yelp_predict

May 29, 2021

1 Yelp Rating Prediction

1.1 Business - Data Cleaning

[3]:
```
import csv
import pandas as pd
import string
import seaborn as sns
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from sklearn.pipeline import Pipeline
from sklearn.linear_model import SGDClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
```

[4]:
```
df_biz = pd.read_json('yelp_academic_dataset_business.json', lines=True)
```

[5]:
```
df_biz.head()
```

[5]:

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes
0	6iYb2HFDywm3zjuRg0shjw	Oskar Blues Taproom	921 Pearl St	Boulder	CO	80302	40.017544	-105.283348	4.0	86	1	{'RestaurantsTableService': 'True', 'WiFi': 'u…
1	tCbdrRPZA0oiIYSmHG3J0w	Flying Elephants at PDX	7000 NE Airport Way	Portland	OR	97218	45.588906	-122.593331	4.0	126	1	{'RestaurantsTakeOut': 'True', 'RestaurantsAtt…
2	bvN78flM8NLprQ1a1y5dRg	The Reclaimory	4720 Hawthorne Ave	Portland	OR	97214	45.511907	-122.613693	4.5	13	1	{'BusinessAcceptsCreditCards': 'True', 'Restau…
3	oaepsyvc0J17qwi8cfrOWg	Great Clips	2566 Enterprise Rd	Orange City	FL	32763	28.914482	-81.295979	3.0	8	1	{'RestaurantsPriceRange2': '1', 'BusinessAccep…
4	PE9uqAjdw0E4-8mjGl3wVA	Crossfit Terminus	1046 Memorial Dr SE	Atlanta	GA	30316	33.747027	-84.353424	4.0	14	1	

```
4       1  {'GoodForKids': 'False', 'BusinessParking': '{…

                                          categories  \
0  Gastropubs, Food, Beer Gardens, Restaurants, B…
1  Salad, Soup, Sandwiches, Delis, Restaurants, C…
2  Antiques, Fashion, Used, Vintage & Consignment…
3                        Beauty & Spas, Hair Salons
4  Gyms, Active Life, Interval Training Gyms, Fit…

                                               hours
0  {'Monday': '11:0-23:0', 'Tuesday': '11:0-23:0'…
1  {'Monday': '5:0-18:0', 'Tuesday': '5:0-17:0', …
2  {'Thursday': '11:0-18:0', 'Friday': '11:0-18:0…
3                                               None
4  {'Monday': '16:0-19:0', 'Tuesday': '16:0-19:0'…
```

[6]:
```
business = df_biz[df_biz['is_open']==1]
business.head()
```

[6]:
```
              business_id                      name              address  \
0  6iYb2HFDywm3zjuRg0shjw      Oskar Blues Taproom          921 Pearl St
1  tCbdrRPZA0oiIYSmHG3J0w  Flying Elephants at PDX  7000 NE Airport Way
2  bvN78flM8NLprQ1a1y5dRg           The Reclaimory   4720 Hawthorne Ave
3  oaepsyvc0J17qwi8cfrOWg              Great Clips   2566 Enterprise Rd
4  PE9uqAjdw0E4-8mjGl3wVA        Crossfit Terminus  1046 Memorial Dr SE

          city state postal_code   latitude   longitude  stars  review_count  \
0      Boulder    CO       80302  40.017544 -105.283348    4.0            86
1    Portland    OR       97218  45.588906 -122.593331    4.0           126
2    Portland    OR       97214  45.511907 -122.613693    4.5            13
3  Orange City    FL       32763  28.914482  -81.295979    3.0             8
4     Atlanta    GA       30316  33.747027  -84.353424    4.0            14

   is_open                                         attributes  \
0        1  {'RestaurantsTableService': 'True', 'WiFi': 'u…
1        1  {'RestaurantsTakeOut': 'True', 'RestaurantsAtt…
2        1  {'BusinessAcceptsCreditCards': 'True', 'Restau…
3        1  {'RestaurantsPriceRange2': '1', 'BusinessAccep…
4        1  {'GoodForKids': 'False', 'BusinessParking': '{…

                                          categories  \
0  Gastropubs, Food, Beer Gardens, Restaurants, B…
1  Salad, Soup, Sandwiches, Delis, Restaurants, C…
2  Antiques, Fashion, Used, Vintage & Consignment…
3                        Beauty & Spas, Hair Salons
4  Gyms, Active Life, Interval Training Gyms, Fit…
```

```
                                                  hours
0  {'Monday': '11:0-23:0', 'Tuesday': '11:0-23:0'…
1  {'Monday': '5:0-18:0', 'Tuesday': '5:0-17:0', …
2  {'Thursday': '11:0-18:0', 'Friday': '11:0-18:0…
3                                                 None
4  {'Monday': '16:0-19:0', 'Tuesday': '16:0-19:0'…
```

