Amazon Fine Food Reviews Analysis

Data Source: https://www.kaggle.com/snap/amazon-fine-food-reviews

EDA: https://nycdatascience.com/blog/student-works/amazon-fine-foods-visualization/

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews: 568,454
Number of users: 256,059
Number of products: 74,258
Timespan: Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

Attribute Information:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

Objective:

Given a review, determine whether the review is positive (rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?

[Ans] We could use Score/Rating. A rating of 4 or 5 can be cosnidered as a positive review. A rating of 1 or 2 can be considered as negative one. A review of rating 3 is considered nuetral and such reviews are ignored from our analysis. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

In [65]:

```python
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remoun
t, call drive.mount("/content/drive", force_remount=True).
```

[1]. Reading Data

[1.1] Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

In order to load the data, We have used the SQLITE dataset as it is easier to query the data and visualise the data efficiently.

Here as we only want to get the global sentiment of the recommendations (positive or negative), we will purposefully ignore all Scores equal to 3. If the score is above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

In [0]:

```python
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

```
import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

In [67]:

```
# using SQLite Table to read data.
con = sqlite3.connect('/content/drive/My Drive/Colab Notebooks/databas
e.sqlite')

# filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 50
```

```
0000 data points
# you can change the number to any other number based on your computing
 power

# filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Sco
re != 3 LIMIT 500000""", con)
# for tsne assignment you can take 5k data points

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score
 != 3 LIMIT 100000""", con)

# Give reviews with Score>3 a positive rating(1), and reviews with a sc
ore<3 a negative rating(0).
def partition(x):
    if x < 3:
        return 0
    return 1

#changing reviews with score less than 3 to be positive and vice-versa
actualScore = filtered_data['Score']
positiveNegative = actualScore.map(partition)
filtered_data['Score'] = positiveNegative
print("Number of data points in our data", filtered_data.shape)
filtered_data.head(3)
```

```
Number of data points in our data (100000, 10)
```

Out[67]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1

In [0]:

```
display = pd.read_sql_query("""
SELECT UserId, ProductId, ProfileName, Time, Score, Text, COUNT(*)
FROM Reviews
GROUP BY UserId
HAVING COUNT(*)>1
""", con)
```

In [69]:

```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[69]:

	UserId	ProductId	ProfileName	Time	Score	Text	COU

	UserId	ProductId	ProfileName	Time	Score	Text	COU
0	#oc-R115TNMSPFT9I7	B007Y59HVM	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ET0	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
2	#oc-R11DNU2NBKQ23Z	B007Y59HVM	Kim Cieszykowski	1348531200	1	This coffee is horrible and unfortunately not ...	2
3	#oc-R11O5J5ZVQE25C	B005HG9ET0	Penguin Chick	1346889600	5	This will be the bottle that you grab from the...	3
4	#oc-R12KPBODL2B5ZD	B007OSBE1U	Christopher P. Presta	1348617600	1	I didnt like this coffee. Instead of telling y...	2

In [70]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[70]:

	UserId	ProductId	ProfileName	Time	Score	Text	(

	UserId	ProductId	ProfileName	Time	Score	Text	(
80638	AZY10LLTJ71NX	B006P7E5ZI	undertheshrine "undertheshrine"	1334707200	5	I was recommended to try green tea extract to ...	5

```
In [71]: display['COUNT(*)'].sum()
```

```
Out[71]: 393063
```

[2] Exploratory Data Analysis

[2.1] Data Cleaning: Deduplication

It is observed (as shown in the table below) that the reviews data had many duplicate entries. Hence it was necessary to remove duplicates in order to get unbiased results for the analysis of the data. Following is an example:

```
In [72]: display= pd.read_sql_query("""
         SELECT *
         FROM Reviews
         WHERE Score != 3 AND UserId="AR5J8UI46CURR"
         ORDER BY ProductID
         """, con)
         display.head()
```

Out[72]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfuln

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfuln
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2

As it can be seen above that same user has multiple reviews with same values for HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary and Text and on doing analysis it was found that

ProductId=B000HDOPZG was Loacker Quadratini Vanilla Wafer Cookies, 8.82-Ounce Packages (Pack of 8)

ProductId=B000HDL1RQ was Loacker Quadratini Lemon Wafer Cookies, 8.82-Ounce Packages (Pack of 8) and so on

It was inferred after analysis that reviews with same parameters other than ProductId belonged to the same product just having different flavour or quantity. Hence in order to reduce redundancy it was decided to eliminate the rows having same parameters.

The method used for the same was that we first sort the data according to ProductId and then just keep the first similar product review and delelte the others. for eg. in the above just the review for ProductId=B000HDL1RQ remains. This method ensures that there is only one representative for each product and deduplication without sorting would lead to possibility of different representatives still existing for the same product.

In [0]:

```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')
```

In [74]:

```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inplace=False)
final.shape
```

Out[74]:

```
(87775, 10)
```

In [75]:

```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[75]:

```
87.775
```

Observation:- It was also seen that in two rows given below the value of HelpfulnessNumerator is greater than HelpfulnessDenominator which is not practically possible hence these two rows too are removed from calcualtions

In [76]:

```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)

display.head()
```

Out[76]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfuln
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2

In [0]:

```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
```

In [78]:

```
#Before starting the next phase of preprocessing lets see the number of
 entries left
print(final.shape)
```

```
#How many positive and negative reviews are present in our dataset?
final['Score'].value_counts()
```

```
(87773, 10)
```

Out[78]:

```
1    73592
0    14181
Name: Score, dtype: int64
```

[3] Preprocessing

[3.1]. Preprocessing Review Text

Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags
2. Remove any punctuations or limited set of special characters like , or . or # etc.
3. Check if the word is made up of english letters and is not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)
5. Convert the word to lowercase
6. Remove Stopwords
7. Finally Snowball Stemming the word (it was obsereved to be better than Porter Stemming)

After which we collect the words used to describe positive and negative reviews

In [0]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [0]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'no
t'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in
 the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'o
urs', 'ourselves', 'you', "you're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselve
s', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'it
s', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'th
is', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'h
ave', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between',
'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',
'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'h
ow', 'all', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 's
o', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should',
"should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
 "shouldn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

In [81]:

```
# Combining all the above stundents
from tqdm import tqdm
from bs4 import BeautifulSoup
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(final['Text'].values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = decontracted(sentance)
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower
() not in stopwords)
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 87773/87773 [00:33<00:00, 2605.20it/s]
```

[4] Featurization

In [82]:

```
#here preprocessed_review is my X and final['Score'] is my Y
print(len(preprocessed_reviews))
print(len(final['Score']))
X=preprocessed_reviews
```

```
Y=final['Score']
#if both are of same lenght then proceed....
```

```
87773
87773
```

In [0]:

```
#here i am performing splittig operation as train test and cv...
from sklearn.model_selection import train_test_split

# X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=
0.33, shuffle=Flase)# this is for time series split
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3
3) # this is random splitting
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_
size=0.33) # this is random splitting
```

[4.1] BAG OF WORDS

In [84]:

```
#BoW
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
vectorizer.fit(X_train) # fitting on train data ,we cant perform fit on
 test or cv

# we use the fitted CountVectorizer to convert the text to vector
X_train_bow = vectorizer.transform(X_train)
X_cv_bow = vectorizer.transform(X_cv)
X_test_bow = vectorizer.transform(X_test)
print("After vectorizations")
print(X_train_bow.shape, y_train.shape)
print(X_cv_bow.shape, y_cv.shape)
print(X_test_bow.shape, y_test.shape)
print("="*100)
#you can also check X_train_bow is of sparse matrix type or not
#below is code for that
print(type(X_train_bow))
#displaying number of unique words in each of splitted dataset
```

```
print("the number of unique words in train: ", X_train_bow.get_shape()[
1])
print("the number of unique words in cv: ", X_cv_bow.get_shape()[1])
print("the number of unique words in test: ", X_test_bow.get_shape()[1
])
```

```
After vectorizations
(39400, 37477) (39400,)
(19407, 37477) (19407,)
(28966, 37477) (28966,)
=======================================================================
=============================
<class 'scipy.sparse.csr.csr_matrix'>
the number of unique words in train:  37477
the number of unique words in cv:  37477
the number of unique words in test:  37477
```

[4.3] TF-IDF

In [85]:

```
#below code for converting to tfidf
#i refered sample solution to write this code
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(X_train)
print("some sample features(unique words in the corpus)",tf_idf_vect.ge
t_feature_names()[0:10])
print('='*50)

X_train_tf_idf = tf_idf_vect.transform(X_train)
X_test_tf_idf = tf_idf_vect.transform(X_test)
X_cv_tf_idf = tf_idf_vect.transform(X_cv)
print("the type of count vectorizer ",type(X_train_tf_idf))
print("the shape of out text TFIDF vectorizer ",X_train_tf_idf.get_shap
e())
print("the number of unique words including both unigrams and bigrams "
, X_train_tf_idf.get_shape()[1])
```

```
some sample features(unique words in the corpus) ['ability', 'able', 'a
ble buy', 'able drink', 'able eat', 'able enjoy', 'able find', 'able fi
nish', 'able get', 'able give']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (39400, 23364)
the number of unique words including both unigrams and bigrams  23364
```

