In [70]:

```
#Decision trees assignment 8
```

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.

[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 [71]:

```
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

from IPython.display import Image
from sklearn.externals.six import StringIO
from sklearn.tree import export_graphviz
from sklearn.tree import DecisionTreeClassifier

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 [72]:

```
# using SQLite Table to read data.
con = sqlite3.connect('C:/Users/sesha/OneDrive/Desktop/IMP/before/MINIPJ/Personal/AMAZON food revi
ew dataset/database.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 500000 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 Score != 3 LIMIT 500000""", co
n)
# 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 score<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[72]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	1	1303862400	Good Quality Dog Food

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	0	1346976000	Not as Advertised
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	1	1219017600	"Delight" says it all

In [73]:

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

In [74]:

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

```
(80668, 7)
```

Out[74]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
0	#oc-R115TNMSPFT9I7	B005ZBZLT4	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ESG	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
2	#oc-R11DNU2NBKQ23Z	B005ZBZLT4	Kim Cieszykowski	1348531200	1	This coffee is horrible and unfortunately not ...	2
3	#oc-R11O5J5ZVQE25C	B005HG9ESG	Penguin Chick	1346889600	5	This will be the bottle that you grab from the...	3
4	#oc-R12KPBODL2B5ZD	B007OSBEV0	Christopher P. Presta	1348617600	1	I didnt like this coffee. Instead of telling y...	2

In [75]:

```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[75]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
80638	AZY10LLTJ71NX	B001ATMQK2	undertheshrine "undertheshrine"	1296691200	5	I bought this 6 pack because for the price tha...	5

In [76]:

```
display['COUNT(*)'].sum()
```

Out[76]:

```
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 [77]:

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

Out[77]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summ
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACK QUADRA VANI WAFE
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACK QUADRA VANI WAFE
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACK QUADRA VANI WAFE
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACK QUADRA VANI WAFE
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACK QUADRA VANI WAFE

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 [78]:

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

In [79]:

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

Out[79]:

```
(87775, 10)
```

In [80]:

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

Out[80]:

```
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 [81]:

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

display.head()
```

Out[81]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1	5	1224892800	Bought This for My Son at College
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2	4	1212883200	Pure cocoa taste with crunchy almonds inside

In [82]:

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

In [83]:

```
#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[83]:

```
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 [84]:

```
# printing some random reviews
sent_0 = final['Text'].values[0]
print(sent_0)
print("="*50)

sent_1000 = final['Text'].values[1000]
print(sent_1000)
print("="*50)

sent_1500 = final['Text'].values[1500]
print(sent_1500)
print("="*50)

sent_4900 = final['Text'].values[4900]
print(sent_4900)
print("="*50)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its ver
y hard to find any chicken products made in the USA but they are out there, but this one isnt.  It
s too bad too because its a good product but I wont take any chances till they know what is going
on with the china imports.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste to
it.  Very little of the 2 lbs that I bought were eaten and I threw the rest away.  I would not buy
the candy again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So if you are afraid of
the fishy smell, don't get it. But I think my dog likes it because of the smell. These treats are
really small in size. They are great for training. You can give your dog several of these without
worrying about him over eating. Amazon's price was much more reasonable than any other retailer. Y
ou can buy a 1 pound bag on Amazon for almost the same price as a 6 ounce bag at other retailers.
It's definitely worth it to buy a big bag if your dog eats them a lot.
==================================================
```

In [85]:

```
# remove urls from text python: https://stackoverflow.com/a/40823105/4084039
sent_0 = re.sub(r"http\S+", "", sent_0)
sent_1000 = re.sub(r"http\S+", "", sent_1000)
sent_150 = re.sub(r"http\S+", "", sent_1500)
sent_4900 = re.sub(r"http\S+", "", sent_4900)

print(sent_0)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its ver
y hard to find any chicken products made in the USA but they are out there, but this one isnt.  It
s too bad too because its a good product but I wont take any chances till they know what is going
on with the china imports.
```

In [86]:

```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an-element
from bs4 import BeautifulSoup

soup = BeautifulSoup(sent_0, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1000, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1500, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_4900, 'lxml')
text = soup.get_text()
print(text)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its ver
y hard to find any chicken products made in the USA but they are out there, but this one isnt.  It
s too bad too because its a good product but I wont take any chances till they know what is going
on with the china imports.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste to
it.  Very little of the 2 lbs that I bought were eaten and I threw the rest away.  I would not buy
the candy again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So if you are afraid of
the fishy smell, don't get it. But I think my dog likes it because of the smell. These treats are
really small in size. They are great for training. You can give your dog several of these without
worrying about him over eating. Amazon's price was much more reasonable than any other retailer. Y
ou can buy a 1 pound bag on Amazon for almost the same price as a 6 ounce bag at other retailers.
It's definitely worth it to buy a big bag if your dog eats them a lot.
```

In [87]:

```
# 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 [88]:

```
sent_1500 = decontracted(sent_1500)
print(sent_1500)
print("="*50)
```

```
was way to hot for my blood, took a bite and did a jig  lol
==================================================
```

In [89]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its ver
y hard to find any chicken products made in the USA but they are out there, but this one isnt.  It
s too bad too because its a good product but I wont take any chances till they know what is going
on with the china imports.
```

In [90]:

```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent_1500 = re.sub('[^A-Za-z0-9]+', ' ', sent_1500)
print(sent_1500)
```

```
was way to hot for my blood took a bite and did a jig lol
```

In [91]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
# <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', 'ours', 'ourselves', 'you', "y
ou're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', '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', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', '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', "d
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn'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 [92]:

```
# Combining all the above stundents
from tqdm import tqdm
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
[01:01<00:00, 1435.13it/s]
```

In [93]:

```
preprocessed_reviews[1500]
```

Out[93]:

```
'way hot blood took bite jig lol'
```

In [94]:

```
print(len(preprocessed_reviews))
final.shape
```

```
87773
```

Out[94]:

```
(87773, 10)
```

[3.2] Preprocessing Review Summary

In [95]:

```
preprocessed_summary = []
# tqdm is for printing the status bar
for summary in tqdm(final['Summary'].values):
    summary = re.sub(r"http\S+", "", summary)
    # remove urls from text python: https://stackoverflow.com/a/40823105/4084039
    summary = BeautifulSoup(summary, 'lxml').get_text()
    # https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an-element
    summary = decontracted(summary)
    summary = re.sub("\S*\d\S*", "", summary).strip() #remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
    summary = re.sub('[^A-Za-z]+', ' ', summary) #remove spacial character: https://stackoverflow.com/a/5843547/4084039
    # https://gist.github.com/sebleier/554280
    summary = ' '.join(e.lower() for e in summary.split() if e.lower() not in stopwords)
    preprocessed_summary.append(summary.strip())
```

```
100%|██████████████████████████████████████████████████████████████████████████| 87773/87773
[00:43<00:00, 2033.87it/s]
```

In [96]:

```
preprocessed_reviews = [i + ' ' + j for i, j in zip(preprocessed_reviews,preprocessed_summary)]
print(preprocessed_reviews[1500])
```

```
way hot blood took bite jig lol hot stuff
```

[4] Featurization

[4.1] BAG OF WORDS WITH TIME SERIES SPLITTING

In [97]:

```
Y = final['Score'].values
X = np.array(preprocessed_reviews)
```

In [98]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
```

```
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=False,random_stat
e=0)
X_train, X_cv, Y_train, Y_cv = train_test_split(X_train, Y_train,
test_size=0.33,shuffle=False,random_state=0)# this is for time series split
#X_train, X_test, y_train, y_test = train_test_split(final['preprocessed_reviews'],
final['Score'], test_size=0.33) # 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 rando
m splitting


print(X_train.shape, Y_train.shape)
print(X_cv.shape, Y_cv.shape)
print(X_test.shape, Y_test.shape)

print("="*100)

from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
vectorizer.fit(X_train) # fit has to happen only on train data

# 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)

print("NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME")
```

```
(39400,) (39400,)
(19407,) (19407,)
(28966,) (28966,)
====================================================================================================

After vectorizations
(39400, 38503) (39400,)
(19407, 38503) (19407,)
(28966, 38503) (28966,)
====================================================================================================

NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME
```

[4.2] Bi-Grams and n-Grams.

In [99]:

```
#bi-gram, tri-gram and n-gram

#removing stop words like "not" should be avoided before building n-grams
# count_vect = CountVectorizer(ngram_range=(1,2))
# please do read the CountVectorizer documentation http://scikit-
learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

# you can choose these numebrs min_df=10, max_features=5000, of your choice
count_vect = CountVectorizer(ngram_range=(1,2), min_df=10, max_features=5000)
final_bigram_counts = count_vect.fit_transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_bigram_counts))
print("the shape of out text BOW vectorizer ",final_bigram_counts.get_shape())
print("the number of unique words including both unigrams and bigrams ", final_bigram_counts.get_s
hape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (87773, 5000)
the number of unique words including both unigrams and bigrams  5000
```

[4.3] TF-IDF

In [100]:

```
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.get_feature_names()[0:10])
print('='*50)

train_tf_idf = tf_idf_vect.transform(X_train)
cv_tf_idf = tf_idf_vect.transform(X_cv)
test_tf_idf = tf_idf_vect.transform(X_test)
print("the type of count vectorizer ",type(train_tf_idf))
print("the shape of out text TRAIN TFIDF vectorizer ",train_tf_idf.get_shape())
print("the shape of out text CV TFIDF vectorizer ",cv_tf_idf.get_shape())
print("the shape of out text TEST TFIDF vectorizer ",test_tf_idf.get_shape())
print("the number of unique words including both unigrams and bigrams in train ", train_tf_idf.get
_shape()[1])
```