[7]:
```
business = df_biz.drop( ['hours','is_open','review_count'], axis=1)
business
business['categories']
```

[7]:
```
0         Gastropubs, Food, Beer Gardens, Restaurants, B…
1         Salad, Soup, Sandwiches, Delis, Restaurants, C…
2         Antiques, Fashion, Used, Vintage & Consignment…
3                                 Beauty & Spas, Hair Salons
4         Gyms, Active Life, Interval Training Gyms, Fit…
                               …
160580    Real Estate, Real Estate Services, Home Servic…
160581        Health Markets, Food, Specialty Food, Grocery
160582    Arts & Entertainment, Paint & Sip, Art Classes…
160583                 Cuban, Sandwiches, Restaurants, Cafes
160584    Restaurants, Middle Eastern, Mediterranean, Pe…
Name: categories, Length: 160585, dtype: object
```

[8]:
```
df_rest = business[business['categories'].str.
↪contains('Restaurants',case=False, na=False)]
```

[9]:
```
df_rest.shape
```

[9]:
```
(50763, 11)
```

[10]:
```
df_rest.head()
```

[10]:
```
               business_id                      name              address  \
0   6iYb2HFDywm3zjuRg0shjw      Oskar Blues Taproom          921 Pearl St
1   tCbdrRPZA0oiIYSmHG3J0w  Flying Elephants at PDX  7000 NE Airport Way
5   D4JtQNTI4X3KcbzacDJsMw      Bob Likes Thai Food          3755 Main St
7   jFYIsSb7r1QeESVUnXPHBw          Boxwood Biscuit         740 S High St
12  HPA_qyMEddpAEtFof02ixg      Mr G's Pizza & Subs         474 Lowell St

          city state postal_code   latitude   longitude  stars  \
0     Boulder    CO       80302  40.017544 -105.283348    4.0
1    Portland    OR       97218  45.588906 -122.593331    4.0
5   Vancouver    BC         V5V  49.251342 -123.101333    3.5
7    Columbus    OH       43206  39.947007  -82.997471    4.5
12    Peabody    MA       01960  42.541155  -70.973438    4.0
```

```
                                              attributes  \
0   {'RestaurantsTableService': 'True', 'WiFi': 'u…
1   {'RestaurantsTakeOut': 'True', 'RestaurantsAtt…
5   {'GoodForKids': 'True', 'Alcohol': 'u'none'', …
7                                                  None
12  {'RestaurantsGoodForGroups': 'True', 'HasTV': …

                                              categories
0   Gastropubs, Food, Beer Gardens, Restaurants, B…
1   Salad, Soup, Sandwiches, Delis, Restaurants, C…
5                                  Restaurants, Thai
7                     Breakfast & Brunch, Restaurants
12                           Food, Pizza, Restaurants
```

1.2 Review - Data Cleaning

[11]:
```
review = pd.read_json('yelp_academic_dataset_review.json', lines=True,␣
↪dtype={'review_id':str,'user_id':str,
                    'business_id':str,'stars':int,'date':str,'text':
↪str,'useful':int,'funny':int,'cool':int},
                     chunksize=1000000)
```