[4.4] Word2Vec

In [86]:

```
#in average w2v the output is of list form and here we write same code
 of all train ,test and cv
#this code is for train data:
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance_train=[]
for sentance in X_train:
    list_of_sentance_train.append(sentance.split())


#training word2vect model
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
# this line of code trains your w2v model on the give list of sentances
w2v_model=Word2Vec(list_of_sentance_train,min_count=5,size=50, workers=
4)
w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])

#this is the actuall code to convert word2vect to avg w2v:
from tqdm import tqdm
import numpy as np
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train = []; # the avg-w2v for each sentence/review is stor
ed in this list
for sent in tqdm(list_of_sentance_train): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
```

```
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_train.append(sent_vec)
sent_vectors_train = np.array(sent_vectors_train)
print(sent_vectors_train.shape)
print(sent_vectors_train[0])
```

```
  0%|          | 123/39400 [00:00<00:31, 1229.35it/s]
```

```
number of words that occured minimum 5 times  11969
sample words  ['ordered', 'pack', 'listed', 'jars', 'warned', 'not', 't
ried', 'yet', 'cups', 'work', 'great', 'keurig', 'year', 'use', 'refill
able', 'filter', 'cup', 'tedious', 'refill', 'every', 'time', 'want',
'coffee', 'especially', 'easy', 'perfect', 'mornings', 'quick', 'lids',
'little', 'tough', 'get', 'sometimes', 'stuck', 'top', 'machines', 'occ
asionally', 'still', 'much', 'less', 'using', 'ready', 'beginning', 'we
ek', 'go', 'morning', 'affordable', 'name', 'brand', 'product']
```

```
100%|██████████| 39400/39400 [01:06<00:00, 594.96it/s]
```

```
(39400, 50)
[-0.26815421 -1.26602095  0.02784101 -0.17142418 -0.4073684  -0.0222731
1
 -0.62498159 -0.17112265 -0.68833753 -0.23359246 -0.61788673  0.2947438
6
 -0.0444341   0.06630311 -0.66914113 -0.3946488   0.01451979  0.5682101
7
 -0.38811492  0.37547644  0.38379669  0.56460062  1.1999604  -0.0526460
7
  0.05996043 -0.62110074 -0.67313385  0.65212659  0.0936182   0.0153474
8
 -0.03616093 -0.66740643  0.42378843  0.62667941 -0.26277018  0.6086988
7
```

```
 -0.30492323  0.01347144  0.60087399 -0.47945883  0.54611267 -0.6290486
9
  0.06553459  0.21326104  0.39729742 -0.17362833 -0.27381471 -0.2720371
4
  0.3702151  -0.61415459]
```

In [87]:

```
#this code is for test data:
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())


#training word2vect model
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
# this line of code trains your w2v model on the give list of sentances
#i made below two statement as comment to avoid data leakage problem
#w2v_model=Word2Vec(list_of_sentance_test,min_count=5,size=50, workers=
4)
#w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])

#this is the actuall code to convert word2vect to avg w2v:
from tqdm import tqdm
import numpy as np
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_test = []; # the avg-w2v for each sentence/review is store
d in this list
for sent in tqdm(list_of_sentance_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_test.append(sent_vec)
sent_vectors_test = np.array(sent_vectors_test)
print(sent_vectors_test.shape)
print(sent_vectors_test[0])
```

```
  0%|          | 0/28966 [00:00<?, ?it/s]
```

```
number of words that occured minimum 5 times  11969
sample words  ['ordered', 'pack', 'listed', 'jars', 'warned', 'not', 't
ried', 'yet', 'cups', 'work', 'great', 'keurig', 'year', 'use', 'refill
able', 'filter', 'cup', 'tedious', 'refill', 'every', 'time', 'want',
'coffee', 'especially', 'easy', 'perfect', 'mornings', 'quick', 'lids',
'little', 'tough', 'get', 'sometimes', 'stuck', 'top', 'machines', 'occ
asionally', 'still', 'much', 'less', 'using', 'ready', 'beginning', 'we
ek', 'go', 'morning', 'affordable', 'name', 'brand', 'product']
```

```
100%|██████████| 28966/28966 [00:47<00:00, 603.79it/s]
```

```
(28966, 50)
[-0.54016105 -0.58607882  0.12218405 -0.79704337  0.92676735  0.0786780
7
 -0.29797174 -0.0714597   0.17487865  0.90715239 -0.75320925  0.3461828
9
 -0.78552649 -0.41352536  0.71847734 -0.00460726 -0.18247823  0.6939709
8
 -0.5188437  -0.46715224  0.235145    0.10868724  0.21110859  0.5606619
9
 -0.43447117 -1.38886678 -0.52228969  1.26074518 -0.20065742 -0.1996445
9
 -0.19153948 -0.06980445  0.4120937  -0.27042362 -0.26295671  0.3451747
6
  0.41006815 -0.47164078 -1.00181173 -0.14287258  0.27088602 -0.4626434
3
  0.11127189  0.9781075   0.32921709  0.04386244 -0.13472525 -0.0771562

7
,
 -0.41462189 -0.27651347]
```

In [88]:

```
#this code is for cv data:
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance_cv=[]
for sentance in X_cv:
    list_of_sentance_cv.append(sentance.split())


#training word2vect model
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
# this line of code trains your w2v model on the give list of sentances
#w2v_model=Word2Vec(list_of_sentance_cv,min_count=5,size=50, workers=4)
#w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])

#this is the actuall code to convert word2vect to avg w2v:
from tqdm import tqdm
import numpy as np
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sentance_cv): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_cv.append(sent_vec)
sent_vectors_cv= np.array(sent_vectors_cv)
print(sent_vectors_cv.shape)
print(sent_vectors_cv[0])
```

```
  0%|          | 0/19407 [00:00<?, ?it/s]
```

```
number of words that occured minimum 5 times  11969
sample words  ['ordered', 'pack', 'listed', 'jars', 'warned', 'not', 't
ried', 'yet', 'cups', 'work', 'great', 'keurig', 'year', 'use', 'refill
able', 'filter', 'cup', 'tedious', 'refill', 'every', 'time', 'want',
'coffee', 'especially', 'easy', 'perfect', 'mornings', 'quick', 'lids',
'little', 'tough', 'get', 'sometimes', 'stuck', 'top', 'machines', 'occ
asionally', 'still', 'much', 'less', 'using', 'ready', 'beginning', 'we
ek', 'go', 'morning', 'affordable', 'name', 'brand', 'product']
```

```
100%|██████████| 19407/19407 [00:32<00:00, 604.92it/s]
```

```
(19407, 50)
[-0.20582975 -0.82129857  0.41615613 -0.13924882 -0.28027636  0.5673650
4
  0.14303225 -0.30543602 -0.52268806  0.03903529 -0.51231233  0.0965161
7
 -0.3542358  -0.09068496 -0.22438389 -0.00415197  0.30980219  0.4611217
7
 -0.0279458   0.03710915  0.34918435 -0.01753645  0.49749179 -0.1662696
5
  0.07541785 -0.78386299 -0.460214    0.6840048  -0.01014782  0.0641961
7
 -0.02441004 -0.44059637  0.02026307  0.25479365 -0.10852001 -0.0918046
5
 -0.14147178 -0.32587181  0.47599197 -0.20584258  0.88359152 -0.5553761
2
  0.31334024  0.5970766   0.55610685 -0.05162499 -0.22576657 -0.2780069
4
  0.0462651  -0.28775515]
```

[4.4.1] TFIDF weighted W2v

In [89]:

```
#this is for train data
i=0
list_of_sentance_train=[]
for sentance in X_train:
    list_of_sentance_train.append(sentance.split())


# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = model.fit_transform(X_train)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))



# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
ll_val = tfidf

tfidf_sent_vectors_train = []; # the tfidf-w2v for each sentence/review
 is stored in this list
row=0;
for sent in tqdm(list_of_sentance_train): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/r
eview
    for word in sent: # for each word in a review/sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
#             tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors_train.append(sent_vec)
    row += 1
tfidf_sent_vectors_train= np.array(sent_vectors_train)
print(tfidf_sent_vectors_train.shape)
print(tfidf_sent_vectors_train[0])
```

```
100%|██████████| 39400/39400 [12:39<00:00, 51.88it/s]
```

```
(39400, 50)
[-0.26815421 -1.26602095  0.02784101 -0.17142418 -0.4073684  -0.0222731
1
 -0.62498159 -0.17112265 -0.68833753 -0.23359246 -0.61788673  0.2947438
6
 -0.0444341   0.06630311 -0.66914113 -0.3946488   0.01451979  0.5682101
7
 -0.38811492  0.37547644  0.38379669  0.56460062  1.1999604  -0.0526460
7
  0.05996043 -0.62110074 -0.67313385  0.65212659  0.0936182   0.0153474
8
 -0.03616093 -0.66740643  0.42378843  0.62667941 -0.26277018  0.6086988
7
 -0.30492323  0.01347144  0.60087399 -0.47945883  0.54611267 -0.6290486
9
  0.06553459  0.21326104  0.39729742 -0.17362833 -0.27381471 -0.2720371
4
  0.3702151  -0.61415459]
```

