Amazon Fine Food Reviews Analysis

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

EDA: https://nycdatascience.com/blog/student-works/amazon-fine-foods-visualization/ (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 [0]:
```
%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 [151]:
```
from google.colab import drive


drive.mount('/content/drive')
```

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

In [152]:
```
%cd ./drive/My Drive
```

```
[Errno 2] No such file or directory: './drive/My Drive'
/content/drive/My Drive
```

In [153]:
```
# using SQLite Table to read data.
con = sqlite3.connect('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 da
ta 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""", con)
# for tsne assignment you can take 5k data points

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 L
IMIT 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[153]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulne
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1
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 [155]:
```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[155]:

	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 [156]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[156]:

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

In [157]:
```
display['COUNT(*)'].sum()
```

Out[157]: 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 [158]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND UserId="AR5J8UI46CURR"
ORDER BY ProductID
""", con)
display.head()
```

Out[158]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpful
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, inp
lace=False, kind='quicksort', na_position='last')
```

In [160]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"
}, keep='first', inplace=False)
final.shape
```

Out[160]: (87775, 10)

In [161]:
```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[161]: 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 [162]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)

display.head()
```

Out[162]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpful
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 [164]:
```
#Before starting the next phase of preprocessing lets see the number of entrie
s left
print(final.shape)

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

```
(87773, 10)
```

Out[164]:
```
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 [165]:
```
# 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 very hard to find any chicken products made in the USA but t
hey are out there, but this one isnt.  Its too bad too because its a good pro
duct but I wont take any chances till they know what is going on with the chi
na 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 an
d 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 grea
t 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 retai
ler. You can buy a 1 pound bag on Amazon for almost the same price as a 6 oun
ce bag at other retailers. It's definitely worth it to buy a big bag if your
dog eats them a lot.
==================================================
```

In [166]:
```
# 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 very hard to find any chicken products made in the USA but t
hey are out there, but this one isnt.  Its too bad too because its a good pro
duct but I wont take any chances till they know what is going on with the chi
na imports.
```

In [167]:
```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-rem
ove-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 very hard to find any chicken products made in the USA but t
hey are out there, but this one isnt.  Its too bad too because its a good pro
duct but I wont take any chances till they know what is going on with the chi
na 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 an
d 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 grea
t 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 retai
ler. You can buy a 1 pound bag on Amazon for almost the same price as a 6 oun
ce bag at other retailers. It's definitely worth it to buy a big bag if your
dog eats them a lot.
```

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 [169]:
```
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 [170]:
```
#remove words with numbers python: https://stackoverflow.com/a/18082370/408403
9
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 very hard to find any chicken products made in the USA but t
hey are out there, but this one isnt.  Its too bad too because its a good pro
duct but I wont take any chances till they know what is going on with the chi
na imports.
```

In [171]:
```
#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 [0]:
```
# 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', "you'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', 'it
self', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 't
hat', "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', 'becau
se', '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', 'a
ll', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'tha
n', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "shoul
d'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', "isn't", 'm
a', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shoul
dn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

In [173]:
```
# 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 i
n stopwords)
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 87773/87773 [00:41<00:00, 2097.32it/s]
```

In [174]:
```
preprocessed_reviews[1500]
```

Out[174]:
```
'way hot blood took bite jig lol'
```

[3.2] Preprocessing Review Summary

In [0]:
```
## Similartly you can do preprocessing for review summary also.
```

[4] Featurization

[4.1] BAG OF WORDS

In [177]:
```
#BoW
count_vect = CountVectorizer() #in scikit-learn
count_vect.fit(preprocessed_reviews)
print("some feature names ", count_vect.get_feature_names()[:10])
print('='*50)

final_counts = count_vect.transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_counts))
print("the shape of out text BOW vectorizer ",final_counts.get_shape())
print("the number of unique words ", final_counts.get_shape()[1])
```

```
some feature names  ['aa', 'aaa', 'aaaa', 'aaaaa', 'aaaaaaaaaaaa', 'aaaaaaaaa
aaaaaa', 'aaaaaaahhhhhh', 'aaaaaaarrrrrggghhh', 'aaaaaawwwwwwwwww', 'aaaaah']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (87773, 54904)
the number of unique words  54904
```

[4.2] Bi-Grams and n-Grams.

In [178]:
```
#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/sta
ble/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_shape()[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 [0]:
```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(preprocessed_reviews)
print("some sample features(unique words in the corpus)",tf_idf_vect.get_featu
re_names()[0:10])
print('='*50)

final_tf_idf = tf_idf_vect.transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_tf_idf))
print("the shape of out text TFIDF vectorizer ",final_tf_idf.get_shape())
print("the number of unique words including both unigrams and bigrams ", final
_tf_idf.get_shape()[1])
```

[4.4] Word2Vec

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

In [0]:
```
# 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/#.W17SRFAzZP
Y
# you can comment this whole cell
# 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(list_of_sentance,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-negati
ve300.bin', binary=True)
        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 ")
```

In [0]:
```
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])
```

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

[4.4.1.1] Avg W2v

In [0]:
```
"""# average Word2Vec
# compute average word2vec for each review.
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this li
st
for sent in tqdm(list_of_sentance): # 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.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))"""
```