[ ]:
```
chunk_list = []
for chunk in review:
    chunk = chunk.drop(['review_id','useful','funny','cool'] , axis=1)
    chunk = chunk.rename(columns={'stars': 'review_stars'})
    chunk_merged = pd.merge(df_rest, chunk, on='business_id', how='inner')
    chunk_list.append(chunk_merged)

df_review = pd.concat(chunk_list, ignore_index=True, join='outer', axis=0)
df_review.head()
```

[1]:
```
df_review.shape
```

```
      ␣
↪---------------------------------------------------------------------------

       NameError                                 Traceback (most recent call␣
↪last)

       <ipython-input-1-374472ebc4e2> in <module>
   ----> 1 df_review.shape


       NameError: name 'df_review' is not defined
```

[83]:
```
df_data = df_review
```

[84]:
```
df_data.shape
```

[84]:
```
(4921128, 15)
```

[85]:
```
df_data.head()
```

[85]:
```
              business_id              name  \
0  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
1  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
2  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
3  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
4  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express

                                    address       city state postal_code  \
0  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL       34746
1  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL       34746
2  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL       34746
3  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL       34746
4  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL       34746

    latitude  longitude  stars  \
0  28.331336 -81.515777    4.0
1  28.331336 -81.515777    4.0
2  28.331336 -81.515777    4.0
3  28.331336 -81.515777    4.0
4  28.331336 -81.515777    4.0

                                        attributes  \
0  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
1  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
2  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
3  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
4  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…

                                        categories                 user_id  \
0  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  T_Zi6O4CUgxg7Rov4RLs0Q
1  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  gD9EpLRQwa-W42G0S11L9A
2  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  BjjFQtAEzU2qWSknC2MAsw
3  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  aTXsmv6NLweGFsxKGl3mbA
4  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  pnwR5WLhh-G99c3yT1Cxsw

   review_stars                                                text  \
0             5  Great customer service and amazing food!! The …
1             1  Do not come here!\n\nDirty, gross and it takes…
2             3  Sushi was meh, service was fast and friendly, …
```

```
3             4  I just saw this place was Eat 24 enabled, so I…
4             5  Little spot in Old Town off of 192. I took Yel…

                  date
0  2017-04-23 02:49:04
1  2017-01-16 02:46:09
2  2018-10-01 17:16:41
3  2015-06-10 19:28:55
4  2017-07-19 17:18:52
```

[86]:
```
df_data = df_data.drop(['postal_code','postal_code','latitude', 'longitude',␣
↪'date'], axis=1)
```

[87]:
```
df_data.head()
```

[87]:
```
              business_id              name  \
0  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
1  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
2  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
3  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express
4  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express

                                    address       city state  stars  \
0  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL    4.0
1  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL    4.0
2  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL    4.0
3  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL    4.0
4  5770 W Irlo Bronson Memorial Hwy, Ste 157  Kissimmee    FL    4.0

                                        attributes  \
0  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
1  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
2  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
3  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…
4  {'RestaurantsTakeOut': 'True', 'RestaurantsTab…

                                        categories                 user_id  \
0  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  T_Zi6O4CUgxg7Rov4RLs0Q
1  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  gD9EpLRQwa-W42G0S11L9A
2  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  BjjFQtAEzU2qWSknC2MAsw
3  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  aTXsmv6NLweGFsxKGl3mbA
4  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  pnwR5WLhh-G99c3yT1Cxsw

   review_stars                                                text
0             5  Great customer service and amazing food!! The …
1             1  Do not come here!\n\nDirty, gross and it takes…
2             3  Sushi was meh, service was fast and friendly, …
```

```
3            4  I just saw this place was Eat 24 enabled, so I…
4            5  Little spot in Old Town off of 192. I took Yel…
```