```
some sample features(unique words in the corpus) ['abdominal', 'ability', 'able', 'able buy',
'able drink', 'able eat', 'able enjoy', 'able find', 'able get', 'able give']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TRAIN TFIDF vectorizer  (39400, 24483)
the shape of out text CV TFIDF vectorizer  (19407, 24483)
the shape of out text TEST TFIDF vectorizer  (28966, 24483)
the number of unique words including both unigrams and bigrams in train  24483
```

[4.4] Word2Vec

In [101]:

```
# Train your own Word2Vec model using your own text corpus
i=0
sent_of_train=[]
for sentance in X_train:
    sent_of_train.append(sentance.split())
```

In [102]:

```
# Train your own Word2Vec model using your own text corpus
i=0
sent_of_test=[]
for sentance in X_test:
    sent_of_test.append(sentance.split())
```

In [103]:

```
# Train your own Word2Vec model using your own text corpus
i=0
sent_of_cv=[]
for sentance in X_cv:
    sent_of_cv.append(sentance.split())
```

In [104]:

```
# Using Google News Word2Vectors

# in this project we are using a pretrained model by google
# its 3.3G file, once you load this into your memory
# it occupies ~9Gb, so please do this step only if you have >12G of ram
# we will provide a pickle file wich contains a dict ,
# and it contains all our courpus words as keys and  model[word] as values
# To use this code-snippet, download "GoogleNews-vectors-negative300.bin"
# from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
# it's 1.9GB in size.


# http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.W17SRFAzZPY
# you can comment this whole cell
# or change these varible according to your need
```

```
# or change these varible according to your need

is_your_ram_gt_16g=False
want_to_use_google_w2v = False
want_to_train_w2v = True

if want_to_train_w2v:
    # min_count = 5 considers only words that occured atleast 5 times
    w2v_model=Word2Vec(sent_of_train,min_count=5,size=50, workers=4)
    print(w2v_model.wv.most_similar('great'))
    print('='*50)
    print(w2v_model.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=Tr
ue)
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
        print("you don't have gogole's word2vec file, keep want_to_train_w2v = True, to train your
own w2v ")
```

```
[('fantastic', 0.8444055318832397), ('awesome', 0.8181746006011963), ('good', 0.8143923282623291),
('excellent', 0.7954152822494507), ('terrific', 0.783186674118042), ('wonderful',
0.7742859125137329), ('perfect', 0.7324698567390442), ('amazing', 0.7161928415298462), ('decent',
0.7071163058280945), ('fabulous', 0.6928749084472656)]
==================================================
[('softest', 0.7129153609275818), ('experienced', 0.7039514780044556), ('best',
0.6989202499389648), ('overrated', 0.6799666881561279), ('tastiest', 0.6547428369522095),
('nastiest', 0.6543765068054199), ('greatest', 0.6496132612228394), ('hottest',
0.6328741312026978), ('biggest', 0.62213534116745), ('closest', 0.6113991141319275)]
```

In [105]:

```
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])
```

```
number of words that occured minimum 5 times  12321
sample words  ['dogs', 'loves', 'chicken', 'product', 'china', 'wont', 'buying', 'anymore',
'hard', 'find', 'products', 'made', 'usa', 'one', 'isnt', 'bad', 'good', 'take', 'chances',
'till', 'know', 'going', 'imports', 'love', 'saw', 'pet', 'store', 'tag', 'attached', 'regarding',
'satisfied', 'safe', 'dog', 'lover', 'infestation', 'literally', 'everywhere', 'flying', 'around',
'kitchen', 'bought', 'hoping', 'least', 'get', 'rid', 'weeks', 'fly', 'stuck', 'buggers', 'success
']
```

[4.4.1] Converting text into vectors using Avg W2V, TFIDF-W2V

[4.4.1.1] Avg W2v

In [106]:

```
# 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(sent_of_cv): # for each review/sentence
    sent_vec_cv = 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_cv += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec_cv /= cnt_words
    sent_vectors_cv.append(sent_vec_cv)
print(len(sent_vectors_cv))
print(len(sent_vectors_cv[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 19407/19407 [02:14<00:00, 144.12it/s]
```

```
19407
50
```

In [107]:

```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train= []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(sent_of_train): # for each review/sentence
    sent_vec_train = 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_train += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec_train /= cnt_words
    sent_vectors_train.append(sent_vec_train)
print(len(sent_vectors_train))
print(len(sent_vectors_train[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 39400/39400 [03:57<00:00, 165.70it/s]
```

```
39400
50
```

In [108]:

```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_test= []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(sent_of_test): # for each review/sentence
    sent_vec_test = 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_test += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec_test /= cnt_words
    sent_vectors_test.append(sent_vec_test)
print(len(sent_vectors_test))
print(len(sent_vectors_test[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 28966/28966 [01:16<00:00, 380.60it/s]
```

```
28966
50
```

[4.4.1.2] TFIDF weighted W2v

In [109]:

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

In [110]:

```
# 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 cell_val = tfidf

tfidf_sent_vectors_cv = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(sent_of_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/review
    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
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 19407/19407 [04
:59<00:00, 64.86it/s]
```

In [111]:

```
# 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 cell_val = tfidf

tfidf_sent_vectors_test = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(sent_of_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/review
    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
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 28966/28966 [07
:28<00:00, 64.62it/s]
```

In [112]:

```
# 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 cell_val = tfidf

tfidf_sent_vectors_train = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(sent_of_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/review
    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
```

