:
```
data_dict
```

Define the Fields for dedupe

In [ ]:
```
def sameOrNotComparator(field_1, field_2) :
    if field_1 and field_2 :
        if field_1 == field_2 :
            return 0
        else:
            return 1
```

In [17]:
```
fields = [
    {'field' : 'name', 'type': 'Exact'},
    {'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': 'name', 'variable name': 'name', 'type': 'String' },
    {'field': 'postalcode', 'variable name': 'postalcode', 'type': 'Exa
    {'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 [18]:
```
deduper = dedupe.Dedupe(fields)
```

In [21]:
```
deduper.prepare_training(data_dict)
```

```
INFO:dedupe.canopy_index:Removing stop word ar
INFO:dedupe.canopy_index:Removing stop word ri
INFO:dedupe.canopy_index:Removing stop word an
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word n
INFO:dedupe.canopy_index:Removing stop word ra
INFO:dedupe.canopy_index:Removing stop word  c
INFO:dedupe.canopy_index:Removing stop word nt
INFO:dedupe.canopy_index:Removing stop word  r
INFO:dedupe.canopy_index:Removing stop word es
INFO:dedupe.canopy_index:Removing stop word re
INFO:dedupe.canopy_index:Removing stop word ta
INFO:dedupe.canopy_index:Removing stop word  b
INFO:dedupe.canopy_index:Removing stop word e
INFO:dedupe.canopy_index:Removing stop word er
INFO:dedupe.canopy_index:Removing stop word  s
INFO:dedupe.canopy_index:Removing stop word s
INFO:dedupe.canopy_index:Removing stop word ca
INFO:dedupe.canopy_index:Removing stop word al
INFO:dedupe.canopy_index:Removing stop word  6
```

In [24]:
```
dedupe.consoleLabel(deduper)
```

```
name : olive garden italian restaurant
category : None
address : None
city : None
postalcode : nan
phone : None
street : None
house_number : None
house : None

name : olive garden italian restaurant
category : None
address : None
city : None
postalcode : nan
phone : None
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 : olive garden italian restaurant
category : None
address : None
city : None
postalcode : nan
phone : None
street : None
house_number : None
house : None

name : olive garden italian restaurant
category : None
address : None
city : None
postalcode : nan
phone : (218) 212-3793
street : None
house_number : None
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

u

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (twoGramFingerprint, name), Sim
plePredicate: (wholeFieldPredicate, postalcode))
name : peter piper pizza
```

```
category : None
address : None
city : None
postalcode : nan
phone : None
street : None
house_number : None
house : None

name : peter piper pizza
category : None
address : 120 s new rd
city : waco
postalcode : 76710
phone : (254) 751-1212
street : s new rd
house_number : 120
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

name : olive garden italian restaurant
category : None
address : None
city : None
postalcode : nan
phone : None
street : None
house_number : None
house : None

name : olive garden italian restaurant
category : None
address : 3915 w war memorial dr
city : peoria
postalcode : 61615
phone : (888) 901-7571
street : w war memorial dr
house_number : 3915
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: (twoGramFingerprint, name), Sim
plePredicate: (wholeFieldPredicate, postalcode))
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, name), Tfid
fNGramCanopyPredicate: (0.8, name))
name : olive garden italian restaurant
category : None
```

```
address : None
city : None
postalcode : nan
phone : None
street : None
house_number : None
house : None

name : olive garden italian restaurant
category : None
address : 2641 n maize rd
city : wichita
postalcode : 67205
phone : (316) 512-7794
street : n maize rd
house_number : 2641
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

n

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
name : premier smog
category : None
address : serving mountain view and the surrounding area
city : phone number
postalcode : nan
phone : None
street : None
house_number : None
house : serving mountain view and the surrounding area

name : premier entertainment professional mobile dj service
category : None
address : serving san diego and the surrounding area
city : phone number
postalcode : nan
phone : None
street : None
house_number : None
house : serving san diego and the surrounding area

3/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : stone werks big rock grille
category : None
address : 5807 worth parkway
city : san antonio
postalcode : 78257
```

```
phone : (210) 558-9898
street : worth parkway
house_number : 5807
house : None

name : stone werks big rock grill
category : None
address : 5807 worth parkway
city : san antonio
postalcode : 78257
phone : (210) 558-9898
street : worth parkway
house_number : 5807
house : None

3/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : la fogata
category : None
address : 3025 e coast hwy
city : corona del mar
postalcode : 92625
phone : (949) 673-2211
street : e coast hwy
house_number : 3025
house : None

name : la fogata restaurant
category : None
address : 8 harbor pointe dr
city : corona del mar
postalcode : 92625
phone : (949) 673-2211
street : harbor pointe dr
house_number : 8
house : None