[4.4.1.2] TFIDF weighted W2v

In [0]:
```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = model.fit_transform(preprocessed_reviews)
# 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 [0]:
```
"""# 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 = []; # the tfidf-w2v for each sentence/review is stored in
this list
row=0;
for sent in tqdm(list_of_sentance): # 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.append(sent_vec)
    row += 1"""
```

[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 (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/receiver-operating-characteristic-curve-roc-curve-and-auc-1/) 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 (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html) 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 (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/confusion-matrix-tpr-fpr-fnr-tnr-1/) with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps. (https://seaborn.pydata.org/generated/seaborn.heatmap.html) (https://seaborn.pydata.org/generated/seaborn.heatmap.html)

(https://seaborn.pydata.org/generated/seaborn.heatmap.html)

(https://seaborn.pydata.org/generated/seaborn.heatmap.html)

7. **Conclusion** (https://seaborn.pydata.org/generated/seaborn.heatmap.html)

   (https://seaborn.pydata.org/generated/seaborn.heatmap.html)
   - 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 (https://seaborn.pydata.org/generated/seaborn.heatmap.html) link (http://zetcode.com/python/prettytable/)

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. (https://soundcloud.com/applied-ai-course/leakage-bow-and-tfidf)

Applying Decision Trees

[5.1] Applying Decision Trees on BOW, SET 1

In [0]:
```
from sklearn.model_selection import train_test_split
#Loading the data into X,Y
X = preprocessed_reviews
Y = final['Score'].values

#splitting the data into Train,Cross validate and test
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.
25)
```

In [180]:
```
from sklearn.model_selection import learning_curve
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.tree import DecisionTreeClassifier

vectorizer = CountVectorizer(min_df = 10)
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)
```

```
After vectorizations
(52663, 9014) (52663,)
(17555, 9014) (17555,)
(17555, 9014) (17555,)
================================================================================
=======================
```

Hyper parameter Tuning

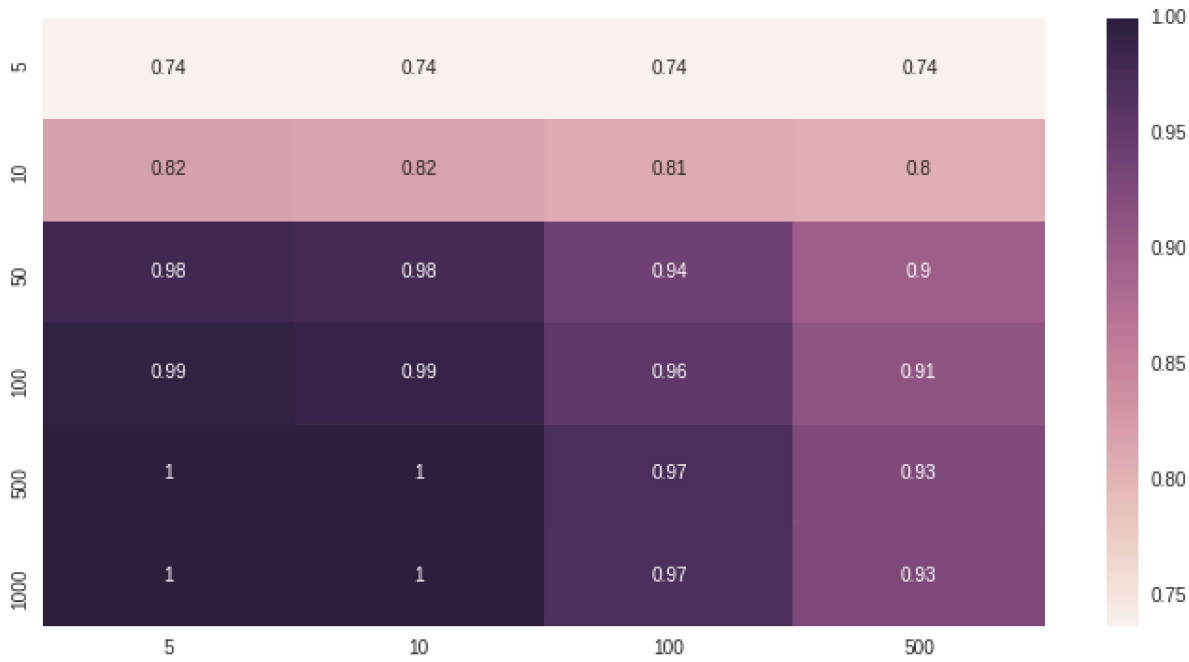
In [181]:
```
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt


#Hyper parameter tuning
train_auc = []
cv_auc = []
C=[ 5, 10, 50, 100, 500, 1000]#best_depth
D=[5, 10, 100, 500]#min_sample_splits
#For loop to find train_auc,cv_auc
for i in C:
  for j in D:
    clf = DecisionTreeClassifier(max_depth=i,min_samples_split=j,class_weight=
'balanced')
    clf.fit(X_train_bow,y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability e
stimates 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(cv_auc)
```

```
[0.7369685336410439, 0.7369010165876181, 0.7369042445434535, 0.73737806672987
4, 0.777631571970798, 0.7792677037299164, 0.7870441439038913, 0.7893485729833
205, 0.7015610738081062, 0.7047306195986356, 0.7722483971665077, 0.8157575720
234272, 0.7210614986709167, 0.7242065586367629, 0.7846918232515722, 0.8166533
604517627, 0.7080257611004369, 0.7096571920721602, 0.7616890049392879, 0.7949
878854694759, 0.7098371352679924, 0.7120919176502563, 0.7632765332569293, 0.7
932298155307929]
```

In [182]:
```
import numpy as np
train_auc=np.array(train_auc)
cv_auc = np.array(cv_auc)

#Train_auc Heatmap
print("Train_AUC Heatmap")

scores = train_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

```
Train_AUC Heatmap
```