[88]:
```
df_data_new = df_data[['business_id', 'name', 'stars', 'categories', 'user_id',␣
↪'review_stars', 'text']]
```

[89]:
```
df_data_new.head()
```

[89]:
```
              business_id              name  stars  \
0  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0
1  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0
2  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0
3  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0
4  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0

                                          categories                 user_id  \
0  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  T_Zi6O4CUgxg7Rov4RLs0Q
1  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  gD9EpLRQwa-W42G0S11L9A
2  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  BjjFQtAEzU2qWSknC2MAsw
3  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  aTXsmv6NLweGFsxKGl3mbA
4  Sushi Bars, Restaurants, Food, Bubble Tea, Jap…  pnwR5WLhh-G99c3yT1Cxsw

   review_stars                                               text
0             5  Great customer service and amazing food!! The …
1             1  Do not come here!\n\nDirty, gross and it takes…
2             3  Sushi was meh, service was fast and friendly, …
3             4  I just saw this place was Eat 24 enabled, so I…
4             5  Little spot in Old Town off of 192. I took Yel…
```

[113]:
```
df_explode = df_data_new.assign(categories = df_data_new.categories.str.
↪split(', ')).explode('categories')
df_explode.sample(3)
```

[113]:
```
                     business_id              name  stars   categories  \
4811422  3942kJZKI8gsJOOBdRsneA  Garlic 'n Lemons    4.0  Ethnic Food
4847871  HaGoOH9GzsBZcmPpiEIJvA           Hopstix    4.0    Breweries
2429200  hVL_gLKPoCTMlS5Pux_XWg            Argosy    4.0      Burgers

                        user_id  review_stars  \
4811422  BlJc3xknuvI1WV2hWHKCHw             5
4847871  sLSrDKncV5S9jAW60AarBw             5
2429200  JlGN8RBNEdwFNU8QAxl7OA             5

                                                      text
4811422  Garlic 'n Lemons is a great spot for shawarma…
4847871  I've been to Hopstix twice: once for dinner on…
2429200  The decor is absolutely beautiful! I love all …
```

[114]:
```
print('Total number of categories:', len(df_explode.categories.value_counts()))
print('Top 10 categories:')
df_explode.categories.value_counts()[:10]
```

```
Total number of categories: 708
Top 10 categories:
```

[114]:
```
Restaurants                4921128
Food                       1602939
Nightlife                  1464071
Bars                       1408685
American (New)              919271
American (Traditional)      898530
Breakfast & Brunch          831710
Sandwiches                  638640
Seafood                     495398
Pizza                       443525
Name: categories, dtype: int64
```

1.3 Tokenize

[90]:
```
STOPWORDS = set(STOPWORDS).union(set(['said', 'mr', 'mrs']))
```

[91]:
```
def tokenize(text):
    return [token for token in simple_preprocess(text) if token not in␣
 ↪STOPWORDS]
```

[116]:
```
df_explode['categories'].value_counts().index
```

[116]:
```
Index(['Restaurants', 'Food', 'Nightlife', 'Bars', 'American (New)',
       'American (Traditional)', 'Breakfast & Brunch', 'Sandwiches', 'Seafood',
       'Pizza',
       …
       'Concept Shops', 'Naturopathic/Holistic',
       'Television Service Providers', 'Embroidery & Crochet',
       'Home Organization', 'Laser Hair Removal', 'Elementary Schools',
       'Baby Gear & Furniture', 'Travel Agents', 'Grilling Equipment'],
      dtype='object', length=708)
```

[117]:
```
rest_yelp = df_explode[df_explode['categories']=='Restaurants']
print(rest_yelp.shape)
rest_yelp.head()
```

```
(4921128, 7)
```

[117]:

	business_id	name	stars	categories	\
0	PBPodbdtLyuQ-sNgOWwkVw	Mr Sushi Express	4.0	Restaurants	