```
100%|████████████████████████████████████████████████████████████████████████████| 39400/39400 [09
:29<00:00, 76.45it/s]
```

[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 [4,6, 8, 9,10,12,14,17] , and the best `min_samples_split` in range [2,10,20,30,40,50])**

   - 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**

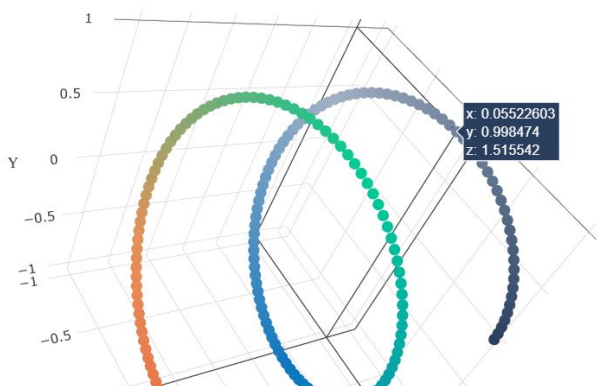
   - Find the top 20 important features from both feature sets Set 1 and Set 2 using `feature_importances_` method of Decision Tree Classifier and print their corresponding feature names

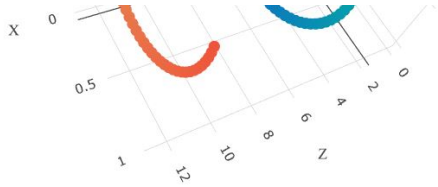
5. **Feature engineering**

   - 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

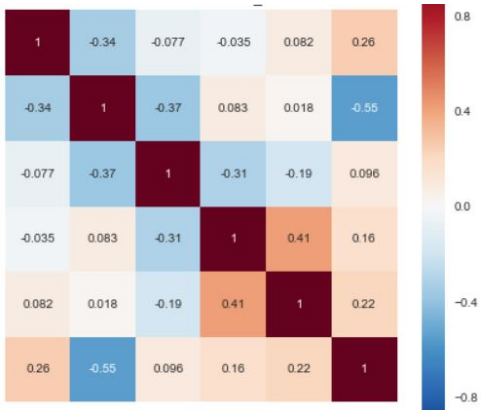




with X-axis as **min_sample_split**, Y-axis as **max_depth**, and Z-axis as **AUC Score** , we have given the notebook which explains how to plot this 3d plot, you can find it in the same drive *3d_scatter_plot.ipynb*

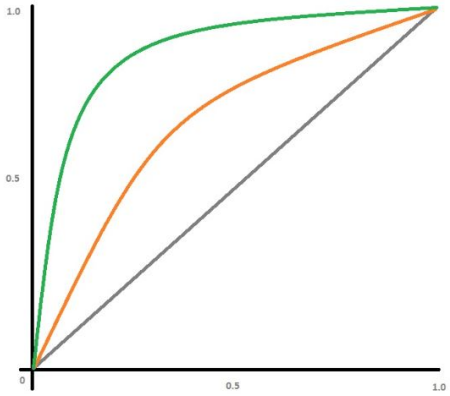
or

- 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



seaborn heat maps with rows as **min_sample_split**, columns as **max_depth**, and values inside the cell representing **AUC Score**

- You choose either of the plotting techniques out of 3d plot or heat map
- 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

Applying Decision Trees

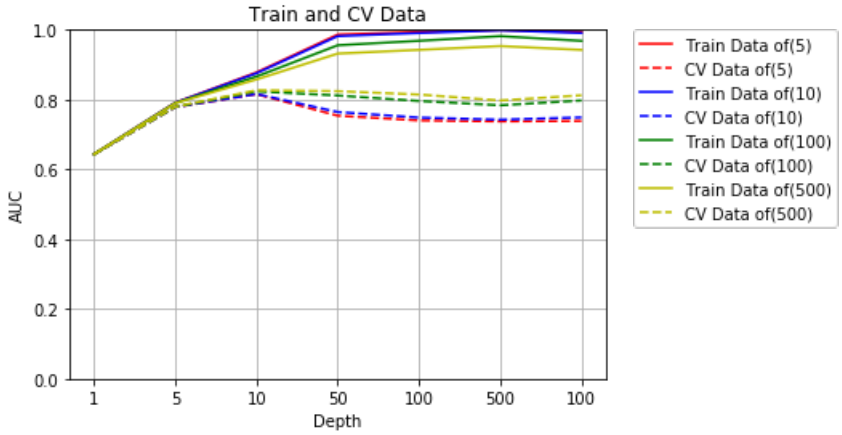
[5.1] Applying Decision Trees on BOW, SET 1

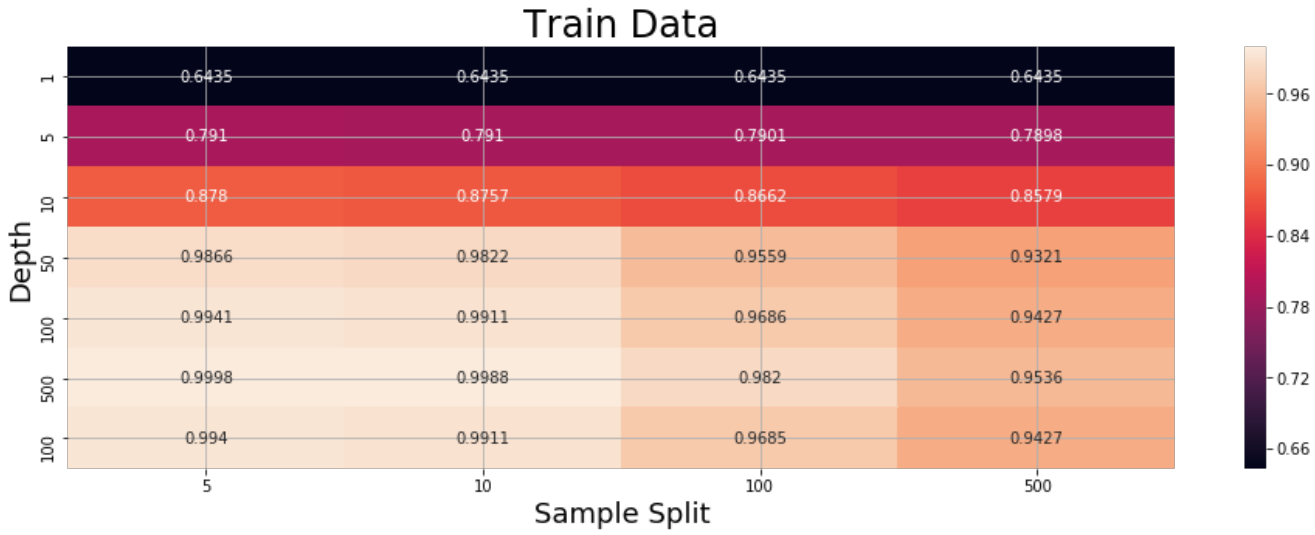
In [113]:

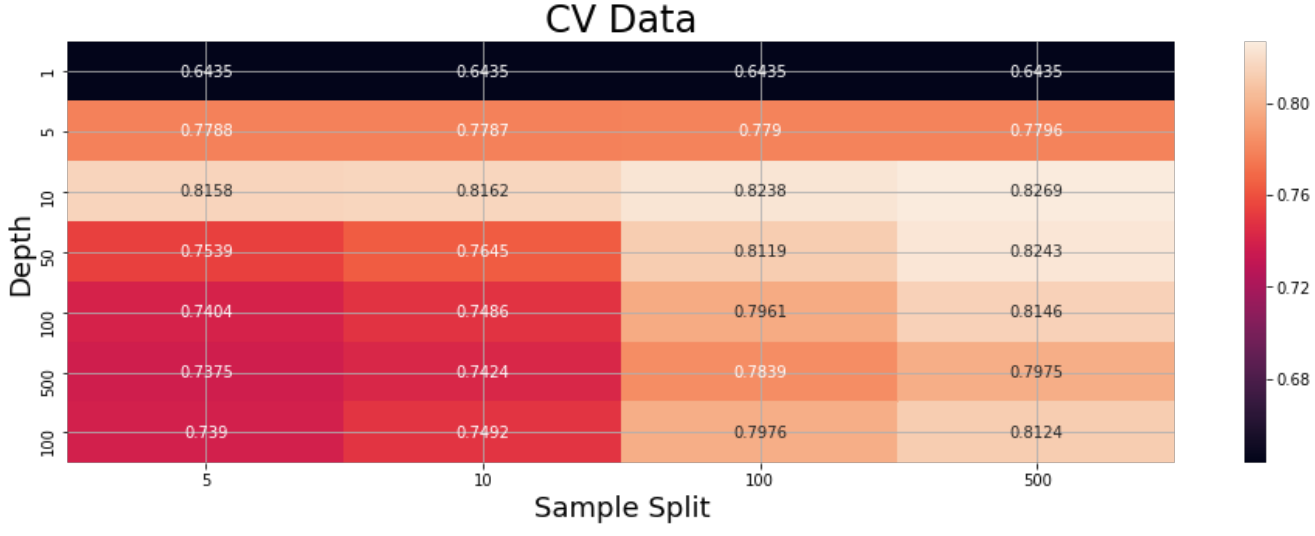
```
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
from sklearn.metrics import roc_auc_score, auc
from sklearn.model_selection import GridSearchCV
```

```
#given depth and min_samples_split and setting them as hyperparam
depth = [1, 5, 10, 50, 100, 500, 100]
min_sample_split = [5, 10, 100, 500]
hyper_param = {'max_depth':depth, 'min_samples_split':min_sample_split}

#declaring the classifier as DecisionTreeClassifier, fitting the classifier and finding the optimu
m values of depth and min_sample_split
clf = DecisionTreeClassifier(class_weight='balanced')
gsv = GridSearchCV(clf,hyper_param,scoring='roc_auc')
gsv.fit(X_train_bow,Y_train)
opt_depth_bow, opt_split_bow  = gsv.best_params_.get('max_depth'), gsv.best_params_.get('min_sample
s_split')

#computing train _auc and cv_auc
train_auc= gsv.cv_results_['mean_train_score']
#train_auc_std= gsv.cv_results_['std_train_score']
cv_auc = gsv.cv_results_['mean_test_score']
#cv_auc_std= gsv.cv_results_['std_test_score']

#plotting the AUC accordingto depth values [5,10,100,500]
x2 = np.arange(len(depth))
plt.plot(x2,train_auc[::4],'r', label = 'Train Data of(5)')
plt.plot(x2,cv_auc[::4],'r--', label = 'CV Data of(5)')

plt.plot(x2,train_auc[1::4],'b', label = 'Train Data of(10)')
plt.plot(x2,cv_auc[1::4],'b--', label = 'CV Data of(10)')

plt.plot(x2,train_auc[2::4],'g', label = 'Train Data of(100)')
plt.plot(x2,cv_auc[2::4],'g--', label = 'CV Data of(100)')

plt.plot(x2,train_auc[3::4],'y', label = 'Train Data of(500)')
plt.plot(x2,cv_auc[3::4],'y--', label = 'CV Data of(500)')

#plotting the graph for the AUC
plt.xticks(x2, depth)
plt.ylim(0,1)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.title("Train and CV Data")
plt.xlabel("Depth")
plt.ylabel("AUC")
plt.show()

#heatmap for train_auc
df_heatmap_train_auc = pd. DataFrame(train_auc. reshape(7, 4), index=depth, columns=min_sample_spli
t)
fig = plt. figure(figsize=(16,5))
heatmap_train_auc = sns. heatmap(df_heatmap_train_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth', size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("Train Data", size=24)
plt. show()

#heatmap for cv_auc
df_heatmap_cv_auc = pd. DataFrame(cv_auc . reshape(7, 4), index=depth, columns=min_sample_split)
fig = plt. figure(figsize=(16,5))
heatmap_cv_auc = sns. heatmap(df_heatmap_cv_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth' , size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("CV Data", size=24)
plt. show()

print("Optimal value of max_depth = ", opt_depth_bow , " Optimal min_samples_split is :",
opt_split_bow)

#Cv auc scores
print("--------------------------")
print("Cv auc scores")
print(cv_auc)
print("Maximun Auc value :",max(cv_auc))

#Verifying the model on Test data
clf = DecisionTreeClassifier(max_depth=opt_depth, min_samples_split=opt_split,class_weight=
'balanced')
```

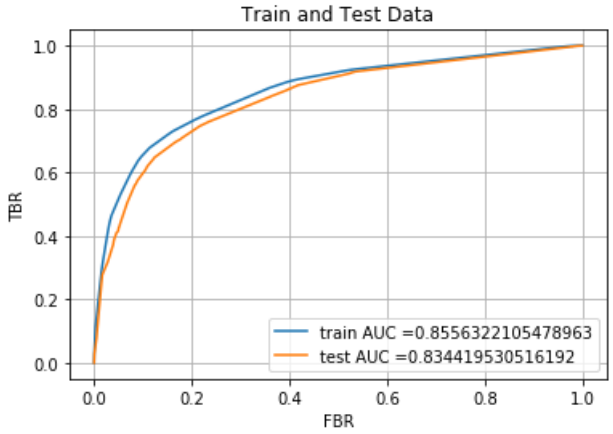
```
clf.fit(X_train_bow,Y_train)