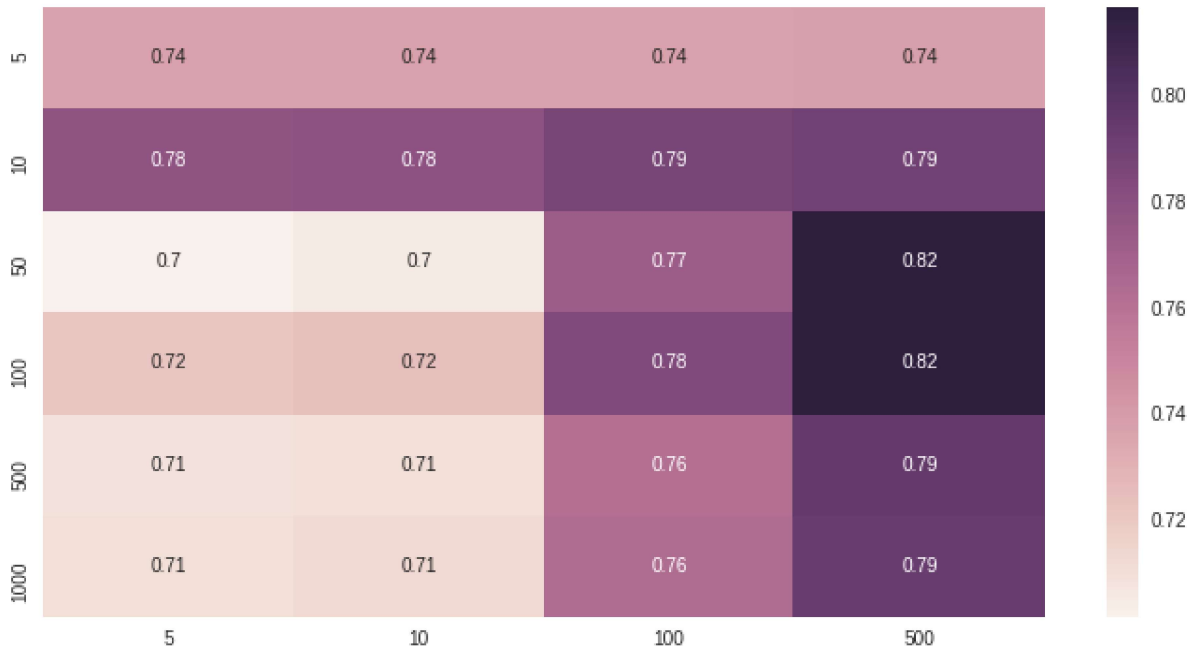
Out[182]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6449bc1f28>



In [184]:
```
print("CV_AUC Heatmap")

#Hyper parameter tuning using seaborn heatmap
scores = cv_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

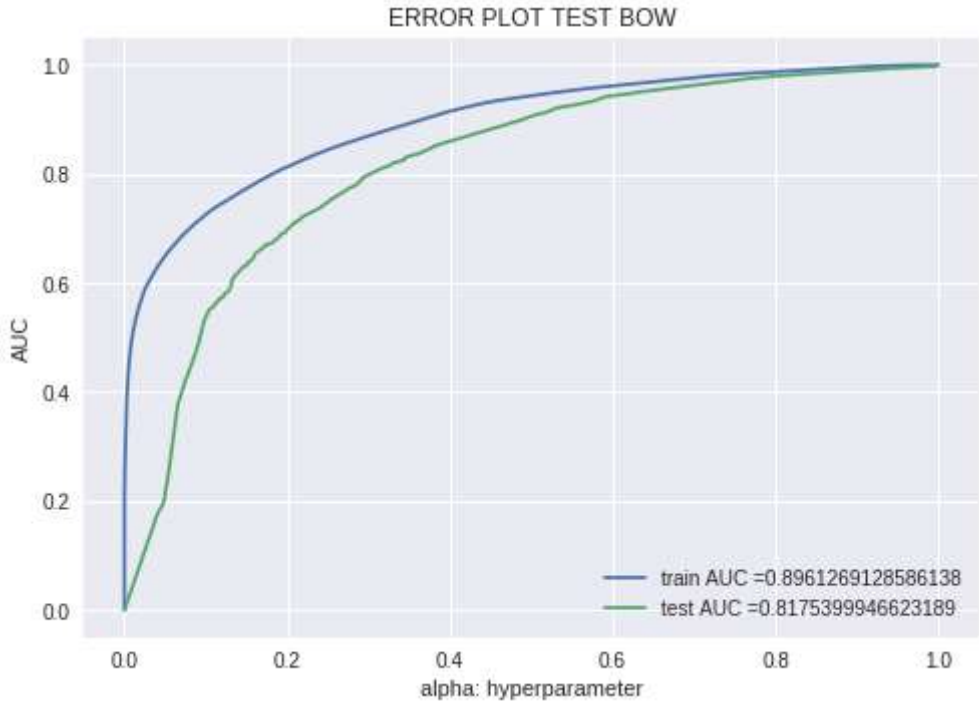
```
CV_AUC Heatmap
```

Out[184]: <matplotlib.axes._subplots.AxesSubplot at 0x7f644a7d11d0>



Testing with test data

In [186]:
```
#Best hyperparameters are max_depth=50,min_samples_split=100

from sklearn.metrics import roc_curve, auc


clf = DecisionTreeClassifier(max_depth=50,min_samples_split=500,class_weight=
'balanced')
clf.fit(X_train_bow,y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estim
ates of the positive class
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(X_trai
n_bow)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test_bo
w)[:,1])

#plotting the results
plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tp
r)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOT TEST BOW")
plt.show()

print("="*100)

#printing confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns;sns.set()
print("Train confusion matrix")

#printing train_confusion matrix using pandas dataframe and seaborn

df_train=pd.DataFrame(confusion_matrix(y_train, clf.predict(X_train_bow)))
sns.heatmap(df_train,annot=True,fmt="d")
plt.show()