4/10 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:(ExistsPredicate: (Exists, house), LevenshteinCan
opyPredicate: (1, name))
name : koya
category : None
address : 508 e old elm rd
city : highland park
postalcode : 60035
phone : (847) 266-0891
street : e old elm rd
house_number : 508
```

```
house : None

name : koya japan
category : None
address : 508 old elm road
city : highland park
postalcode : 60035
phone : (847) 266-0891
street : old elm road
house_number : 508
house : None

5/10 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:(ExistsPredicate: (Exists, house), LevenshteinCan
opyPredicate: (1, name))
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, city), Tfid
fTextCanopyPredicate: (0.8, phone))
name : jax salon - moved
category : None
address : 1158 chestnut st
city : menlo park
postalcode : 94025
phone : (650) 323-4247
street : chestnut st
house_number : 1158
house : None

name : jax salon
category : None
address : 1610 el camino real
city : menlo park
postalcode : 94025
phone : (650) 323-4247
street : el camino real
house_number : 1610
house : None

6/10 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:(ExistsPredicate: (Exists, house), LevenshteinCan
opyPredicate: (1, name))
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, city), Tfid
fTextCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), TfidfNGramCanopyPredicate: (0.8, phone))
name : starbucks
category : None
```

```
address : 2 n central ave
city : saint louis
postalcode : 63105
phone : (314) 863-8070
street : n central ave
house_number : 2
house : None

name : starbucks coffee
category : None
address : 343 s kirkwood rd
city : saint louis
postalcode : 63122
phone : (314) 821-2377
street : s kirkwood rd
house_number : 343
house : None

7/10 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:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
name : pappasito's cantina
category : None
address : 11831 east freeway
city : houston
postalcode : 77029
phone : (713) 455-8378
street : east freeway
house_number : 11831
house : None

name : pappasito's cantina
category : None
address : 1600 lamar st hilton americas
city : houston
postalcode : 77010
phone : (713) 353-4400
street : lamar st
house_number : 1600
house : hilton americas

7/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : stage 62 delicatessen restaurant
category : None
address : 9105 strada pl
```

```
city : naples
postalcode : 34108
phone : (239) 597-2800
street : strada pl
house_number : 9105
house : None

name : stage 62 delicatessen & restaurant
category : None
address : 9105 strada pl ste 3125
city : naples
postalcode : 34108
phone : (239) 597-2800
street : strada pl
house_number : 9105
house : None

7/10 positive, 4/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : aaron's barbecue & grill
category : None
address : 403 harlin dr
city : gainesville
postalcode : 65655
phone : (417) 679-0061
street : harlin dr
house_number : 403
house : None

name : aaron's barbecue and grill
category : None
address : 403 harlin dr
city : gainesville
postalcode : 65655
phone : (417) 679-0061
street : harlin dr
house_number : 403
house : None

8/10 positive, 4/10 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:(ExistsPredicate: (Exists, house), SimplePredica
te: (wholeFieldPredicate, name))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simpl
ePredicate: (fingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (hundredIntegersOddPredicate,
name), SimplePredicate: (suffixArray, phone))
name : mykonos greek restaurant
category : None
```

```
address : 172 plandome road
city : manhasset
postalcode : 11030
phone : (516) 365-0113
street : plandome road
house_number : 172
house : None

name : mykonos restaurant
category : None
address : 172 plandome road
city : manhasset
postalcode : 11030
phone : (516) 365-0113
street : plandome road
house_number : 172
house : None

9/10 positive, 4/10 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:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (hundredIntegersOddPredicate, n
ame), SimplePredicate: (suffixArray, phone))
INFO:dedupe.training:(TfidfNGramCanopyPredicate: (0.8, address), Tfidf
NGramCanopyPredicate: (0.8, phone))
name : bonsai japaness cuisine
category : None
address : 3401 el camino real
city : atherton
postalcode : 94027
phone : (650) 367-6547
street : el camino real
house_number : 3401
house : None

name : bonsai japanese cuisine
category : None
address : 3401 w el camino real
city : atherton
postalcode : 94027
phone : (650) 367-6547
street : w el camino real
house_number : 3401
house : None

10/10 positive, 4/10 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), TfidfNG
ramCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
INFO:dedupe.training:(SimplePredicate: (hundredIntegersOddPredicate, n
ame), SimplePredicate: (suffixArray, phone))
name : pappas bbq
category : None
address : 12420 east freeway
city : houston
postalcode : 77015
phone : (832) 214-4078
street : east freeway
house_number : 12420
house : None

name : pappas bar-b-q
category : None
address : 12424 east freeway
city : houston
postalcode : 77015
phone : (832) 214-4078
street : east freeway
house_number : 12424
house : None