In [90]:

```
#this is for test data
i=0
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())


# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
#model = TfidfVectorizer()
tf_idf_matrix = model.transform(X_test)
# we are converting a dictionary with word as a key, and the idf as a v
```

```
alue
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))



# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
ll_val = tfidf

tfidf_sent_vectors_test = []; # the tfidf-w2v for each sentence/review
 is stored in this list
row=0;
for sent in tqdm(list_of_sentance_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/r
eview
    for word in sent: # for each word in a review/sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
#              tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors_test.append(sent_vec)
    row += 1
tfidf_sent_vectors_test= np.array(sent_vectors_test)
print(tfidf_sent_vectors_test.shape)
print(tfidf_sent_vectors_test[0])
```

```
100%|██████████| 28966/28966 [09:37<00:00, 50.20it/s]
```

```
(28966, 50)
[-0.54016105 -0.58607882  0.12218405 -0.79704337  0.92676735  0.0786780
7
```

```
 -0.29797174 -0.0714597   0.17487865  0.90715239 -0.75320925  0.3461828
9
 -0.78552649 -0.41352536  0.71847734 -0.00460726 -0.18247823  0.6939709
8
 -0.5188437  -0.46715224  0.235145    0.10868724  0.21110859  0.5606619
9
 -0.43447117 -1.38886678 -0.52228969  1.26074518 -0.20065742 -0.1996445
9
 -0.19153948 -0.06980445  0.4120937  -0.27042362 -0.26295671  0.3451747
6
  0.41006815 -0.47164078 -1.00181173 -0.14287258  0.27088602 -0.4626434
3
  0.11127189  0.9781075   0.32921709  0.04386244 -0.13472525 -0.0771562
7
 -0.41462189 -0.27651347]
```

In [91]:

```
#this is for cv data
i=0
list_of_sentance_cv=[]
for sentance in X_cv:
    list_of_sentance_cv.append(sentance.split())


# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
#model = TfidfVectorizer()
tf_idf_matrix = model.transform(X_cv)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))



# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
ll_val = tfidf

tfidf_sent_vectors_cv = []; # the tfidf-w2v for each sentence/review is
```

```
 stored in this list
row=0;
for sent in tqdm(list_of_sentance_cv): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/r
eview
    for word in sent: # for each word in a review/sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
#             tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors_cv.append(sent_vec)
    row += 1
tfidf_sent_vectors_cv= np.array(sent_vectors_cv)
print(tfidf_sent_vectors_cv.shape)
print(tfidf_sent_vectors_cv[0])
```

```
100%|██████████| 19407/19407 [06:34<00:00, 49.24it/s]
```

```
(19407, 50)
[-0.20582975 -0.82129857  0.41615613 -0.13924882 -0.28027636  0.5673650
4
  0.14303225 -0.30543602 -0.52268806  0.03903529 -0.51231233  0.0965161
7
 -0.3542358  -0.09068496 -0.22438389 -0.00415197  0.30980219  0.4611217
7
 -0.0279458   0.03710915  0.34918435 -0.01753645  0.49749179 -0.1662696
5
  0.07541785 -0.78386299 -0.460214    0.6840048  -0.01014782  0.0641961
7
 -0.02441004 -0.44059637  0.02026307  0.25479365 -0.10852001 -0.0918046
5
 -0.14147178 -0.32587181  0.47599197 -0.20584258  0.88359152 -0.5553761
2
  0.31334024  0.5970766   0.55610685 -0.05162499 -0.22576657 -0.2780069
4
  0.0462651  -0.28775515]
```

[5] Assignment 8: Decision Trees

1. **Apply Decision Trees on these feature sets**

   - SET 1:Review text, preprocessed one converted into vectors using (BOW)
   - SET 2:Review text, preprocessed one converted into vectors using (TFIDF)
   - SET 3:Review text, preprocessed one converted into vectors using (AVG W2v)
   - SET 4:Review text, preprocessed one converted into vectors using (TFIDF W2v)

2. **The hyper paramter tuning (best `depth` in range [1, 5, 10, 50, 100, 500, 100], and the best `min_samples_split` in range [5, 10, 100, 500])**

   - Find the best hyper parameter which will give the maximum AUC value
   - Find the best hyper paramter using k-fold cross validation or simple cross validation data
   - Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this task of hyperparameter tuning

3. **Graphviz**

   - Visualize your decision tree with Graphviz. It helps you to understand how a decision is being made, given a new vector.
   - Since feature names are not obtained from word2vec related models, visualize only BOW & TFIDF decision trees using Graphviz
   - Make sure to print the words in each node of the decision tree instead of printing its index.

- Just for visualization purpose, limit max_depth to 2 or 3 and either embed the generated images of graphviz in your notebook, or directly upload them as .png files.

4. **Feature importance**

- Find the top 20 important features from both feature sets Set 1 and Set 2 using `feature_importances_` method of Decision Tree Classifier and print their corresponding feature names

5. **Feature engineering**

- To increase the performance of your model, you can also experiment with with feature engineering like :
  - Taking length of reviews as another feature.
  - Considering some features from review summary as well.

6. **Representation of results**

- You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure.
- Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
- Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps.

7. **Conclusion**

- You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link

Note: Data Leakage

1. There will be an issue of data-leakage if you vectorize the entire data and then split it into train/cv/test.
2. To avoid the issue of data-leakag, make sure to split your data first and then vectorize it.
3. While vectorizing your data, apply the method fit_transform() on you train data, and apply the method transform() on cv/test data.
4. For more details please go through this link.

Applying Decision Trees

[5.1] Applying Decision Trees on BOW, SET 1

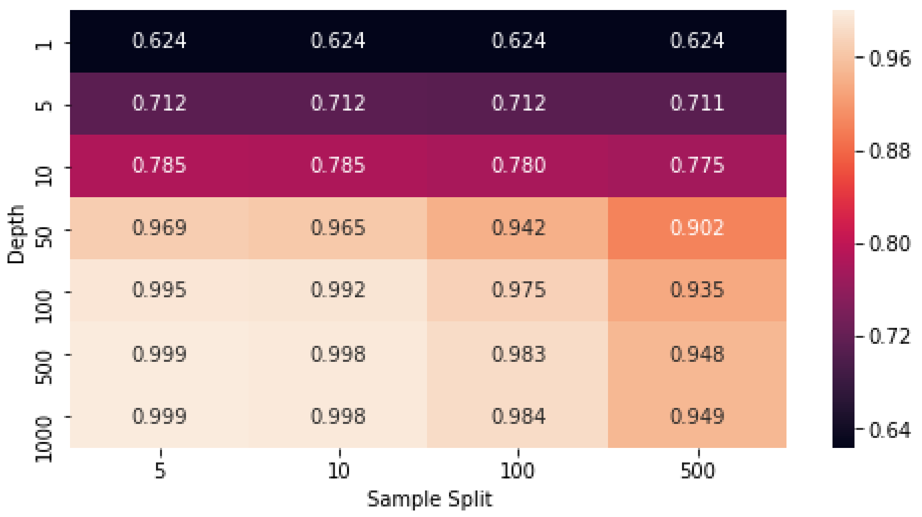
auc plot for train and cv (cross validation:)(bow)

In [0]:

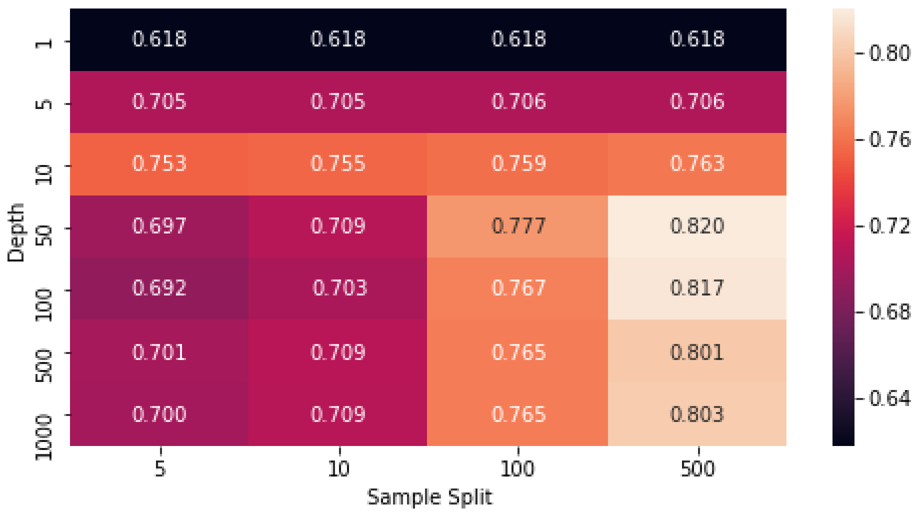
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score
depth = [1, 5, 10, 50, 100, 500, 1000]
min_samples_split = [5, 10, 100, 500]
train_auc = []
cv_auc = []
for d in depth:
  for m in min_samples_split:
    clf = DecisionTreeClassifier(max_depth = d, min_samples_split = m)
    clf.fit(X_train_bow, y_train);
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probab
ility estimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(X_train_bow)[:,1]
    y_cv_pred =  clf.predict_proba(X_cv_bow)[:,1]
    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))
```