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(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.legend()
plt.xlabel("FBR")
plt.ylabel("TBR")
plt.title("Train and Test Data")
plt.show()
```







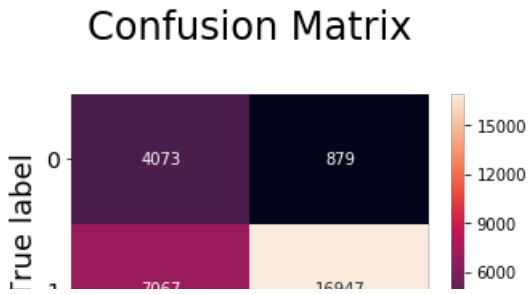
```
Optimal value of max_depth =  10  Optimal min_samples_split is : 500
--------------------------
Cv auc scores
[0.64347723 0.64347723 0.64347723 0.64347723 0.77879993 0.77873995
 0.77903146 0.77959447 0.81583421 0.81619655 0.82382027 0.8268789
 0.75392632 0.76446727 0.8118528  0.8243061  0.74036674 0.74859479
 0.79614846 0.81461008 0.73753759 0.74244687 0.78385586 0.79753997
 0.73903541 0.74915238 0.79759517 0.81243854]
Maximun Auc value : 0.826878903410373
```



In [114]:

```
#Confusion Matrix

print("Train confusion matrix")
print(confusion_matrix(Y_train, clf.predict(X_train_bow)))
print("Test confusion matrix")
print(confusion_matrix(Y_test, clf.predict(X_test_bow)))

cm = confusion_matrix(Y_train, clf.predict(X_train_bow))
cm = confusion_matrix(Y_test, clf.predict(X_test_bow))
tn, fp, fn, tp = cm.ravel()
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

# Code for drawing seaborn heatmaps
class_names = ['0','1']
df_heatmap = pd.DataFrame(cm, index=class_names, columns=class_names )
fig = plt.figure(figsize=(5,3))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
plt.ylabel('True label',size=18)
plt.xlabel('Predict label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```

```
Train confusion matrix
[[ 5176   998]
 [ 8946 24280]]
Test confusion matrix
[[ 4073   879]
 [ 7067 16947]]
```



[5.1.1] Top 20 important features from BOW

In [115]:

```
clf = DecisionTreeClassifier(max_depth= 50, min_samples_split=500,class_weight= 'balanced')
clf.fit(X_train_bow,Y_train)
feat = vectorizer.get_feature_names()
n=20
coefs = sorted(zip(clf.feature_importances_, feat))
top = coefs[:-(n + 1):-1]

print("Feature importances\tFeatures")
for (coef1, feat1) in top:
    print("%.4f\t\t\t%-15s" % (coef1, feat1))
```

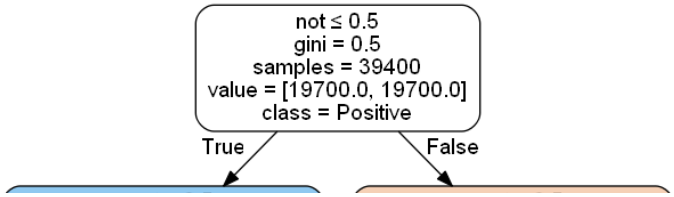
```
Feature importances Features
0.1786   not
0.1206   great
0.0628   best
0.0584   delicious
0.0329   love
0.0281   good
0.0248   perfect
0.0244   loves
0.0241   excellent
0.0212   disappointed
0.0160   bad
0.0108   wonderful
0.0100   yummy
0.0097   nice
0.0094   favorite
0.0076   tasty
0.0074   awful
0.0071   horrible
0.0070   worst
0.0063   easy
```

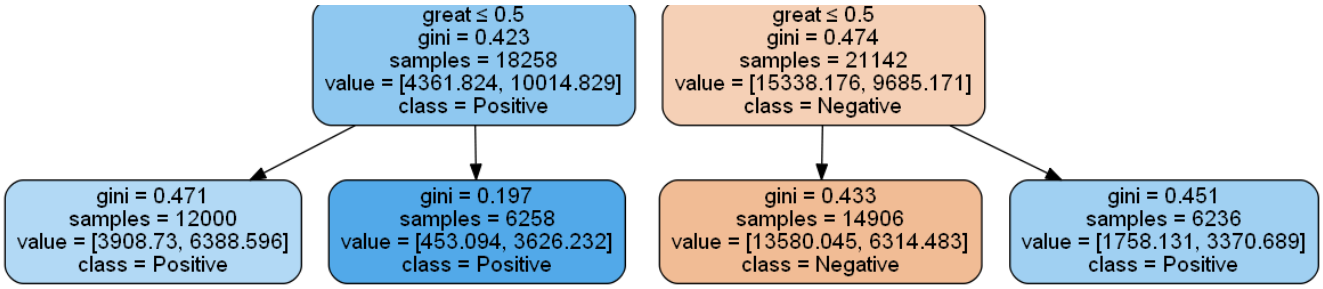
[5.1.2] Graphviz visualization of Decision Tree on BOW

In [116]:

```
#install Graphviz
#set the environment variable's path
#install pydotplus
import pydot
clf = DecisionTreeClassifier(max_depth= 2, min_samples_split=500,class_weight= 'balanced')
clf.fit(X_train_bow,Y_train)
names=vectorizer.get_feature_names()
dot_data = StringIO()
tree.export_graphviz(clf, out_file=dot_data,feature_names=names,
                     class_names=['Negative','Positive'],
                     filled=True, rounded=True,
                     special_characters=True)
graph = pydot.graph_from_dot_data(dot_data.getvalue())[0]
Image(graph.create_png())
```

Out[116]:



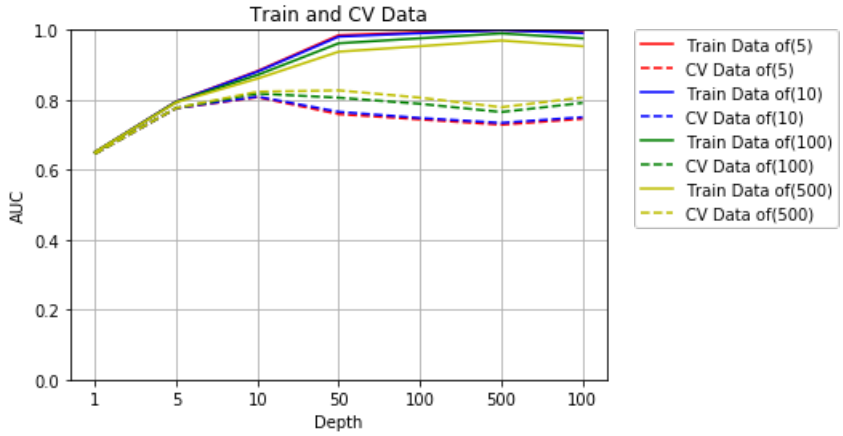


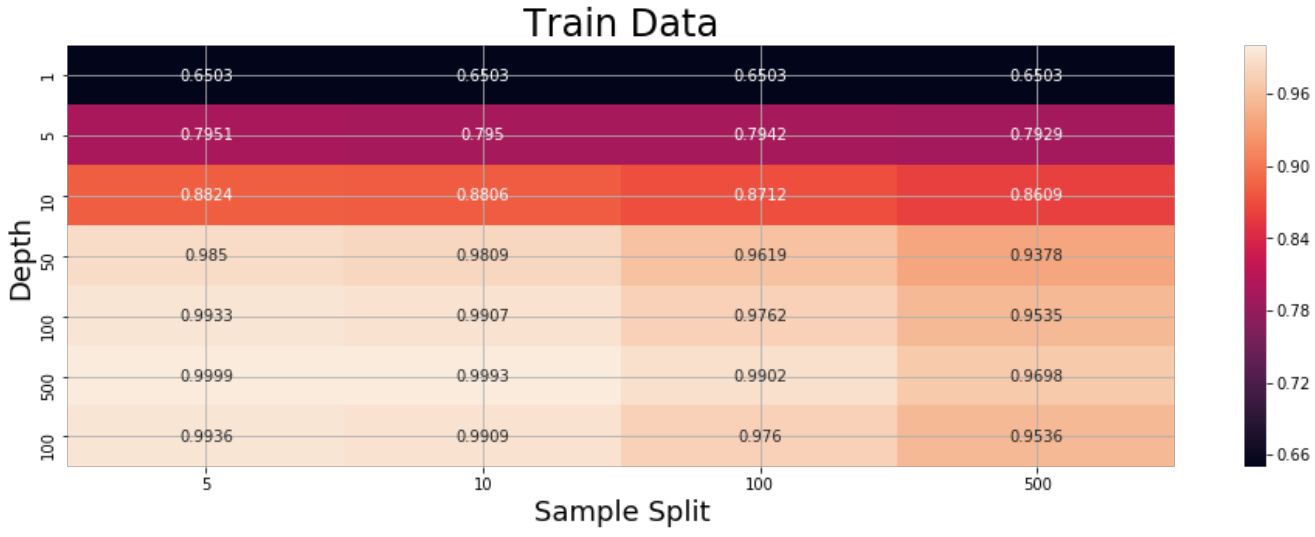
[5.2] Applying Decision Trees on TFIDF, SET 2