#printing train_confusion matrix using pandas dataframe and seaborn
print("Test confusion matrix")

df_test=pd.DataFrame(confusion_matrix(y_test,clf.predict(X_test_bow)))
sns.heatmap(df_test,annot=True,fmt="d")
plt.show()
```




```
==============================================================================
========================
Train confusion matrix
```

```
Test confusion matrix
```

[5.1.1] Top 20 important features from SET 1

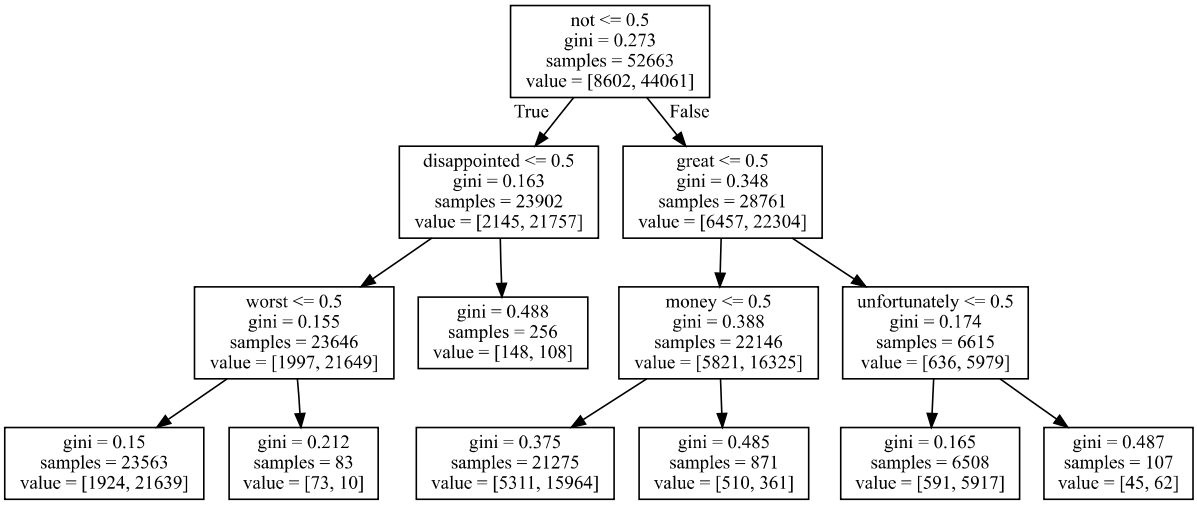
In [187]:
```
#Finding feature importance values and sorting them according to their index
pos=np.argsort((clf.feature_importances_))[::-1][0:20]

#Mapping the top 20 positive  values against respective feature names
print(np.take(vectorizer.get_feature_names(), pos))
```

```
['not' 'great' 'best' 'delicious' 'love' 'perfect' 'good' 'disappointed'
 'loves' 'bad' 'excellent' 'nice' 'money' 'wonderful' 'away' 'worst'
 'favorite' 'terrible' 'awful' 'disappointing']
```

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

In [188]:
```
#Training DecisionTreeClassifier with depth 3
clf = DecisionTreeClassifier(max_depth=3,min_samples_split=500)
clf.fit(X_train_bow,y_train)

import graphviz
from sklearn import tree
#Using graphviz to plot the graph using clf(Decision tree classifier)
data = tree.export_graphviz(clf,out_file=None,feature_names=vectorizer.get_fea
ture_names())
graph = graphviz.Source(data)
graph
```

Out[188]:



[5.2] Applying Decision Trees on TFIDF, SET 2

In [189]:
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(X_train) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_tf = tf_idf_vect.transform(X_train)
X_cv_tf=tf_idf_vect.transform(X_cv)
X_test_tf = tf_idf_vect.transform(X_test)


print("After vectorizations")
print(X_train_tf.shape, y_train.shape)
print(X_cv_tf.shape, y_cv.shape)
print(X_test_tf.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(52663, 31186) (52663,)
(17555, 31186) (17555,)
(17555, 31186) (17555,)
==============================================================================
=======================
```