```
print("AUC SCORES")
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
print("*"*20, "train data", "*"*20)
train_auc= np.array(train_auc)
train_auc= train_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(train_auc,annot=True, fmt=".3f", xticklabels=min_samples_sp
lit,yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
print("*"*20, "cv data", "*"*20)
cv_auc= np.array(cv_auc)
cv_auc= cv_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(cv_auc,annot=True,fmt=".3f", xticklabels=min_samples_split,
yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
```

```
AUC SCORES
******************** train data ********************
```



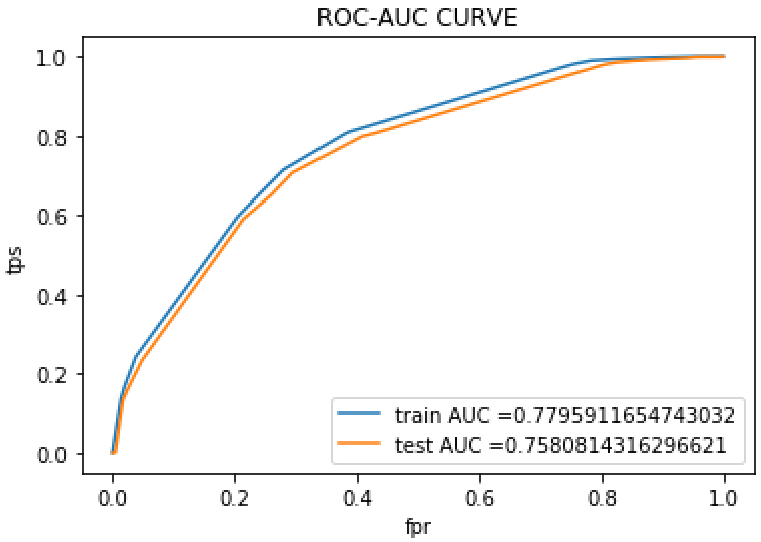
```
******************** cv data ********************
```



In [115]:

```python
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth = 10, min_samples_split = 100)
clf.fit(X_train_bow,y_train)
train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(X_train_bow)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test_bow)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tps")
plt.title("ROC-AUC CURVE")
plt.show()
```

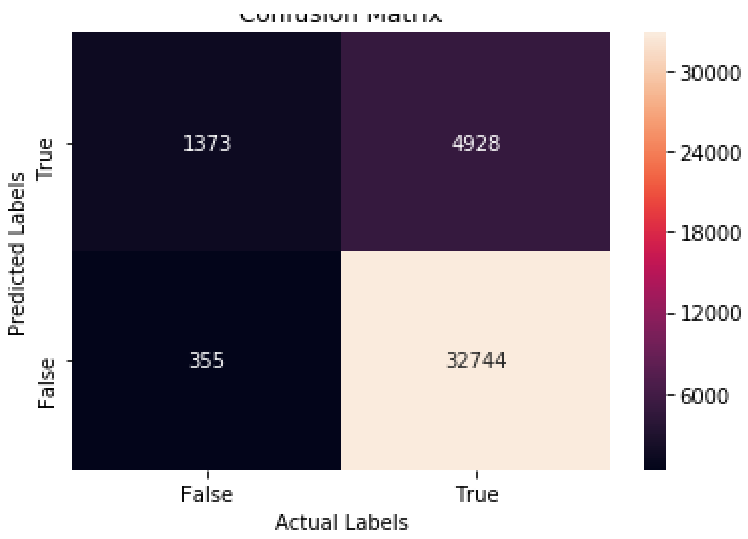


In [117]:

```
#for seaborn confusion matrix :https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-confusion-matrix-with-labels
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
ax= plt.subplot()
arr1=confusion_matrix(y_train, clf.predict(X_train_bow))
df_1= pd.DataFrame(arr1, range(2),range(2))
plt.figure(figsize = (5,2))
sn.heatmap(df_1, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Train confusion matrix
```

Confusion Matrix

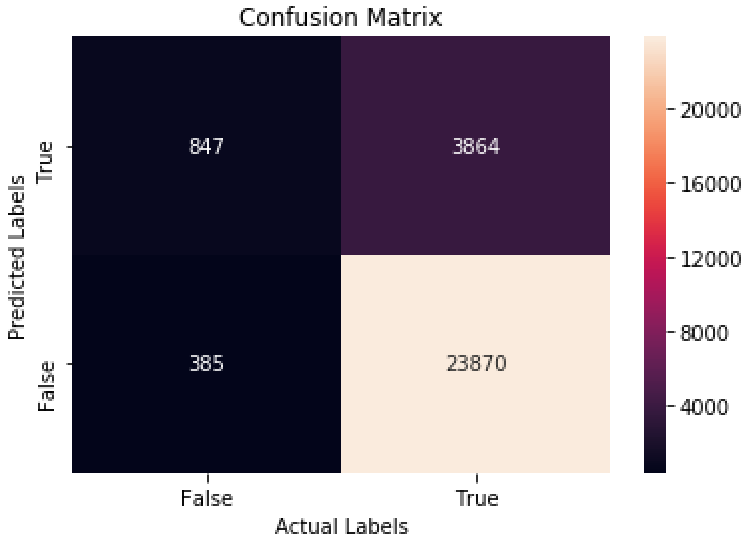


```
<Figure size 360x144 with 0 Axes>
```

In [118]:

```
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-con
fusion-matrix-with-labels
print("Test confusion matrix")
ax= plt.subplot()
arr2=confusion_matrix(y_test, clf.predict(X_test_bow))
df_2= pd.DataFrame(arr2, range(2),range(2))
plt.figure(figsize = (5,2))
sn.heatmap(df_2, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Test confusion matrix
```



```
<Figure size 360x144 with 0 Axes>
```

[5.1.1] Top 20 important features from SET 1

In [133]:

```
#i used below link as reference for printing feature importance with it
s name
#https://stackoverflow.com/questions/11116697/how-to-get-most-informati
ve-features-for-scikit-learn-classifiers
feature_names = vectorizer.get_feature_names()
coefs = sorted(zip(clf.feature_importances_, feature_names))
top = coefs[:-(20 + 1):-1]
print("feature_importances\tfeatures")
for (coef, feat) in top:
  print("%f\t\t%s" % (coef, feat))
```

```
feature_importances     features
0.165375                not
0.105634                great
0.093442                disappointed
0.085036                worst
0.069556                money
```

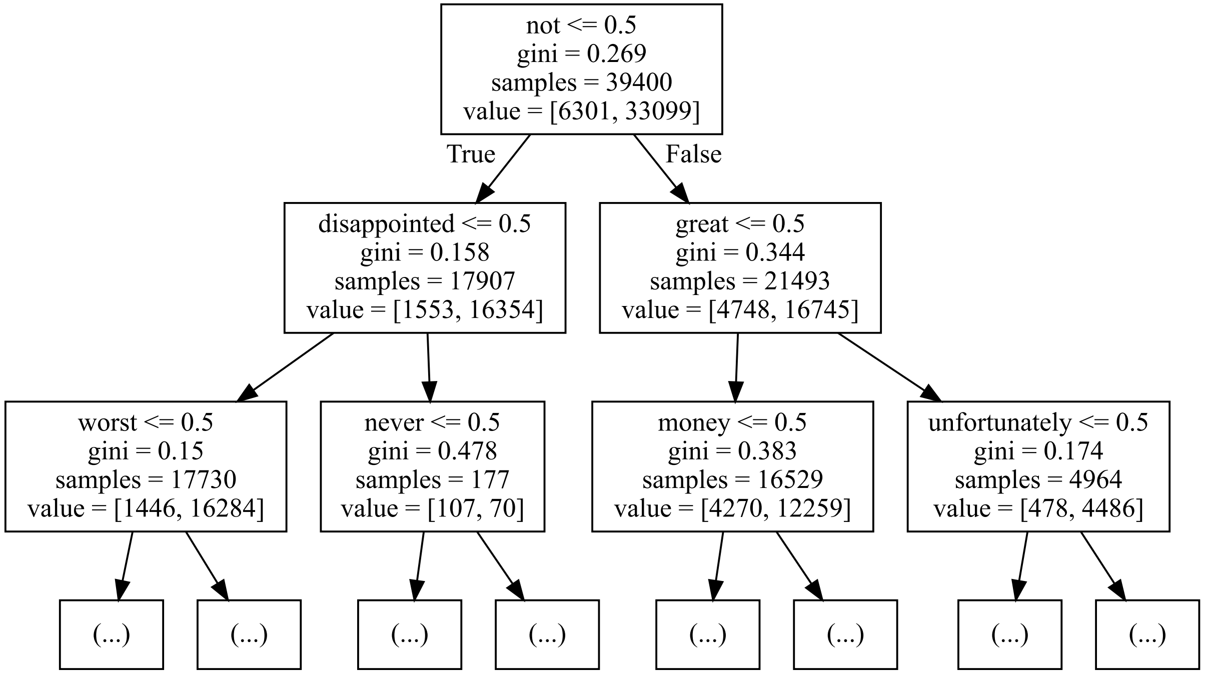
```
0.060998                horrible
0.041766                love
0.038867                delicious
0.035666                terrible
0.033254                good
0.029734                awful
0.021067                waste
0.020906                threw
0.017783                bad
0.016185                disappointing
0.013838                refund
0.010439                best
0.007571                try
0.007351                unfortunately
0.007202                bit
```

[5.1.2] Graphviz visualization of Decision Tree on BOW, SET 1

In [135]:

```
from sklearn import tree
from graphviz import Source
import graphviz
feature = vectorizer.get_feature_names()
Source(tree.export_graphviz(clf, out_file = None, feature_names = featu
re,max_depth=2))
```

Out[135]:



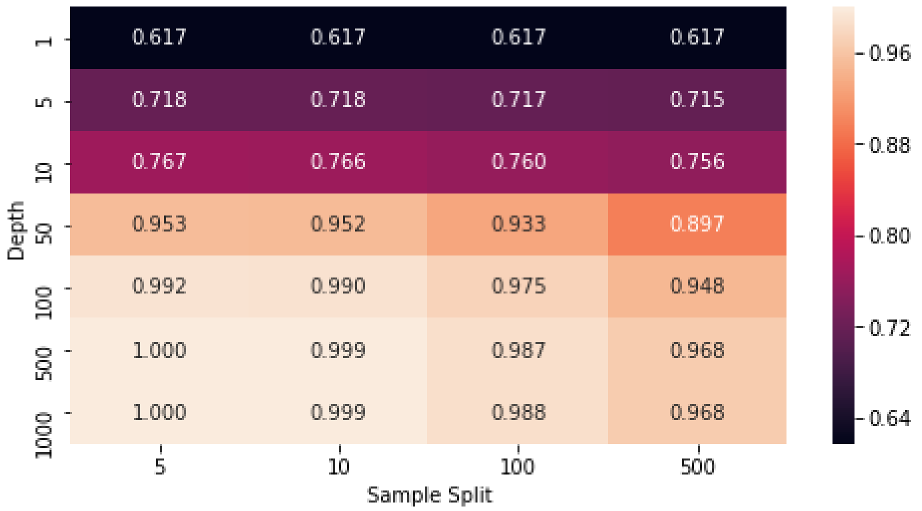
[5.2] Applying Decision Trees on TFIDF, SET 2

In [136]:

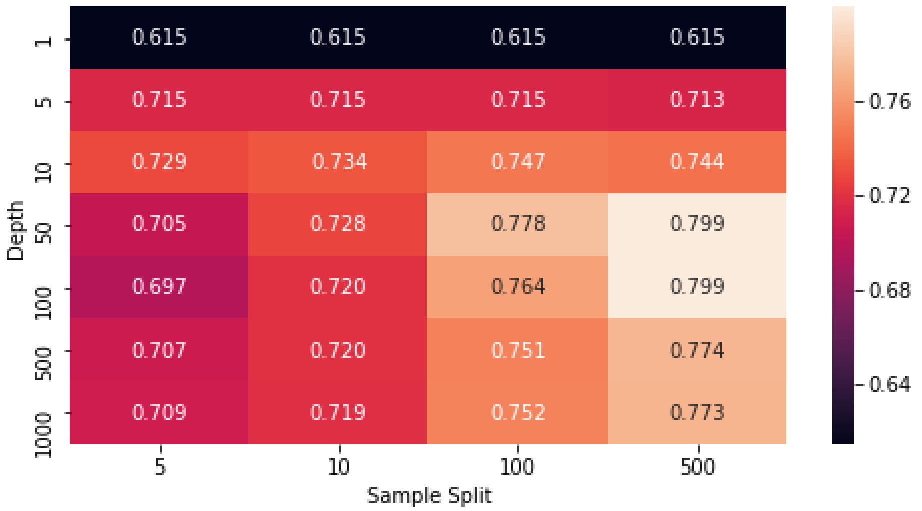
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score
depth = [1, 5, 10, 50, 100, 500, 1000]
min_samples_split = [5, 10, 100, 500]
train_auc = []
cv_auc = []
for d in depth:
  for m in min_samples_split:
    clf = DecisionTreeClassifier(max_depth = d, min_samples_split = m)
    clf.fit(X_train_tf_idf, y_train);
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probab
ility estimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(X_train_tf_idf)[:,1]
    y_cv_pred =  clf.predict_proba(X_cv_tf_idf)[:,1]
    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

print("AUC SCORES")
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
print("*"*20, "train data", "*"*20)
train_auc= np.array(train_auc)
train_auc= train_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(train_auc,annot=True, fmt=".3f", xticklabels=min_samples_sp
lit,yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
print("*"*20, "cv data", "*"*20)
cv_auc= np.array(cv_auc)
cv_auc= cv_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(cv_auc,annot=True,fmt=".3f", xticklabels=min_samples_split,
yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
```

```
AUC SCORES
******************** train data ********************
```



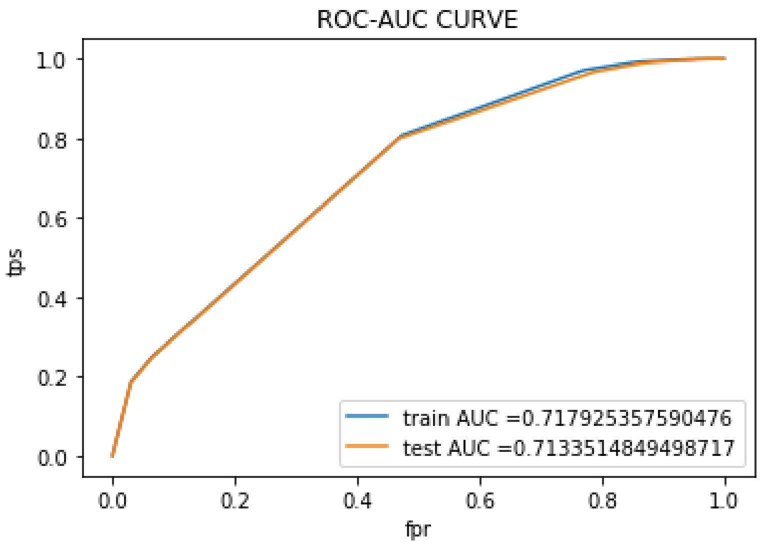
```
****************** cv data ******************
```



In [138]:

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth = 5, min_samples_split = 10)
clf.fit(X_train_tf_idf,y_train)
train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba
(X_train_tf_idf)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_
test_tf_idf)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tps")
plt.title("ROC-AUC CURVE")
plt.show()
```

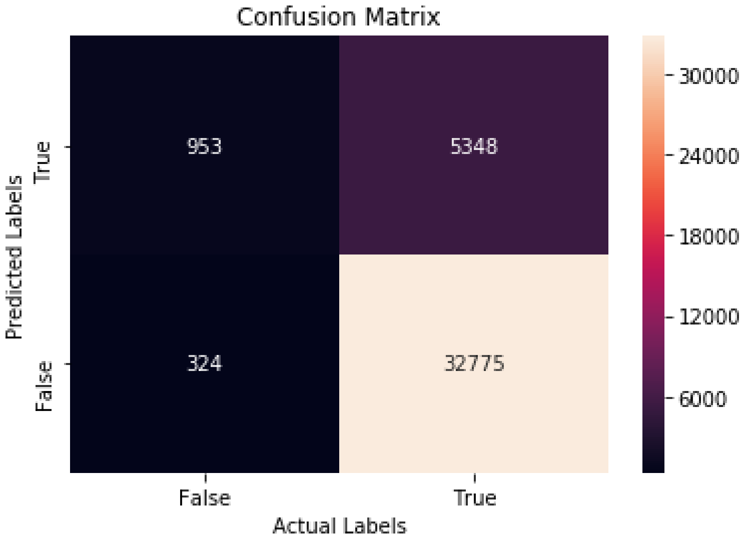


In [140]:

```
#for seaborn confusion matrix :https://stackoverflow.com/questions/3557
2000/how-can-i-plot-a-confusion-matrix
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-con
fusion-matrix-with-labels
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
```

```
print("Train confusion matrix")
ax= plt.subplot()
arr1=confusion_matrix(y_train, clf.predict(X_train_tf_idf))
df_1= pd.DataFrame(arr1, range(2),range(2))
plt.figure(figsize = (5,2))
sn.heatmap(df_1, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Train confusion matrix
```



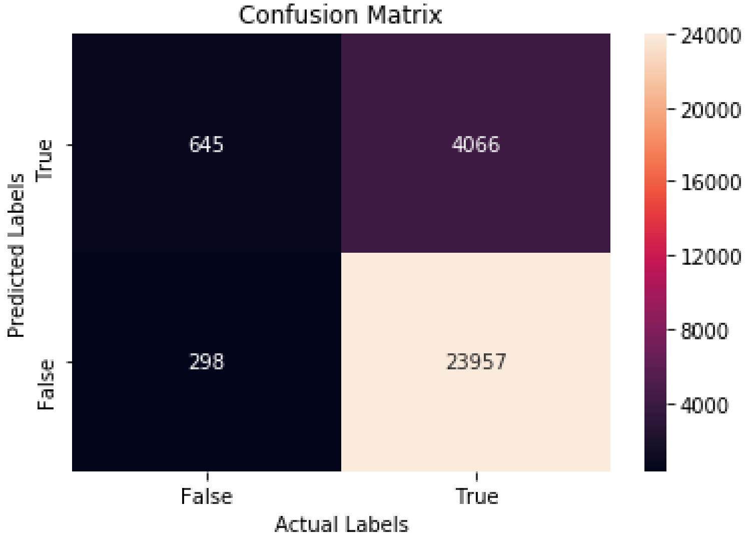
```
<Figure size 360x144 with 0 Axes>
```

In [141]:

```
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-confusion-matrix-with-labels
print("Test confusion matrix")
ax= plt.subplot()
arr2=confusion_matrix(y_test, clf.predict(X_test_tf_idf))
df_2= pd.DataFrame(arr2, range(2),range(2))
plt.figure(figsize = (5,2))
```

```
sn.heatmap(df_2, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Test confusion matrix
```



```
<Figure size 360x144 with 0 Axes>
```

[5.2.1] Top 20 important features from SET 2

In [144]:

```
#i used below link as reference for printing feature importance with it
s name
#https://stackoverflow.com/questions/11116697/how-to-get-most-informati
ve-features-for-scikit-learn-classifiers
feature_names = tf_idf_vect.get_feature_names()
coefs = sorted(zip(clf.feature_importances_, feature_names))
top = coefs[:-(20 + 1):-1]
print("feature_importances\tfeatures")
```

```
for (coef, feat) in top:
  print("%f\t\t%s" % (coef, feat))
```

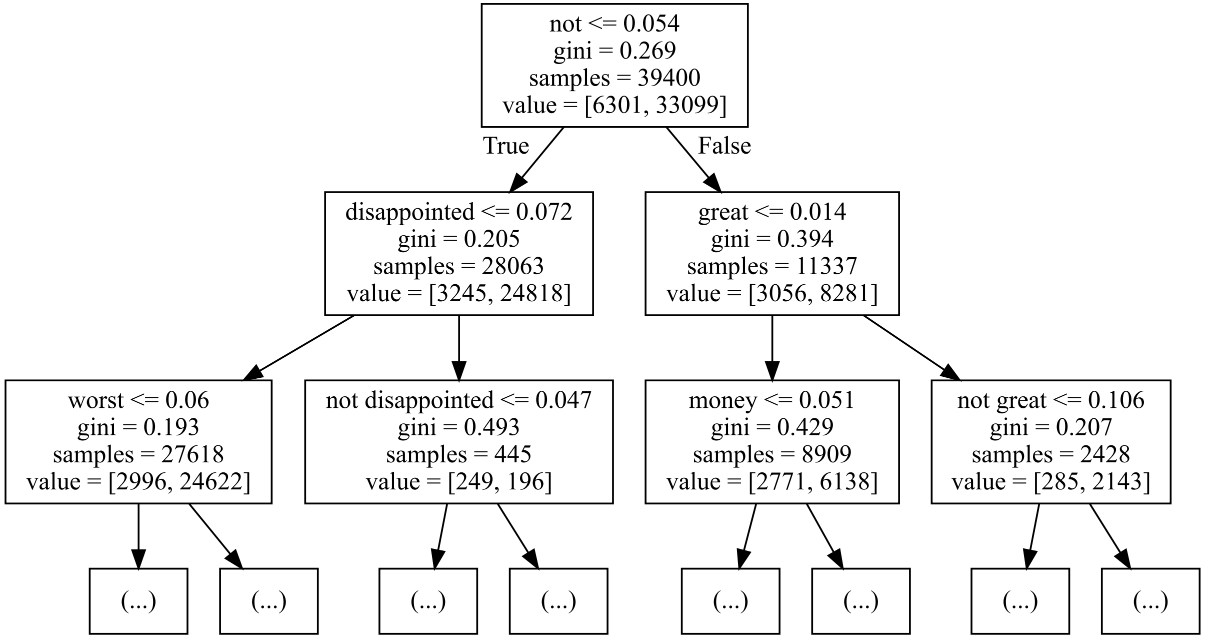
```
feature_importances     features
0.291371                not
0.147719                great
0.116799                disappointed
0.102439                worst
0.078174                money
0.072717                not buy
0.069801                horrible
0.032372                not disappointed
0.015842                not worth
0.010614                best
0.010466                waste
0.009799                not great
0.009720                never disappointed
0.006492                little disappointed
0.005443                little
0.004712                bit
0.003090                really
0.002954                touch
0.002915                tastes
0.002629                not get
```

[5.2.2] Graphviz visualization of Decision Tree on TFIDF, SET 2

In [146]:

```
from sklearn import tree
from graphviz import Source
import graphviz
feature = tf_idf_vect.get_feature_names()
Source(tree.export_graphviz(clf, out_file = None, feature_names = featu
re,max_depth=2))
```

Out[146]:



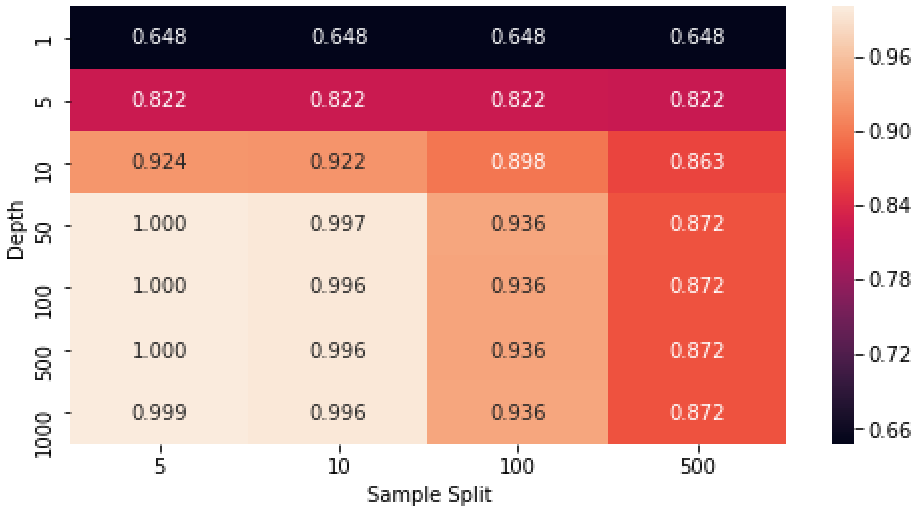
[5.3] Applying Decision Trees on AVG W2V, SET 3

In [147]:

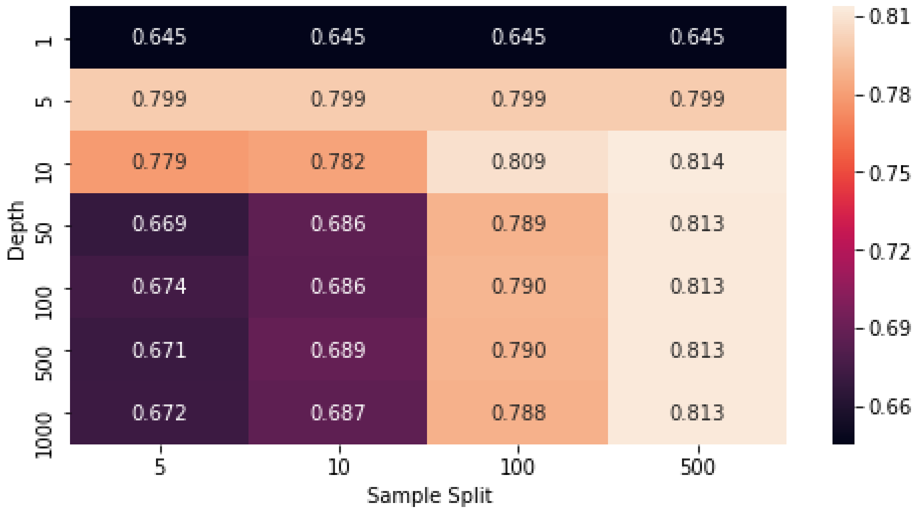
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score
depth = [1, 5, 10, 50, 100, 500, 1000]
min_samples_split = [5, 10, 100, 500]
train_auc = []
cv_auc = []
for d in depth:
  for m in min_samples_split:
    clf = DecisionTreeClassifier(max_depth = d, min_samples_split = m)
    clf.fit(sent_vectors_train, y_train);
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probab
ility estimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(sent_vectors_train)[:,1]
    y_cv_pred =  clf.predict_proba(sent_vectors_cv)[:,1]
    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

print("AUC SCORES")
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
print("*"*20, "train data", "*"*20)
train_auc= np.array(train_auc)
train_auc= train_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(train_auc,annot=True, fmt=".3f", xticklabels=min_samples_split,yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
print("*"*20, "cv data", "*"*20)
cv_auc= np.array(cv_auc)
cv_auc= cv_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(cv_auc,annot=True,fmt=".3f", xticklabels=min_samples_split, yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
```

```
AUC SCORES
******************** train data ********************
```