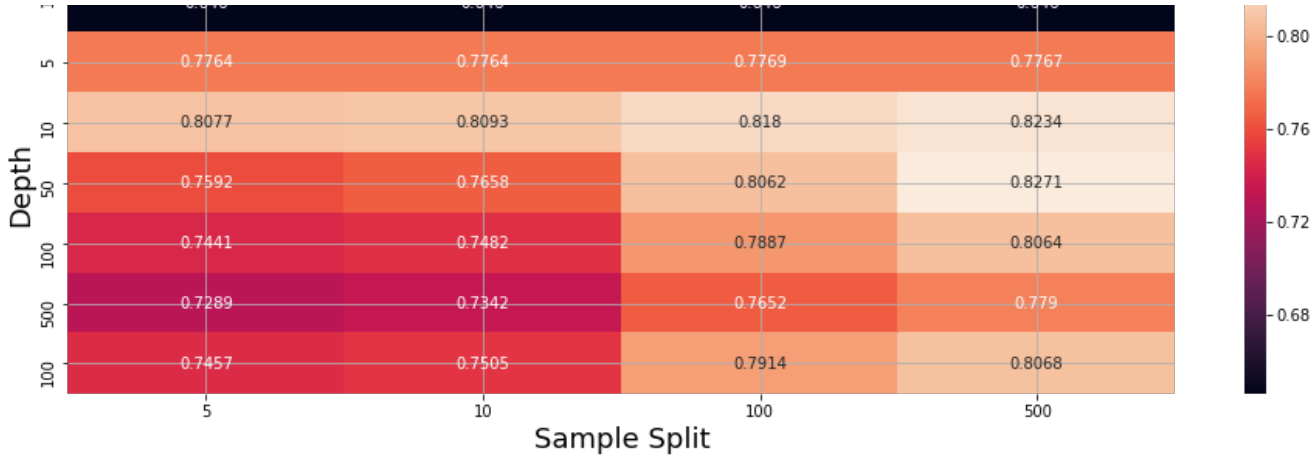
In [117]:

```
#given depth and min_samples_split and setting them as hyperparam
depth = [1, 5, 10, 50, 100, 500, 100]
min_sample_split = [5, 10, 100, 500]
hyper_param = {'max_depth':depth, 'min_samples_split':min_sample_split}

#declaring the classifier as DecisionTreeClassifier, fitting the classifier and finding the optimu
m values of depth and min_sample_split
clf = DecisionTreeClassifier(class_weight='balanced')
gsv = GridSearchCV(clf,hyper_param,scoring='roc_auc')
gsv.fit(train_tf_idf,Y_train)
opt_depth_tfidf, opt_split_tfidf  = gsv.best_params_.get('max_depth'), gsv.best_params_.get('min_sa
mples_split')

#computing train _auc and cv_auc
train_auc= gsv.cv_results_['mean_train_score']
#train_auc_std= gsv.cv_results_['std_train_score']
cv_auc = gsv.cv_results_['mean_test_score']
#cv_auc_std= gsv.cv_results_['std_test_score']

#plotting the AUC accordingto depth values [5,10,100,500]
x2 = np.arange(len(depth))
plt.plot(x2,train_auc[::4],'r', label = 'Train Data of(5)')
plt.plot(x2,cv_auc[::4],'r--', label = 'CV Data of(5)')

plt.plot(x2,train_auc[1::4],'b', label = 'Train Data of(10)')
plt.plot(x2,cv_auc[1::4],'b--', label = 'CV Data of(10)')

plt.plot(x2,train_auc[2::4],'g', label = 'Train Data of(100)')
plt.plot(x2,cv_auc[2::4],'g--', label = 'CV Data of(100)')

plt.plot(x2,train_auc[3::4],'y', label = 'Train Data of(500)')
plt.plot(x2,cv_auc[3::4],'y--', label = 'CV Data of(500)')

#plotting the graph for the AUC
plt.xticks(x2, depth)
plt.ylim(0,1)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.title("Train and CV Data")
plt.xlabel("Depth")
plt.ylabel("AUC")
plt.show()

#heatmap for train_auc
df_heatmap_train_auc = pd. DataFrame(train_auc. reshape(7, 4), index=depth, columns=min_sample_spli
t)
fig = plt. figure(figsize=(16,5))
heatmap_train_auc = sns. heatmap(df_heatmap_train_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth', size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("Train Data", size=24)
plt. show()

#heatmap for cv_auc
df_heatmap_cv_auc = pd. DataFrame(cv_auc . reshape(7, 4), index=depth, columns=min_sample_split)
fig = plt. figure(figsize=(16,5))
heatmap_cv_auc = sns. heatmap(df_heatmap_cv_auc, annot=True, fmt='.4g')
plt.grid(True)
```

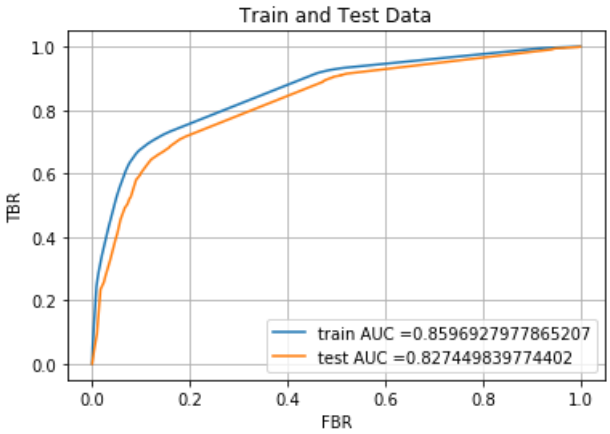
```
plt. ylabel('Depth' , size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("CV Data", size=24)
plt. show()

print("Optimal value of max_depth = ", opt_depth_tfidf , " Optimal min_samples_split is :",
opt_split_tfidf)

#Cv auc scores
print("---------------------------")
print("Cv auc scores")
print(cv_auc)
print("Maximun Auc value :",max(cv_auc))

#Verifying the model on Test data
clf = DecisionTreeClassifier(max_depth=opt_depth, min_samples_split=opt_split,class_weight=
'balanced')
clf.fit(train_tf_idf,Y_train)

train_fpr, train_tpr, thresholds = roc_curve(Y_train, clf.predict_proba(train_tf_idf)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, clf.predict_proba(test_tf_idf)[:,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(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.legend()
plt.xlabel("FBR")
plt.ylabel("TBR")
plt.title("Train and Test Data")
plt.show()
```








```
Optimal value of max_depth =  50  Optimal min_samples_split is : 500
--------------------------
Cv auc scores
[0.6459541  0.6459541  0.6459541  0.6459541  0.77636324 0.77636403
 0.77685901 0.77670814 0.80772168 0.80934011 0.81802085 0.82344792
 0.75919403 0.7658191  0.80621991 0.82710795 0.74408263 0.74817321
 0.78874815 0.80641187 0.72891543 0.73421533 0.76521201 0.77899722
 0.74574484 0.75053323 0.79137574 0.80683786]
Maximun Auc value : 0.827107950993988
```

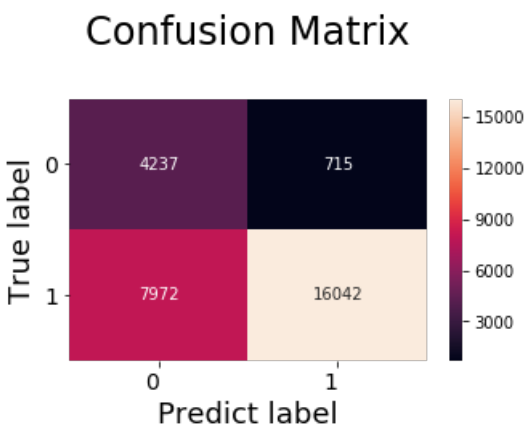


In [118]:

```
#Confusion Matrix

print("Train confusion matrix")
print(confusion_matrix(Y_train, clf.predict(train_tf_idf)))
print("Test confusion matrix")
print(confusion_matrix(Y_test, clf.predict(test_tf_idf)))

cm = confusion_matrix(Y_train, clf.predict(train_tf_idf))
cm = confusion_matrix(Y_test, clf.predict(test_tf_idf))
tn, fp, fn, tp = cm.ravel()
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

# Code for drawing seaborn heatmaps
class_names = ['0','1']
df_heatmap = pd.DataFrame(cm, index=class_names, columns=class_names )
fig = plt.figure(figsize=(5,3))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
plt.ylabel('True label',size=18)
plt.xlabel('Predict label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```

```
Train confusion matrix
```

```
[[ 5466   708]
 [10139 23087]]
Test confusion matrix
[[ 4237   715]
 [ 7972 16042]]
```



[5.2.1] Top 20 important features from SET 2

In [119]:

```
clf = DecisionTreeClassifier(max_depth= 50, min_samples_split=500,class_weight= 'balanced')
clf.fit(train_tf_idf,Y_train)
feat = tf_idf_vect.get_feature_names()
n=20
coefs = sorted(zip(clf.feature_importances_, feat))
top = coefs[:-(n + 1):-1]

print("Feature importances\tFeatures")
for (coef1, feat1) in top:
    print("%.4f\t\t\t%-15s" % (coef1, feat1))
```