Hyper Parameter Tuning

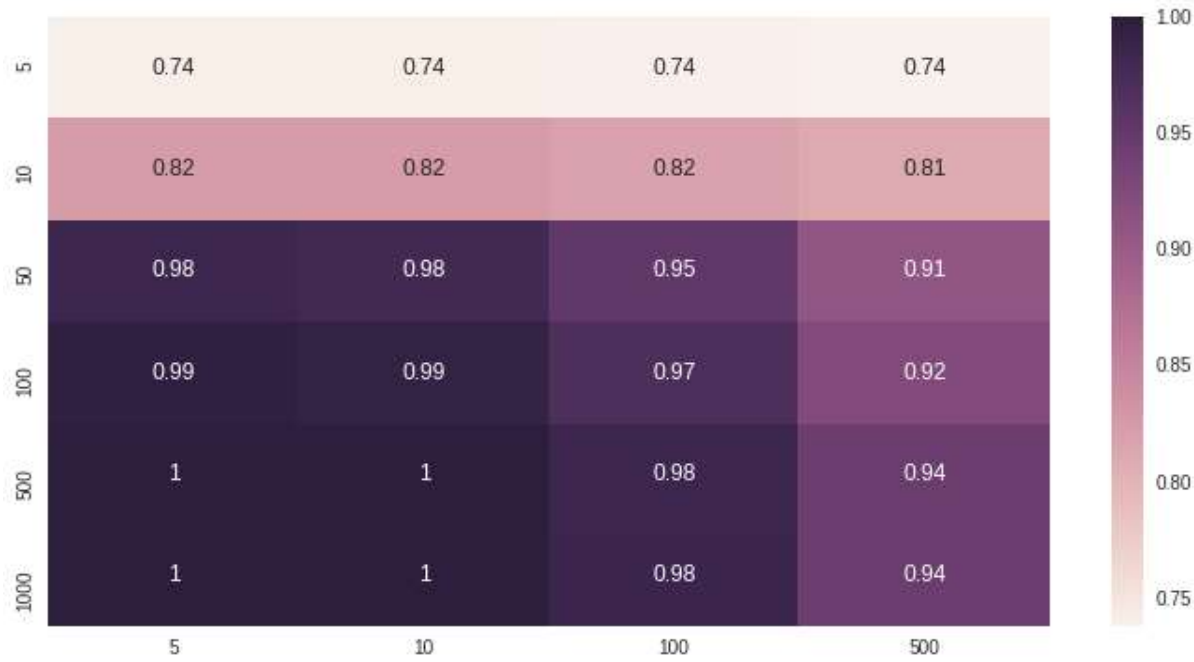
In [190]:
```
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt


#Hyper parameter tuning
train_auc = []
cv_auc = []
C=[ 5, 10, 50, 100, 500, 1000]#best_depth
D=[5, 10, 100, 500]#min_sample_splits
#For loop to find train_auc,cv_auc
for i in C:
  for j in D:
    clf = DecisionTreeClassifier(max_depth=i,min_samples_split=j,class_weight=
'balanced')
    clf.fit(X_train_tf,y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability e
stimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(X_train_tf)[:,1]
    y_cv_pred =  clf.predict_proba(X_cv_tf)[:,1]

    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

print(cv_auc)
```

```
[0.7352993981813868, 0.7352993981813868, 0.7360230985155428, 0.73589854605995
96, 0.7782818025573067, 0.7780278004812035, 0.7821977057408225, 0.78739014885
81038, 0.7067097983752508, 0.7163970656678643, 0.7581376030196582, 0.81097450
24382234, 0.7105941461425953, 0.7217630083429035, 0.7619300951920494, 0.80313
47059599798, 0.7017041839413445, 0.7066348680696402, 0.746294711729552, 0.780
6591981669335, 0.7042367055350188, 0.7086733160429454, 0.7431754140424081, 0.
7791260541546418]
```

In [191]:
```
import numpy as np
train_auc=np.array(train_auc)
cv_auc = np.array(cv_auc)

#Train_auc Heatmap
print("Train_AUC Heatmap")

scores = train_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

```
Train_AUC Heatmap
```

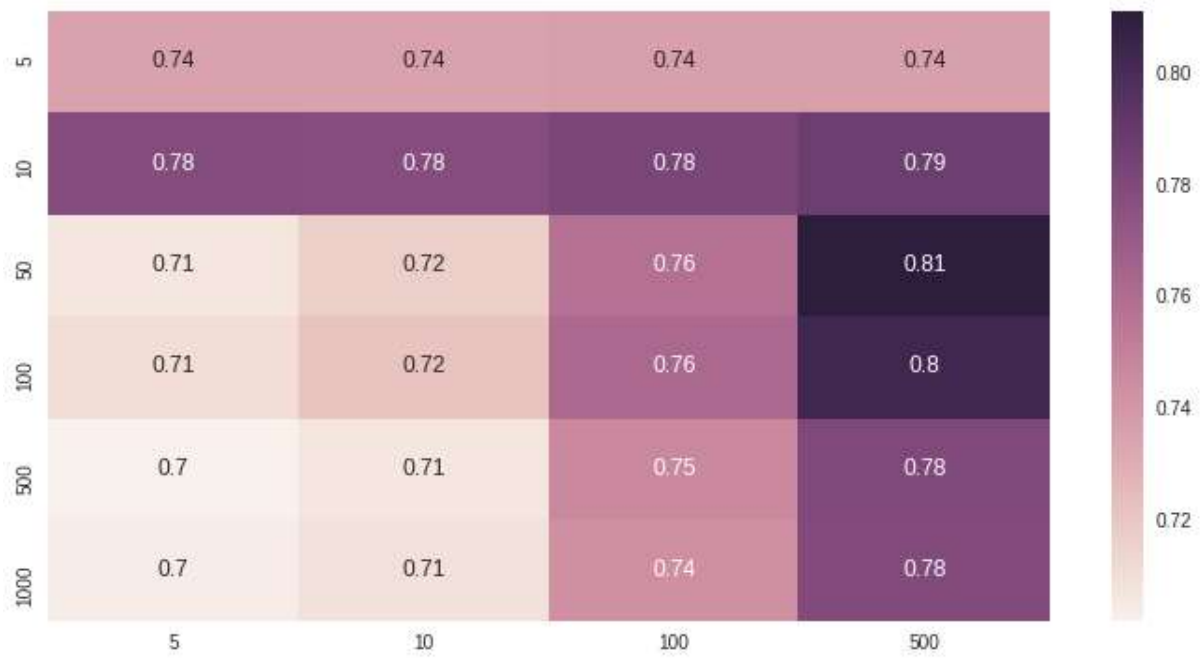
Out[191]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6449227710>



In [193]:
```
print("CV_AUC Heatmap")

#Hyper parameter tuning using seaborn heatmap
scores = cv_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