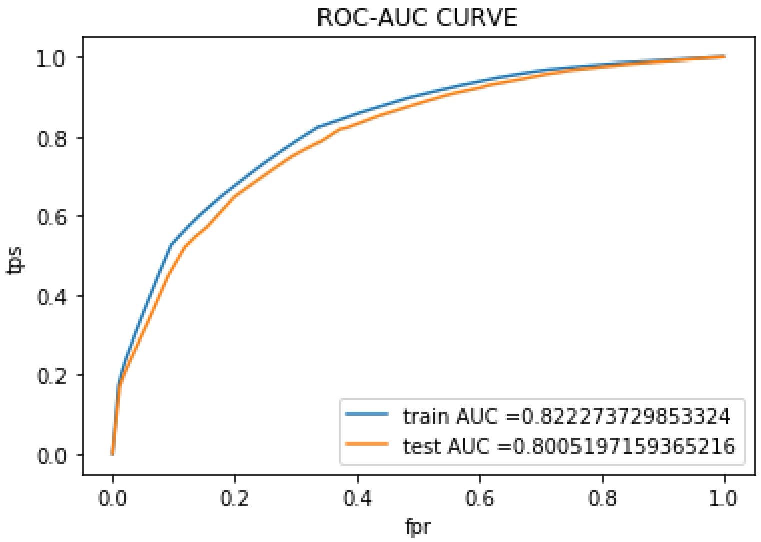
```
******************* cv data *******************
```



In [148]:

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth = 5, min_samples_split = 5)
clf.fit(sent_vectors_train,y_train)
train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba
(sent_vectors_train)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(se
nt_vectors_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tps")
plt.title("ROC-AUC CURVE")
plt.show()
```

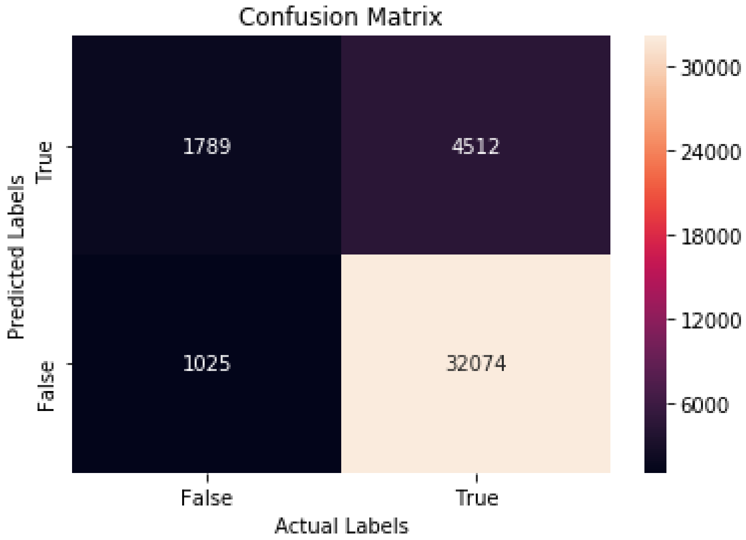


In [149]:

```
#for seaborn confusion matrix :https://stackoverflow.com/questions/3557
2000/how-can-i-plot-a-confusion-matrix
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-con
fusion-matrix-with-labels
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
```

```
print("Train confusion matrix")
ax= plt.subplot()
arr1=confusion_matrix(y_train, clf.predict(sent_vectors_train))
df_1= pd.DataFrame(arr1, range(2),range(2))
plt.figure(figsize = (5,2))
sn.heatmap(df_1, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Train confusion matrix
```



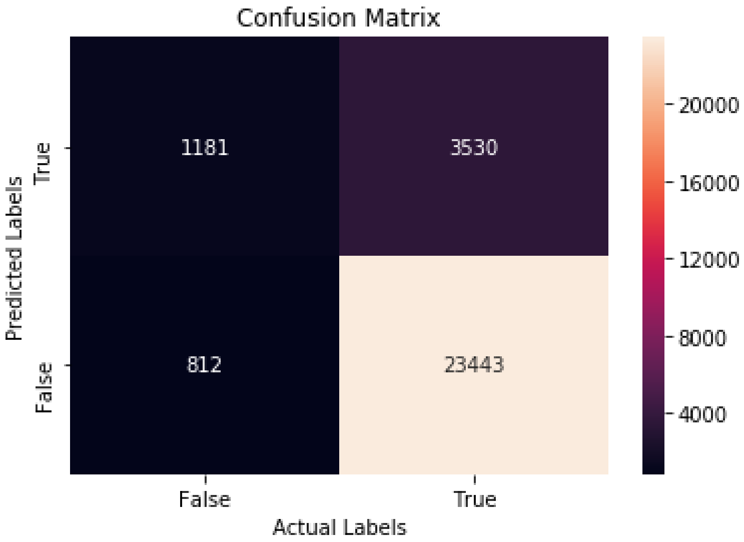
```
<Figure size 360x144 with 0 Axes>
```

In [150]:

```
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-con
fusion-matrix-with-labels
print("Test confusion matrix")
ax= plt.subplot()
arr2=confusion_matrix(y_test, clf.predict(sent_vectors_test))
df_2= pd.DataFrame(arr2, range(2),range(2))
plt.figure(figsize = (5,2))
```

```
sn.heatmap(df_2, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Test confusion matrix
```



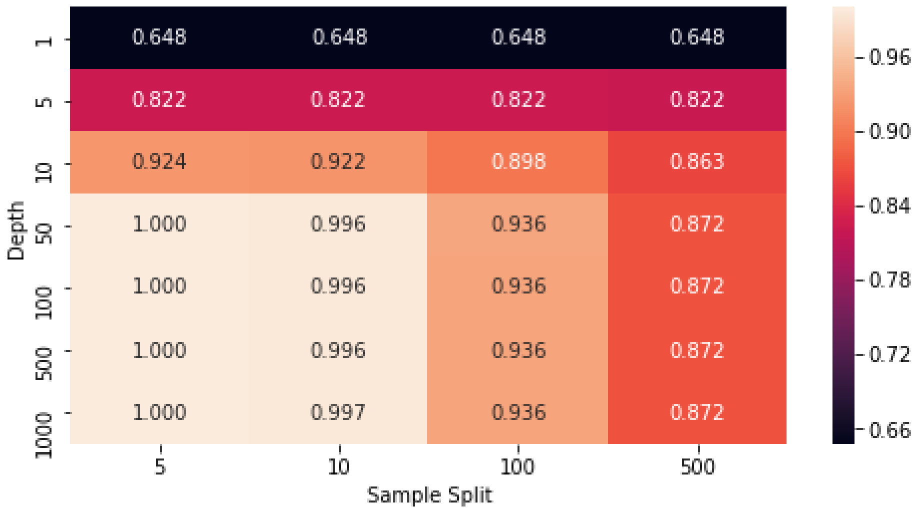
```
<Figure size 360x144 with 0 Axes>
```

[5.4] Applying Decision Trees on TFIDF W2V, SET 4

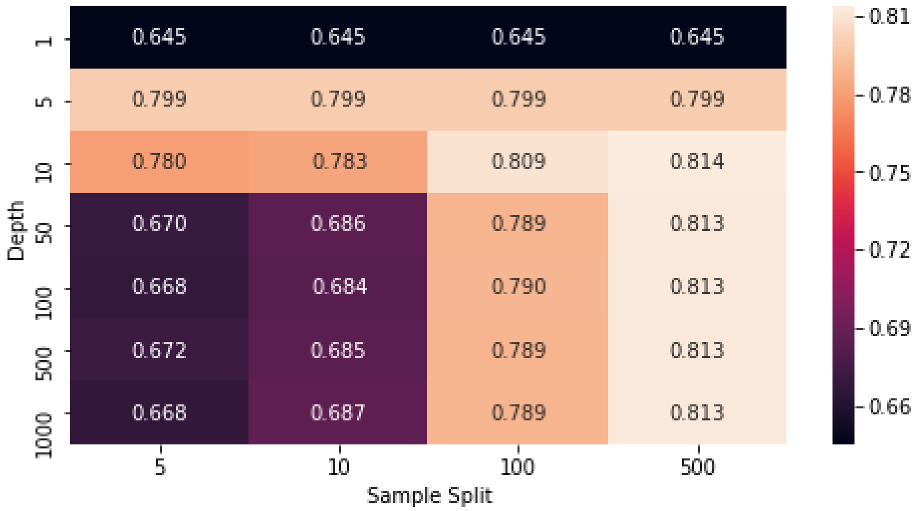
In [151]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score
depth = [1, 5, 10, 50, 100, 500, 1000]
min_samples_split = [5, 10, 100, 500]
train_auc = []
cv_auc = []
for d in depth:
  for m in min_samples_split:
    clf = DecisionTreeClassifier(max_depth = d, min_samples_split = m)
    clf.fit(tfidf_sent_vectors_train, y_train);
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probab
ility estimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(tfidf_sent_vectors_train)[:,1]
    y_cv_pred =  clf.predict_proba(tfidf_sent_vectors_cv)[:,1]
    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

print("AUC SCORES")
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
print("*"*20, "train data", "*"*20)
train_auc= np.array(train_auc)
train_auc= train_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(train_auc,annot=True, fmt=".3f", xticklabels=min_samples_sp
lit,yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
print("*"*20, "cv data", "*"*20)
cv_auc= np.array(cv_auc)
cv_auc= cv_auc.reshape(len(depth),len(min_samples_split))
plt.figure(figsize=(8,4))
sns.heatmap(cv_auc,annot=True,fmt=".3f", xticklabels=min_samples_split,
yticklabels=depth)
plt.xlabel('Sample Split')
plt.ylabel('Depth')
plt.show()
```

```
AUC SCORES
******************** train data ********************
```