11/10 positive, 4/10 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:(ExistsPredicate: (Exists, house), LevenshteinCan
opyPredicate: (1, name))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(SimplePredicate: (fingerprint, address), TfidfNG
ramCanopyPredicate: (0.8, phone))
name : culinaria the best of mexico
category : None
address : 15900 la cantera pkwy
city : san antonio
postalcode : 78256
phone : (210) 582-6255
street : la cantera pkwy
house_number : 15900
house : None

name : culinaria 5k beer & wine run
category : None
address : 15900 la cantera pkwy
city : san antonio
postalcode : 78256
```

```
phone : None
street : la cantera pkwy
house_number : 15900
house : None

12/10 positive, 4/10 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, city), Si
mplePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
name : chez marie
category : None
address : 633 old post road
city : bedford
postalcode : 10506
phone : (914) 234-3992
street : old post road
house_number : 633
house : None

name : bedford playhouse apartments
category : None
address : 633 old post road
city : bedford
postalcode : 10506
phone : None
street : old post road
house_number : 633
house : None

13/10 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, n
ame), TfidfTextCanopyPredicate: (0.8, street))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
name : ihop
category : None
address : 1586 northern boulevard
city : manhasset
postalcode : 11030
phone : (516) 365-2732
street : northern boulevard
house_number : 1586
house : None
```

```
name : international house of pancakes
category : None
address : 1586 northern boulevard
city : manhasset
postalcode : 11030
phone : (516) 365-2732
street : northern boulevard
house_number : 1586
house : None

13/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
y

name : ristorante aida
category : None
address : 130 almshouse road
city : richboro
postalcode : 18954
phone : (215) 355-6660
street : almshouse road
house_number : 130
house : None

name : ristorante denicola
category : None
address : 130 almshouse rd ste 405
city : richboro
postalcode : 18954
phone : (215) 355-9066
street : almshouse rd
house_number : 130
house : None

14/10 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: (firstTokenPredicate, city), S
implePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredica
te: (wholeFieldPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), SimplePredicate: (sameSevenCharStartPredicate, name))
name : john p. fields
category : None
address : 26 n central ave
city : clayton
postalcode : 63105
phone : (314) 862-1886
street : n central ave
house_number : 26
```

```
house : None

name : john p fields restaurant
category : None
address : 26 n central ave
city : saint louis
postalcode : 63105
phone : (314) 862-1886
street : n central ave
house_number : 26
house : None

14/10 positive, 6/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : la hacienda super taqurias
category : None
address : 3211 orange avenue
city : fort pierce
postalcode : 34947
phone : None
street : orange avenue
house_number : 3211
house : None

name : la hacienda
category : None
address : 2403 falcon road
city : altus
postalcode : 73521
phone : (580) 379-4234
street : falcon road
house_number : 2403
house : None

15/10 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:(SimplePredicate: (fingerprint, address), TfidfNG
ramCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, n
ame), TfidfNGramCanopyPredicate: (0.8, address))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
INFO:dedupe.training:(SimplePredicate: (hundredIntegersOddPredicate, n
ame), SimplePredicate: (suffixArray, phone))
name : pizzaexpress
category : None
address : 901 e st
```

```
city : snyder
postalcode : 73566
phone : (580) 569-2007
street : e st
house_number : 901
house : None

name : pizza express
category : None
address : 701 e st
city : snyder
postalcode : 73566
phone : (580) 569-2007
street : e st
house_number : 701
house : None

15/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

name : blackwood country steak house
category : None
address : 1004 n san jacinto st
city : houston
postalcode : 77002
phone : (713) 221-9000
street : n san jacinto st
house_number : 1004
house : None

name : blackwood country steak restaurant
category : None
address : 6901 schneider st
city : houston
postalcode : 77093
phone : (713) 221-9000
street : schneider st
house_number : 6901
house : None

16/10 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: (2, address), TfidfN
GramCanopyPredicate: (0.8, phone))
INFO:dedupe.training:(SimplePredicate: (sameSevenCharStartPredicate, n
ame), TfidfNGramCanopyPredicate: (0.8, address))
INFO:dedupe.training:(SimplePredicate: (commonTwoTokens, city), Simple
Predicate: (fingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
```