```
Feature importances Features
0.1564   not
0.1226   great
0.0616   best
0.0540   delicious
0.0377   love
0.0362   good
0.0252   excellent
0.0211   perfect
0.0203   loves
0.0194   disappointed
0.0135   bad
0.0129   wonderful
0.0121   favorite
0.0097   not great
0.0093   easy
0.0092   not good
0.0092   yummy
0.0088   awesome
0.0078   nice
0.0071   tasty
```

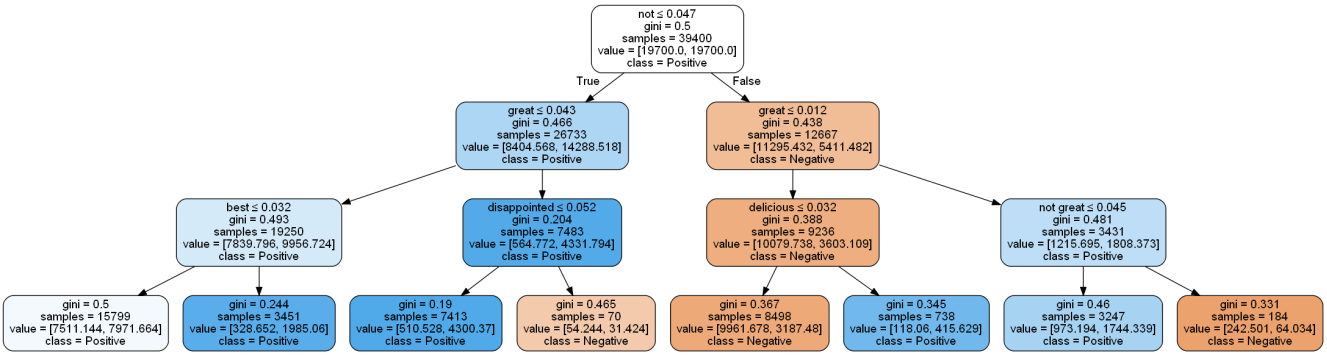
[5.2.2] Graphviz visualization of Decision Tree on TFIDF, SET 2

In [120]:

```
#install Graphviz
#set the environment variable's path
#install pydotplus
import pydot
clf = DecisionTreeClassifier(max_depth= 3, min_samples_split=500,class_weight= 'balanced')
clf.fit(train_tf_idf,Y_train)
names=tf_idf_vect.get_feature_names()
```

```
dot_data = StringIO()
tree.export_graphviz(clf, out_file=dot_data,feature_names=names,
                     class_names=['Negative','Positive'],
                     filled=True, rounded=True,
                     special_characters=True)
graph = pydot.graph_from_dot_data(dot_data.getvalue())[0]
Image(graph.create_png())
```

Out[120]:

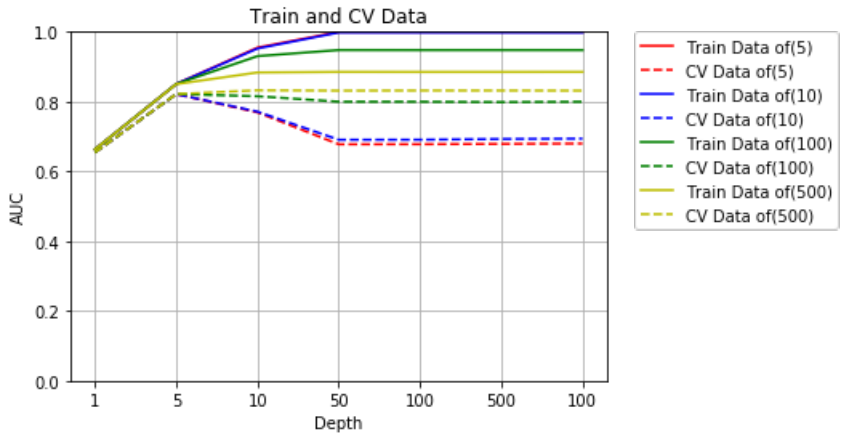


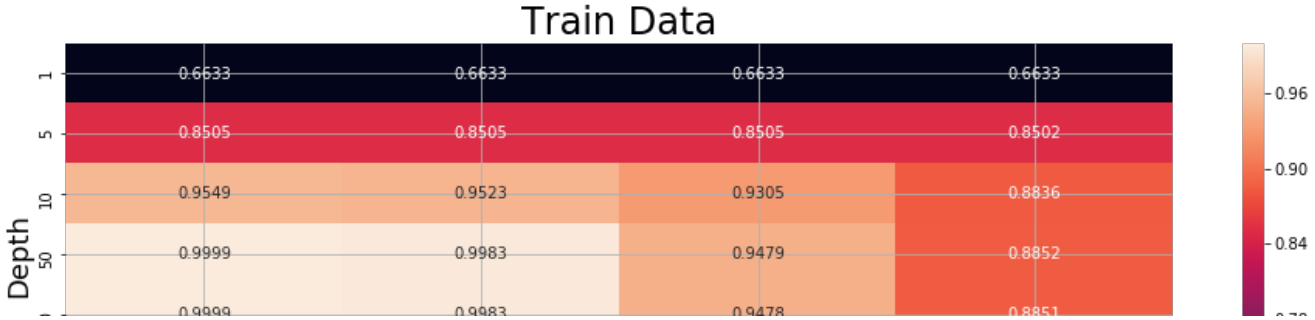
[5.3] Applying Decision Trees on AVG W2V, SET 3

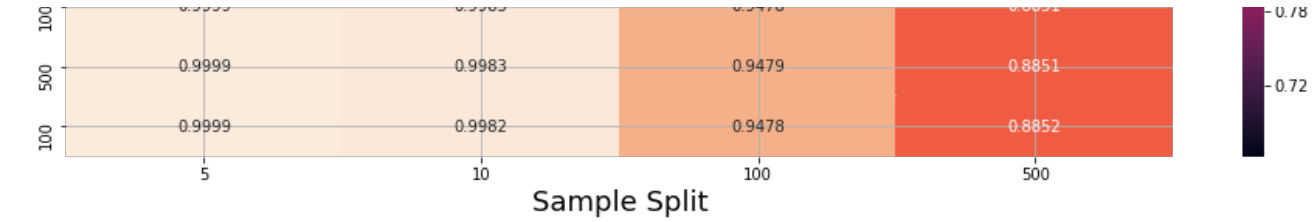
In [121]:

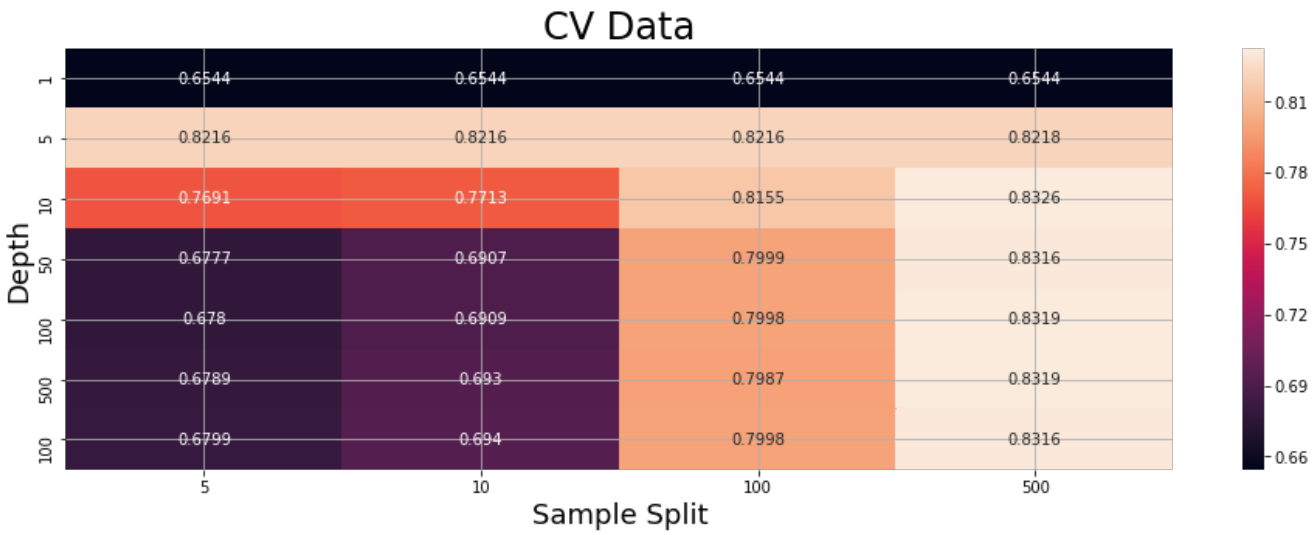
```
#given depth and min_samples_split and setting them as hyperparam
depth = [1, 5, 10, 50, 100, 500, 100]
min_sample_split = [5, 10, 100, 500]
hyper_param = {'max_depth':depth, 'min_samples_split':min_sample_split}

#declaring the classifier as DecisionTreeClassifier, fitting the classifier and finding the optimu
m values of depth and min_sample_split
clf = DecisionTreeClassifier(class_weight='balanced')
gsv = GridSearchCV(clf,hyper_param,scoring='roc_auc')
gsv.fit(sent_vectors_train,Y_train)
opt_depth_avgw2v, opt_split_avgw2v  = gsv.best_params_.get('max_depth'), gsv.best_params_.get('min_
samples_split')

#computing train _auc and cv_auc
train_auc= gsv.cv_results_['mean_train_score']
#train_auc_std= gsv.cv_results_['std_train_score']
cv_auc = gsv.cv_results_['mean_test_score']
#cv_auc_std= gsv.cv_results_['std_test_score']

#plotting the AUC accordingto depth values [5,10,100,500]
x2 = np.arange(len(depth))
plt.plot(x2,train_auc[::4],'r', label = 'Train Data of(5)')
plt.plot(x2,cv_auc[::4],'r--', label = 'CV Data of(5)')

plt.plot(x2,train_auc[1::4],'b', label = 'Train Data of(10)')
plt.plot(x2,cv_auc[1::4],'b--', label = 'CV Data of(10)')

plt.plot(x2,train_auc[2::4],'g', label = 'Train Data of(100)')
plt.plot(x2,cv_auc[2::4],'g--', label = 'CV Data of(100)')

plt.plot(x2,train_auc[3::4],'y', label = 'Train Data of(500)')
plt.plot(x2,cv_auc[3::4],'y--', label = 'CV Data of(500)')