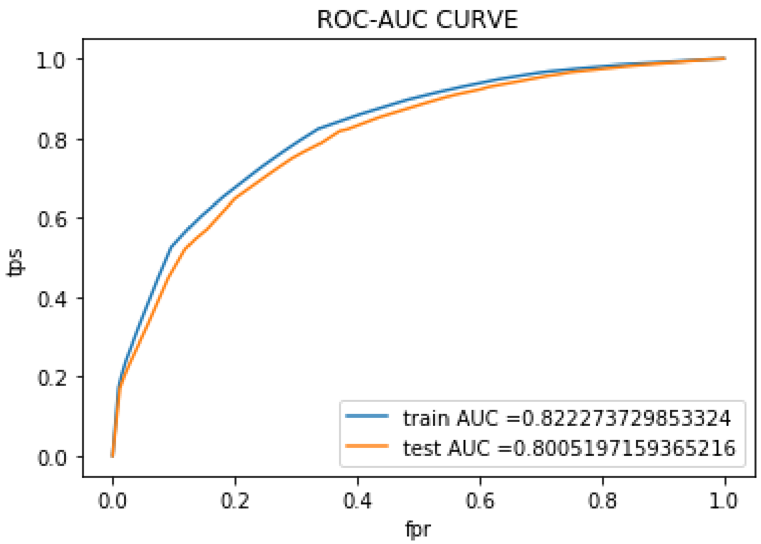
```
******************** cv data ********************
```



In [152]:

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth = 5, min_samples_split = 100)
clf.fit(tfidf_sent_vectors_train,y_train)
train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba
(tfidf_sent_vectors_train)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(tf
idf_sent_vectors_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tps")
plt.title("ROC-AUC CURVE")
plt.show()
```

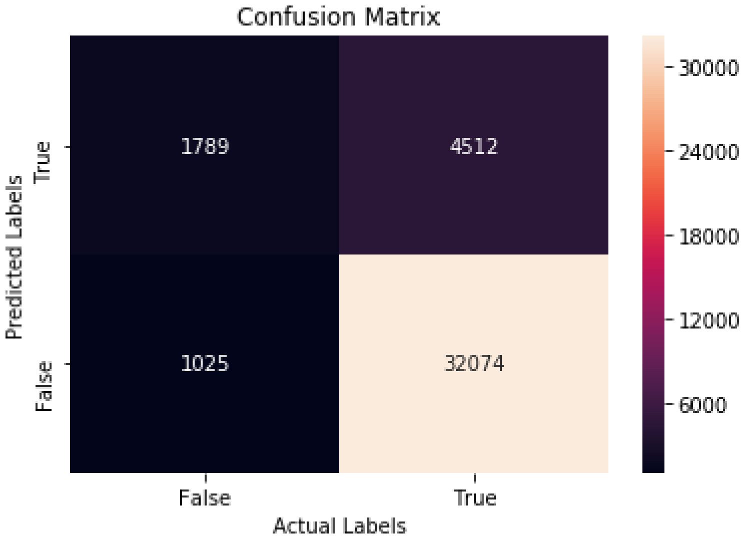


In [153]:

```
#for seaborn confusion matrix :https://stackoverflow.com/questions/3557
2000/how-can-i-plot-a-confusion-matrix
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-con
fusion-matrix-with-labels
import seaborn as sn
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
```

```
print("Train confusion matrix")
ax= plt.subplot()
arr1=confusion_matrix(y_train, clf.predict(tfidf_sent_vectors_train))
df_1= pd.DataFrame(arr1, range(2),range(2))
plt.figure(figsize = (5,2))
sn.heatmap(df_1, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Train confusion matrix
```



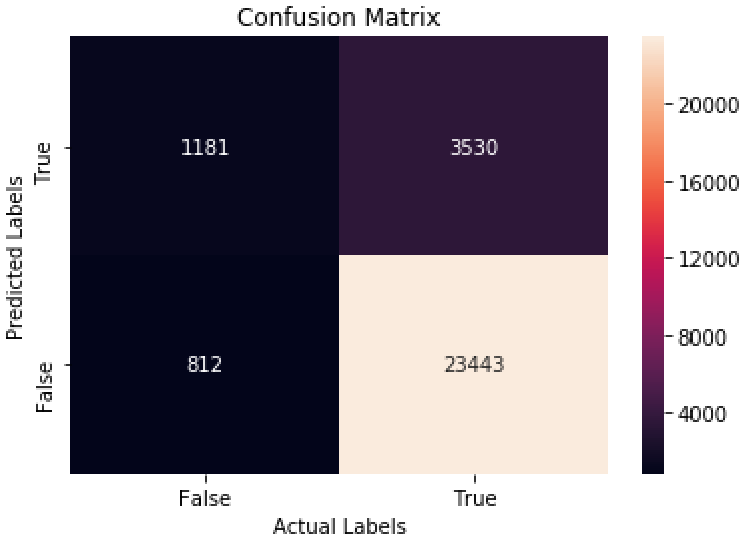
```
<Figure size 360x144 with 0 Axes>
```

In [154]:

```
#refernce:https://stackoverflow.com/questions/19233771/sklearn-plot-confusion-matrix-with-labels
print("Test confusion matrix")
ax= plt.subplot()
arr2=confusion_matrix(y_test, clf.predict(tfidf_sent_vectors_test))
df_2= pd.DataFrame(arr2, range(2),range(2))
plt.figure(figsize = (5,2))
```

```
sn.heatmap(df_2, annot=True,fmt="d",ax=ax)
ax.set_title('Confusion Matrix');
ax.set_xlabel('Actual Labels')
ax.set_ylabel('Predicted Labels')
ax.xaxis.set_ticklabels(['False', 'True']);
ax.yaxis.set_ticklabels(['True', 'False']);
```

```
Test confusion matrix
```



```
<Figure size 360x144 with 0 Axes>
```

[6] Conclusions

In [156]:

```
from prettytable import PrettyTable
print('auc performace table:')
x = PrettyTable()
x.field_names = ["Vectorizer", "Depth", 'min split value',"auc"]
x.add_row(["BoW", "10", 100,0.758])
x.add_row(["tfidf", "5",10, 0.71])
x.add_row(["avg w2v", "5", 5,0.80])
```

```
x.add_row(["tfidfw2v", "5",100, 0.80])
print(x)
```

```
auc performace table:
+------------+-------+-----------------+-------+
| Vectorizer | Depth | min split value |  auc  |
+------------+-------+-----------------+-------+
|    BoW     |   10  |       100       | 0.758 |
|   tfidf    |   5   |        10       |  0.71 |
|  avg w2v   |   5   |        5        |  0.8  |
|  tfidfw2v  |   5   |       100       |  0.8  |
+------------+-------+-----------------+-------+
```