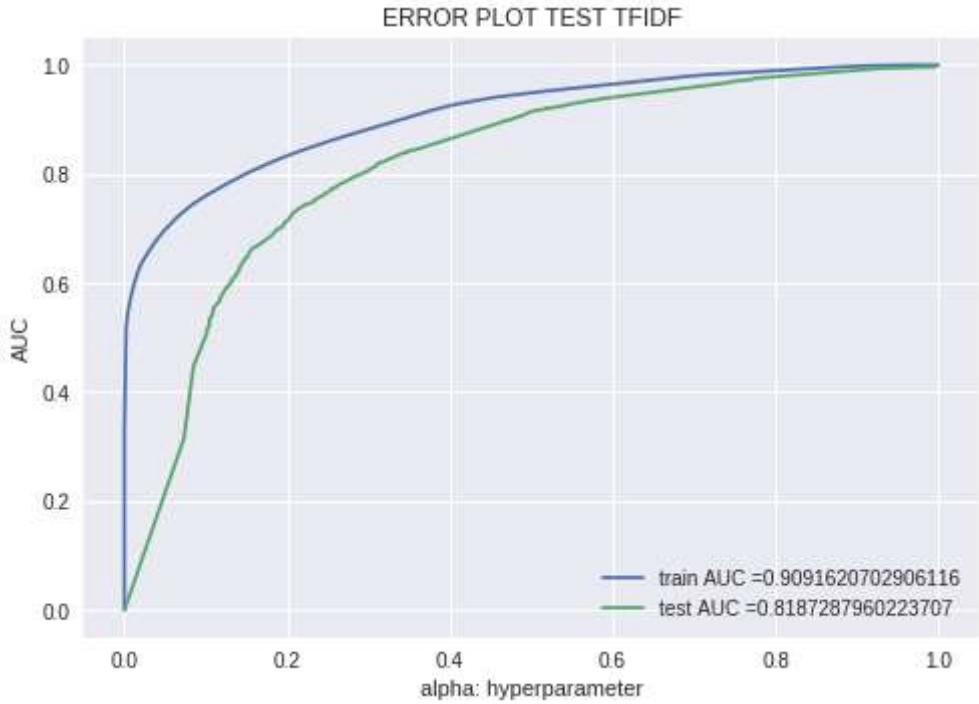
```
CV_AUC Heatmap
```

Out[193]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f645cf67550>
```



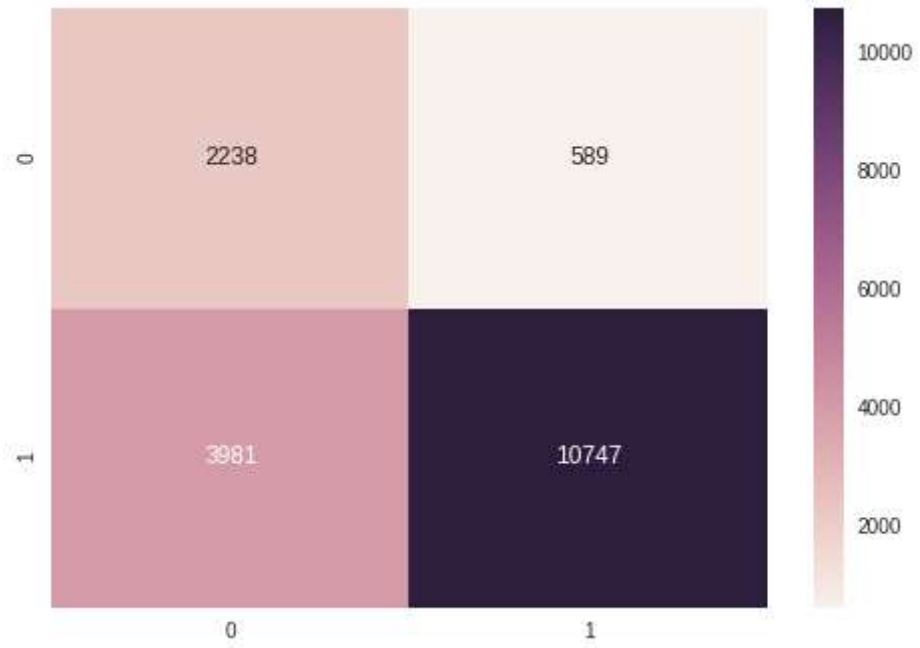
Testing with Test Data

In [194]:
```
#Best hyperparameters are max_depth=100,min_samples_split=100

from sklearn.metrics import roc_curve, auc


clf = DecisionTreeClassifier(max_depth=50,min_samples_split=500,class_weight=
'balanced')
clf.fit(X_train_tf,y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estim
ates of the positive class
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(X_trai
n_tf)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test_tf
)[:,1])

#plotting the results
plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tp
r)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOT TEST TFIDF")
plt.show()

print("="*100)

#printing confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns;sns.set()
print("Train confusion matrix")

#printing train_confusion matrix using pandas dataframe and seaborn

df_train=pd.DataFrame(confusion_matrix(y_train, clf.predict(X_train_tf)))
sns.heatmap(df_train,annot=True,fmt="d")
plt.show()

#printing train_confusion matrix using pandas dataframe and seaborn
print("Test confusion matrix")

df_test=pd.DataFrame(confusion_matrix(y_test,clf.predict(X_test_tf)))
sns.heatmap(df_test,annot=True,fmt="d")
plt.show()
```




```
================================================================================
========================
Train confusion matrix
```