#plotting the graph for the AUC
plt.xticks(x2, depth)
plt.ylim(0,1)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.title("Train and CV Data")
plt.xlabel("Depth")
plt.ylabel("AUC")
plt.show()

#heatmap for train_auc
df_heatmap_train_auc = pd. DataFrame(train_auc. reshape(7, 4), index=depth, columns=min_sample_spli
t)
```

```
fig = plt. figure(figsize=(16,5))
heatmap_train_auc = sns. heatmap(df_heatmap_train_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth', size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("Train Data", size=24)
plt. show()

#heatmap for cv_auc
df_heatmap_cv_auc = pd. DataFrame(cv_auc . reshape(7, 4), index=depth, columns=min_sample_split)
fig = plt. figure(figsize=(16,5))
heatmap_cv_auc = sns. heatmap(df_heatmap_cv_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth' , size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("CV Data", size=24)
plt. show()

print("Optimal value of max_depth = ", opt_depth_avgw2v , " Optimal min_samples_split is :", opt_sp
lit_avgw2v)

#Cv auc scores
print("--------------------------")
print("Cv auc scores")
print(cv_auc)
print("Maximun Auc value :",max(cv_auc))

#Verifying the model on Test data
clf = DecisionTreeClassifier(max_depth=opt_depth, min_samples_split=opt_split,class_weight=
'balanced')
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(sent_vectors_test)[:,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(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.legend()
plt.xlabel("FBR")
plt.ylabel("TBR")
plt.title("Train and Test Data")
plt.show()
```

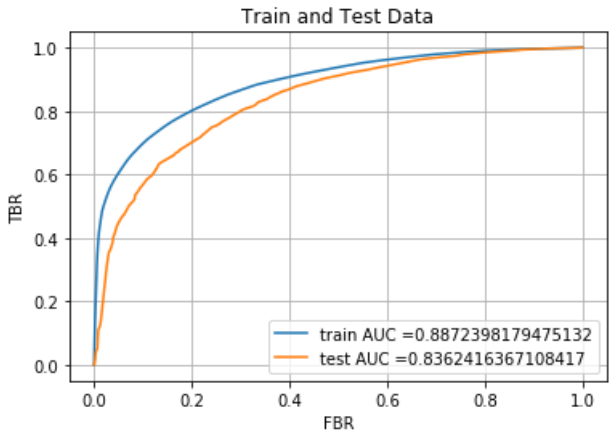









```
Optimal value of max_depth =  10  Optimal min_samples_split is : 500
---------------------------
Cv auc scores
[0.65444389 0.65444389 0.65444389 0.65444389 0.82159954 0.82159954
 0.82159954 0.82182527 0.76911792 0.77130575 0.81547157 0.83259328
 0.67773849 0.69073149 0.7998604  0.83160646 0.67798265 0.69089492
 0.79977838 0.83194043 0.67888546 0.6929659  0.7986971  0.8319169
 0.67989072 0.69402055 0.79975244 0.83162999]
Maximun Auc value : 0.8325932753896986
```

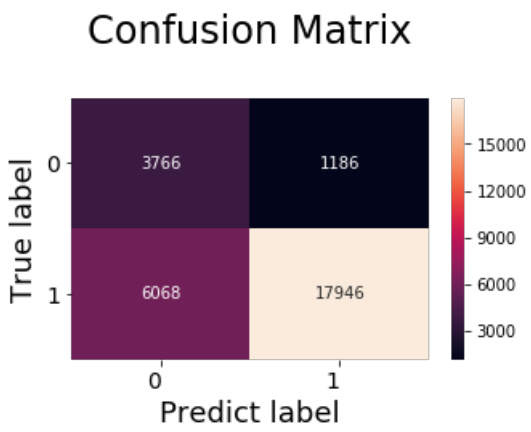


In [122]:

```
#Confusion Matrix

print("Train confusion matrix")
print(confusion_matrix(Y_train, clf.predict(sent_vectors_train)))
print("Test confusion matrix")
print(confusion_matrix(Y_test, clf.predict(sent_vectors_test)))

cm = confusion_matrix(Y_train, clf.predict(sent_vectors_train))
cm = confusion_matrix(Y_test, clf.predict(sent_vectors_test))
tn, fp, fn, tp = cm.ravel()
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

# Code for drawing seaborn heatmaps
class_names = ['0','1']
df_heatmap = pd.DataFrame(cm, index=class_names, columns=class_names )
```

```
fig = plt.figure(figsize=(5,3))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
plt.ylabel('True label',size=18)
plt.xlabel('Predict label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```

```
Train confusion matrix
[[ 5203   971]
 [ 7801 25425]]
Test confusion matrix
[[ 3766  1186]
 [ 6068 17946]]
```

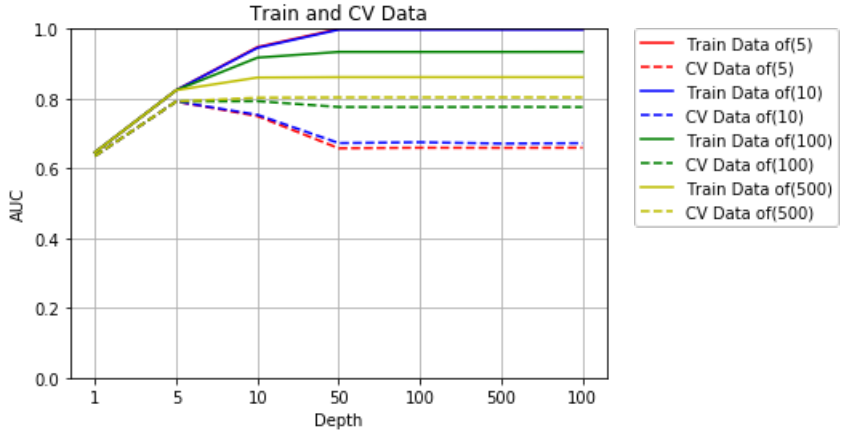


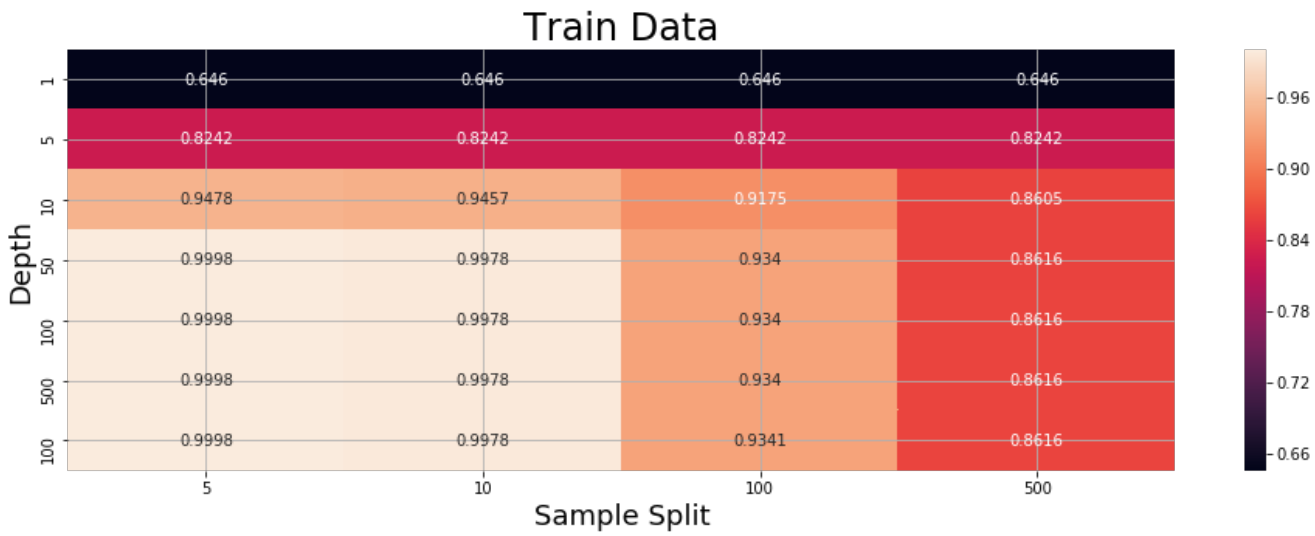
[5.4] Applying Decision Trees on TFIDF W2V, SET 4