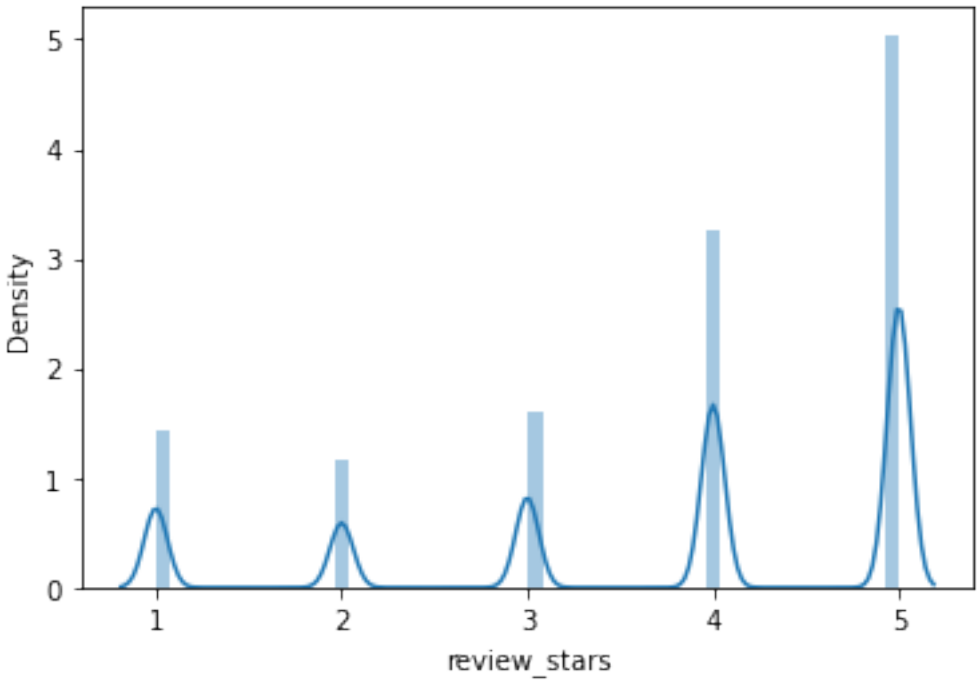
```
1  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0  Restaurants
2  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0  Restaurants
3  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0  Restaurants
4  PBPodbdtLyuQ-sNgOWwkVw  Mr Sushi Express    4.0  Restaurants

                  user_id  review_stars  \
0  T_Zi6O4CUgxg7Rov4RLs0Q             5
1  gD9EpLRQwa-W42G0S11L9A             1
2  BjjFQtAEzU2qWSknC2MAsw             3
3  aTXsmv6NLweGFsxKGl3mbA             4
4  pnwR5WLhh-G99c3yT1Cxsw             5

                                                text
0  Great customer service and amazing food!! The …
1  Do not come here!\n\nDirty, gross and it takes…
2  Sushi was meh, service was fast and friendly, …
3  I just saw this place was Eat 24 enabled, so I…
4  Little spot in Old Town off of 192. I took Yel…
```

[118]:
```
sns.distplot(rest_yelp['review_stars']);
```

```
C:\Users\palla\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```



[ ]:
```
rest_yelp['tokens'] = rest_yelp['text'].apply(tokenize)
rest_yelp['tokens'].head()
```

1.4 Classification

[97]:
```
SDC = SGDClassifier()
RFC = RandomForestClassifier()
VECT = TfidfVectorizer(stop_words='english', ngram_range=(1,1))
```

[99]:
```
pipe = Pipeline([('vect', VECT), ('rfc', RFC)])
```

[100]:
```
pipe.fit(rest_yelp['text'], rest_yelp['review_stars'])
```

[100]:
```
Pipeline(steps=[('vect', TfidfVectorizer(stop_words='english')),
                ('rfc', RandomForestClassifier())])
```

[112]:
```
pipe.predict(["Food is good"])[0]
```

[112]:
```
4
```

[ ]:
```

```