```
Test confusion matrix
```



[5.2.1] Top 20 important features from SET 2

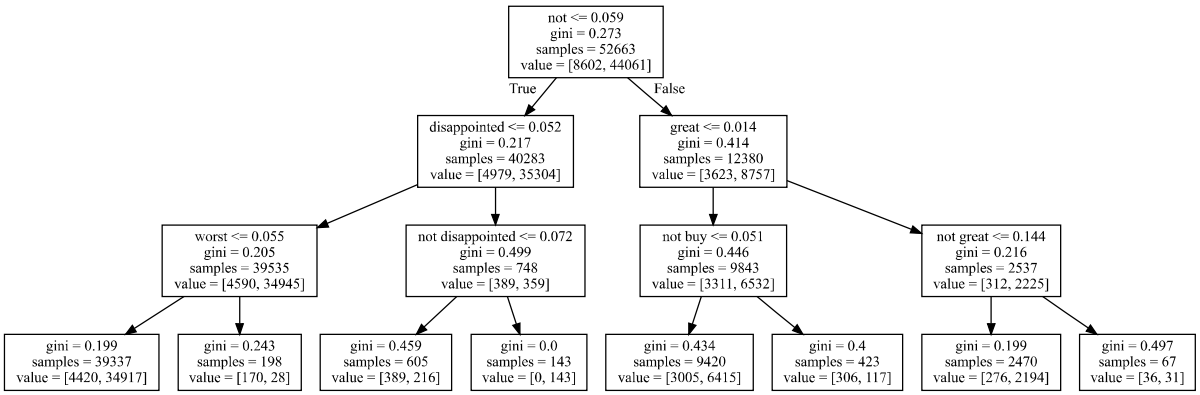
In [195]:
```
#Finding feature importance values and sorting them according to their index
pos=np.argsort((clf.feature_importances_))[::-1][0:20]

#Mapping the top 20 positive  values against respective feature names
print(np.take(tf_idf_vect.get_feature_names(), pos))
```

```
['not' 'great' 'best' 'delicious' 'love' 'good' 'disappointed' 'perfect'
 'bad' 'loves' 'nice' 'money' 'excellent' 'wonderful' 'away' 'thought'
 'favorite' 'taste' 'product' 'unfortunately']
```

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

In [196]:
```
#Training DecisionTreeClassifier with depth 3
clf = DecisionTreeClassifier(max_depth=3,min_samples_split=500)
clf.fit(X_train_tf,y_train)

import graphviz
from sklearn import tree
#Using graphviz to plot the graph using clf(Decision tree classifier)
data = tree.export_graphviz(clf, out_file=None,feature_names=tf_idf_vect.get_f
eature_names())
graph = graphviz.Source(data)
graph
```

Out[196]:



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

In [197]:
```
i=0
list_of_sentance_train=[]
for sentance in X_train:
    list_of_sentance_train.append(sentance.split())

# Training W2V 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])
```

```
number of words that occured minimum 5 times  13833
sample words  ['love', 'tea', 'glad', 'amazon', 'carries', 'hard', 'find', 's
tores', 'one', 'place', 'know', 'way', 'expensive', 'buying', 'online', 'grea
t', 'herbal', 'strongly', 'cinnamon', 'naturally', 'sweet', 'not', 'need', 'a
dd', 'sweetner', 'also', 'bags', 'go', 'long', 'use', 'large', 'saucepan', 'f
ull', 'water', 'taste', 'somewhat', 'addicting', 'regular', 'boring', 'drinki
ng', 'reluctant', 'try', 'product', 'reviewing', 'reviews', 'people', 'though
t', 'heck', 'much', 'satisfy']
```

In [198]:
```
from tqdm import tqdm
import numpy as np

# converting train data text

# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train = []; # the avg-w2v for each sentence/review is stored in t
his 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, 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_train.append(sent_vec)
sent_vectors_train = np.array(sent_vectors_train)
print(sent_vectors_train.shape)
print(sent_vectors_train[0])
```

```
100%|██████████| 52663/52663 [01:37<00:00, 540.63it/s]

(52663, 50)
[ 0.69288093  0.76653135 -0.72746875  0.34054917  0.62242039 -0.11439697
  1.03746221 -1.12394882  0.06518087  0.75054939 -0.74410841  0.49055043
 -0.41531718  0.36573346  0.11305755 -0.04073057  0.46660199  0.30366126
 -0.35180624  0.40303014 -0.00458239 -0.50786457 -0.25522987 -0.13143575
  0.43307014  0.17720003 -0.04563277  0.16447643  0.79687722 -0.1653764
  0.44469374  0.43164848  0.74709735 -0.12298748 -0.90663211  0.42945048
 -0.18320209 -0.82732702  0.11346005 -0.49105868  0.32900779  0.9339572
  0.17727358 -0.13444528  0.59060319 -0.3544106   0.66449837 -0.00312158
 -0.03863223  0.70187556]
```

In [199]:
```
i=0
list_of_sentance_cv=[]
for sentance in X_cv:
    list_of_sentance_cv.append(sentance.split())
# 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])
```

```
100%|██████████| 17555/17555 [00:32<00:00, 536.86it/s]

(17555, 50)
[ 7.01921682e-01  3.04028464e-02 -7.94399215e-01 -4.80168554e-01
  8.21114877e-01  5.27055061e-01  1.28054740e+00 -1.22825691e-01
 -9.65965051e-03  3.06154662e-01 -5.81392206e-01  8.21606805e-01
 -1.19935475e-01  7.11479799e-01 -3.56254956e-01 -7.30671803e-01
  4.36932142e-01  8.14131679e-01  3.66898831e-01  1.39132197e-01
  2.63982420e-01  1.42895712e-01 -5.86853711e-01  1.89456004e-01
  4.17327814e-01 -3.44072148e-01 -6.42458412e-01  3.06149805e-01
  9.66563489e-01  1.67941286e-01  2.66331573e-01  8.66942067e-02
  2.94369854e-01  1.26106640e-01 -5.92979560e-01  5.84392663e-01
 -5.25343039e-01 -4.48585304e-01  5.10560757e-02 -1.03923317e+00
  1.49218999e-01  3.85795903e-01 -8.34666939e-01 -1.26234401e-02
  2.11017085e-01  9.67513128e-04  3.31895515e-01  6.18266383e-01
  2.30029214e-01  4.76786066e-01]
```

In [200]:
```
i=0
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in th
is 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, 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_test.append(sent_vec)
sent_vectors_test = np.array(sent_vectors_test)
print(sent_vectors_test.shape)
print(sent_vectors_test[0])
```