```
INFO:dedupe.training:(SimplePredicate: (hundredIntegersOddPredicate, n
ame), SimplePredicate: (suffixArray, phone))
name : raj mahal indian restaurant
category : None
address : 2740 w 12th st
city : erie
postalcode : 16505
phone : (814) 838-1055
street : w 12th st
house_number : 2740
house : None

name : rag magal indian restaurant
category : None
address : 5618 peach st
city : erie
postalcode : 16509
phone : (814) 838-1055
street : peach st
house_number : 5618
house : None

17/10 positive, 7/10 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, city), S
implePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredica
te: (wholeFieldPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, St
reetName), SimplePredicate: (sameSevenCharStartPredicate, name))
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, name), Sim
plePredicate: (commonThreeTokens, street))
name : sidwells family restaurant
category : None
address : 500 w pearl st
city : tremont
postalcode : 61568
phone : (309) 925-5300
street : w pearl st
house_number : 500
house : None

name : subway
category : None
address : 600 w pearl st
city : tremont
postalcode : 61568
phone : (309) 925-7600
street : w pearl st
house_number : 600
house : None

17/10 positive, 8/10 negative
```

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

n

name : arby's
category : None
address : 9418 n interstate 20
city : sweetwater
postalcode : 79556
phone : (325) 235-1450
street : n interstate 20
house_number : 9418
house : None

name : love's travel stop
category : None
address : 9418 n interstate 20
city : sweetwater
postalcode : 79556
phone : (325) 235-1240
street : n interstate 20
house_number : 9418
house : None

17/10 positive, 9/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

name : 54th street bar and grill
category : None
address : 5715-5727 rim pass dr
city : san antonio
postalcode : 78257
phone : None
street : rim pass dr
house_number : 5715-5727
house : None

name : 54th street restaurant & drafthouse
category : None
address : 17122 w interstate 10
city : san antonio
postalcode : 78257
phone : (210) 690-5424
street : w interstate 10
house_number : 17122
house : None

17/10 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 [ ]:
```
deduper2 = deduper
```

In [25]:
```
deduper.train(recall=1)
```

```
INFO:rlr.crossvalidation:using cross validation to find optimum alph
a...
INFO:rlr.crossvalidation:optimum alpha: 0.010000, score 0.378272292103
0144
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (firstTokenPredicate, city), Si
mplePredicate: (twoGramFingerprint, phone))
INFO:dedupe.training:(ExistsPredicate: (Exists, house), SimplePredicat
e: (wholeFieldPredicate, name))
INFO:dedupe.training:(PartialPredicate: (commonTwoTokens, address, Str
eetName), SimplePredicate: (sameSevenCharStartPredicate, name))
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, name), Simp
lePredicate: (commonThreeTokens, street))
```

In [26]:
```
deduper.predicates
```

Out[26]:
```
((SimplePredicate: (firstTokenPredicate, city),
  SimplePredicate: (twoGramFingerprint, phone)),
 (ExistsPredicate: (Exists, house),
  SimplePredicate: (wholeFieldPredicate, name)),
 (PartialPredicate: (commonTwoTokens, address, StreetName),
  SimplePredicate: (sameSevenCharStartPredicate, name)),
 (SimplePredicate: (commonThreeTokens, name),
  SimplePredicate: (commonThreeTokens, street)))
```

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

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

In [27]:
```
threshold = deduper.threshold(data_dict, recall_weight=1)
```

```
INFO:dedupe.blocking:10000, 1.4191872 seconds
INFO:dedupe.api:Maximum expected recall and precision
INFO:dedupe.api:recall: 0.997
INFO:dedupe.api:precision: 0.917
INFO:dedupe.api:With threshold: 0.480
```

In [29]:
```
clustered_dupes = deduper.match(data_dict, threshold)
```

```
INFO:dedupe.blocking:10000, 1.9783572 seconds
```

In [30]:
```
print('# duplicate sets', len(clustered_dupes))
```

```
# duplicate sets 2001
```

In [31]:
```
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 [32]:
```
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 [33]:
```
df = pd.read_csv(output_file)
```

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

In [34]:
```
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 [36]:
```
results = pd.read_csv(matches_file)
results['yp-y'] = results.apply(lambda row: f"{row['yellow_pages_id']}-
```

In [ ]:
```
results.columns
```

In [ ]:
```
results['duplicate'][0]
```

In [40]:
```
results = results[results['duplicate'] == 1]#[['yellow_pages_id', 'yelp_
```

In [42]:
```
len(results)
```

Out[42]: 126

In [43]:
```
res_set = set(results['yp-y'].values.tolist())
pred_set = set(predictions['yp-y'].values.tolist())
```

In [44]:
```
tp = len(res_set & pred_set)
fn = len(res_set-pred_set)
fpos = len(pred_set-res_set)

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

```
tp: 20 fp: 2045 fn: 106
```