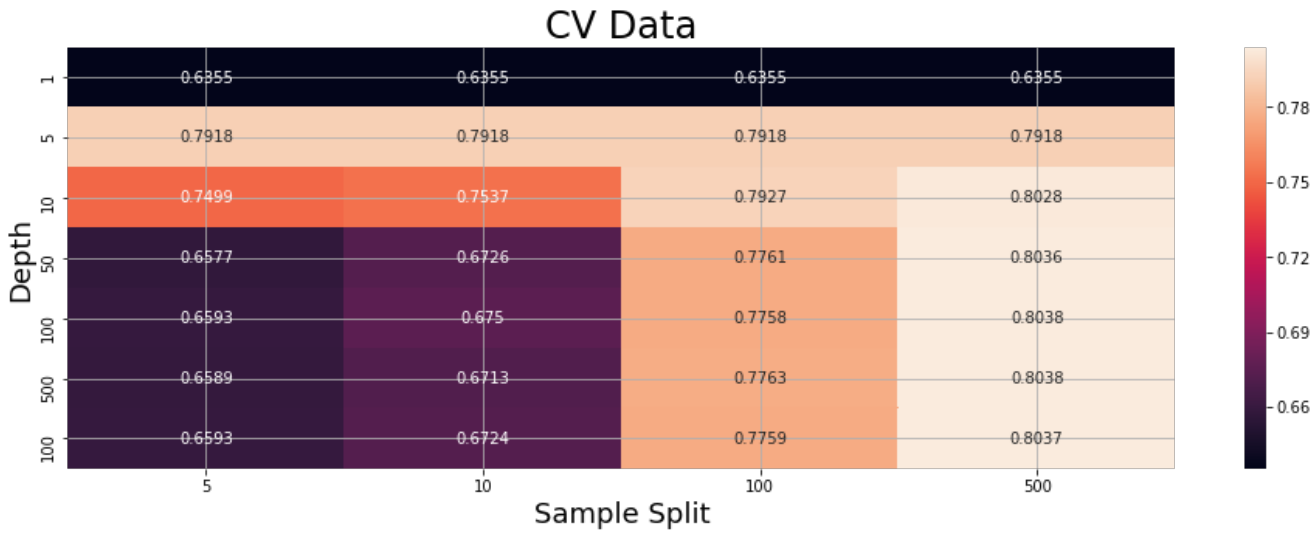
In [123]:

```
#given depth and min_samples_split and setting them as hyperparam
depth = [1, 5, 10, 50, 100, 500, 100]
min_sample_split = [5, 10, 100, 500]
hyper_param = {'max_depth':depth, 'min_samples_split':min_sample_split}

#declaring the classifier as DecisionTreeClassifier, fitting the classifier and finding the optimu
m values of depth and min_sample_split
clf = DecisionTreeClassifier(class_weight='balanced')
gsv = GridSearchCV(clf,hyper_param,scoring='roc_auc')
gsv.fit(tfidf_sent_vectors_train,Y_train)
opt_depth_tfidf_w2v, opt_split_tfidf_w2v  = gsv.best_params_.get('max_depth'), gsv.best_params_.get
('min_samples_split')

#computing train _auc and cv_auc
train_auc= gsv.cv_results_['mean_train_score']
#train_auc_std= gsv.cv_results_['std_train_score']
cv_auc = gsv.cv_results_['mean_test_score']
#cv_auc_std= gsv.cv_results_['std_test_score']

#plotting the AUC accordingto depth values [5,10,100,500]
x2 = np.arange(len(depth))
plt.plot(x2,train_auc[::4],'r', label = 'Train Data of(5)')
plt.plot(x2,cv_auc[::4],'r--', label = 'CV Data of(5)')

plt.plot(x2,train_auc[1::4],'b', label = 'Train Data of(10)')
plt.plot(x2,cv_auc[1::4],'b--', label = 'CV Data of(10)')

plt.plot(x2,train_auc[2::4],'g', label = 'Train Data of(100)')
plt.plot(x2,cv_auc[2::4],'g--', label = 'CV Data of(100)')

plt.plot(x2,train_auc[3::4],'y', label = 'Train Data of(500)')
plt.plot(x2,cv_auc[3::4],'y--', label = 'CV Data of(500)')

#plotting the graph for the AUC
```

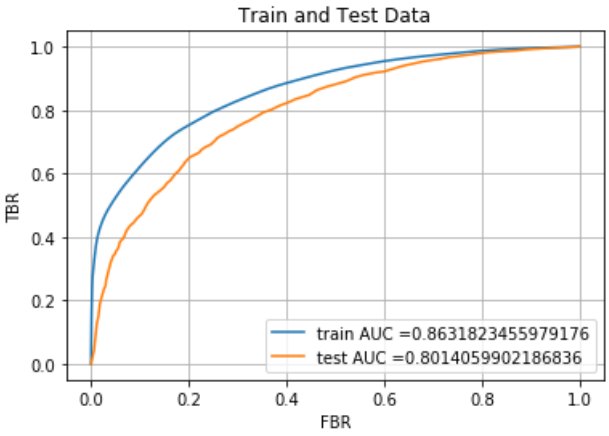
```
plt.xticks(x2, depth)
plt.ylim(0,1)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.title("Train and CV Data")
plt.xlabel("Depth")
plt.ylabel("AUC")
plt.show()

#heatmap for train_auc
df_heatmap_train_auc = pd. DataFrame(train_auc. reshape(7, 4), index=depth, columns=min_sample_spli
t)
fig = plt. figure(figsize=(16,5))
heatmap_train_auc = sns. heatmap(df_heatmap_train_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth', size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("Train Data", size=24)
plt. show()

#heatmap for cv_auc
df_heatmap_cv_auc = pd. DataFrame(cv_auc . reshape(7, 4), index=depth, columns=min_sample_split)
fig = plt. figure(figsize=(16,5))
heatmap_cv_auc = sns. heatmap(df_heatmap_cv_auc, annot=True, fmt='.4g')
plt.grid(True)
plt. ylabel('Depth' , size=18)
plt. xlabel('Sample Split' , size=18)
plt. title("CV Data", size=24)
plt. show()

print("Optimal value of max_depth = ", opt_depth_tfidf_w2v , " Optimal min_samples_split is :", opt
_split_tfidf_w2v)

#Cv auc scores
print("--------------------------")
print("Cv auc scores")
print(cv_auc)
print("Maximun Auc value :",max(cv_auc))

#Verifying the model on Test data
clf = DecisionTreeClassifier(max_depth=opt_depth, min_samples_split=opt_split,class_weight=
'balanced')
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(tfidf_sent_vectors_test)[:,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(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.grid(True)
plt.legend()
plt.xlabel("FBR")
plt.ylabel("TBR")
plt.title("Train and Test Data")
plt.show()
```








```
Optimal value of max_depth =  500  Optimal min_samples_split is : 500
--------------------------
Cv auc scores
[0.63552519 0.63552519 0.63552519 0.63552519 0.79180301 0.79180301
 0.79180301 0.79180301 0.79180301 0.74990476 0.75366597 0.79267136 0.80279995
 0.65774189 0.67256287 0.7761017  0.80363203 0.65927336 0.67496945
 0.77578708 0.80376266 0.65886871 0.67131852 0.77631291 0.80380118
 0.6593185  0.67241343 0.77592835 0.80372242]
Maximun Auc value : 0.803801182585852
```

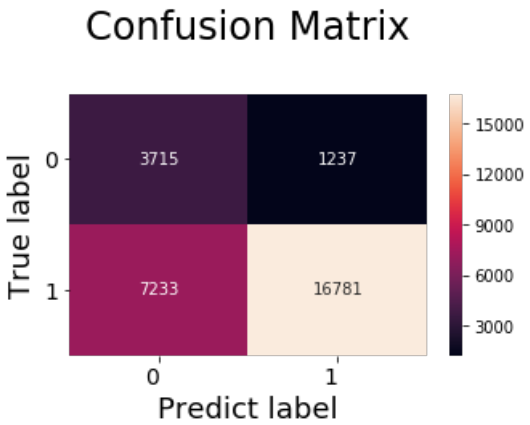


In [124]:

```
#Confusion Matrix
```

```
print("Train confusion matrix")
print(confusion_matrix(Y_train, clf.predict(tfidf_sent_vectors_train)))
print("Test confusion matrix")
print(confusion_matrix(Y_test, clf.predict(tfidf_sent_vectors_test)))

cm = confusion_matrix(Y_train, clf.predict(tfidf_sent_vectors_train))
cm = confusion_matrix(Y_test, clf.predict(tfidf_sent_vectors_test))
tn, fp, fn, tp = cm.ravel()
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

# Code for drawing seaborn heatmaps
class_names = ['0','1']
df_heatmap = pd.DataFrame(cm, index=class_names, columns=class_names )
fig = plt.figure(figsize=(5,3))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14)
plt.ylabel('True label',size=18)
plt.xlabel('Predict label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```

```
Train confusion matrix
[[ 5097  1077]
 [ 9005 24221]]
Test confusion matrix
[[ 3715  1237]
 [ 7233 16781]]
```



[6] Conclusions

In [125]:

```
from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["S.NO","MODEL", "DEPTH","SAMPLE SPLIT"]

auc = [0.83,0.82,0.83,0.80]

x.add_row(["1","BAG OF WORDS",opt_depth_bow, opt_split_bow])
x.add_row(["2","TFIDF",opt_depth_tfidf, opt_split_tfidf])
x.add_row(["3","AVG W2V",opt_depth_avgw2v, opt_split_avgw2v])
x.add_row(["4","TFIDF W2V",opt_depth_tfidf_w2v, opt_split_tfidf_w2v])

x.add_column("AUC",auc)

# Printing the Table
print(x)
```

```
+------+--------------+-------+--------------+------+
| S.NO |    MODEL     | DEPTH | SAMPLE SPLIT | AUC  |
```

1	BAG OF WORDS	10	500	0.83
2	TFIDF	50	500	0.82
3	AVG W2V	10	500	0.83
4	TFIDF W2V	500	500	0.8

Observations

1. Bag of Words and AvgW2v gave the same AUC which is the highest and the least AUC is given by TFIDFW2V.
2. The optimum depth kept changing, but is the same for Bag of Words and AVGW2V andthe maximum depth is for TFIDF W2V.
3. Here we have seen with 100k datapoints, with more datapoints our auc might increase.