```
100%|██████████| 17555/17555 [00:33<00:00, 527.87it/s]

(17555, 50)
[ 0.00119119  1.05643972  0.79127266 -0.5209213   0.2162802  -0.29389119
 -0.38001179 -0.07901735 -0.25881543  0.34590223  0.09021694  0.63917509
  0.26305723 -0.38567108  0.02769856  0.22844634  0.03484211  0.64933102
 -0.22252587  0.67431445 -0.09554618  0.19268577  0.21643843 -0.0554288
  0.21913252 -0.22060314 -0.89718413 -0.20908981  0.35253083 -0.00449383
  0.96154297  1.03892573 -0.14291135 -0.08887237 -0.78004939  0.1135225
 -0.38199087 -0.07728498  0.39753774  0.40650368 -0.17096166  0.35606751
  0.23634751  0.23727409  0.05583161  0.54505504 -0.26937854  0.25495738
  0.11844105  0.27902913]
```

In [201]:
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score


print("After vectorizations")
print(sent_vectors_train.shape, y_train.shape)
print(sent_vectors_cv.shape, y_cv.shape)
print(sent_vectors_test.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(52663, 50) (52663,)
(17555, 50) (17555,)
(17555, 50) (17555,)
==============================================================================
=======================
```

Hyper parameter Tuning

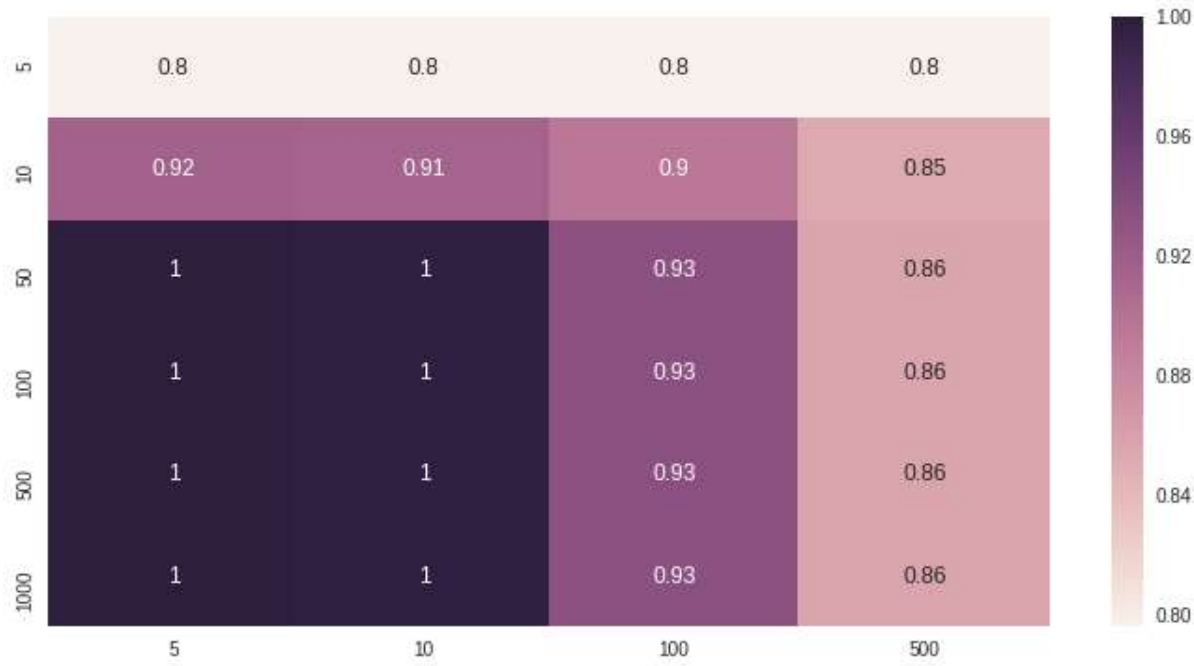
In [202]:
```
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt


#Hyper parameter tuning
train_auc = []
cv_auc = []
C=[ 5, 10, 50, 100, 500, 1000]#best_depth
D=[5, 10, 100, 500]#min_sample_splits
#For loop to find train_auc,cv_auc
for i in C:
  for j in D:
    clf = DecisionTreeClassifier(max_depth=i,min_samples_split=j,class_weight=
'balanced')
    clf.fit(sent_vectors_train,y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability e
stimates 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(cv_auc)
```

```
[0.7794338268562784, 0.7794338268562784, 0.7794338268562784, 0.77879283828780
78, 0.777204782205524, 0.7771903239090442, 0.8041207666893417, 0.811911223310
3284, 0.6545212639565519, 0.6710917972708235, 0.773899465892363, 0.8111801097
239875, 0.6554300869441927, 0.6700075723179933, 0.7720981683473966, 0.8111801
097239875, 0.6530599695771913, 0.6704612044384466, 0.7708966323608195, 0.8111
886275998422, 0.6575894912093558, 0.6728627790328485, 0.772007405102517, 0.81
11801097239875]
```

In [203]:
```
import numpy as np
train_auc=np.array(train_auc)
cv_auc = np.array(cv_auc)

#Train_auc Heatmap
print("Train_AUC Heatmap")

scores = train_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

```
Train_AUC Heatmap
```

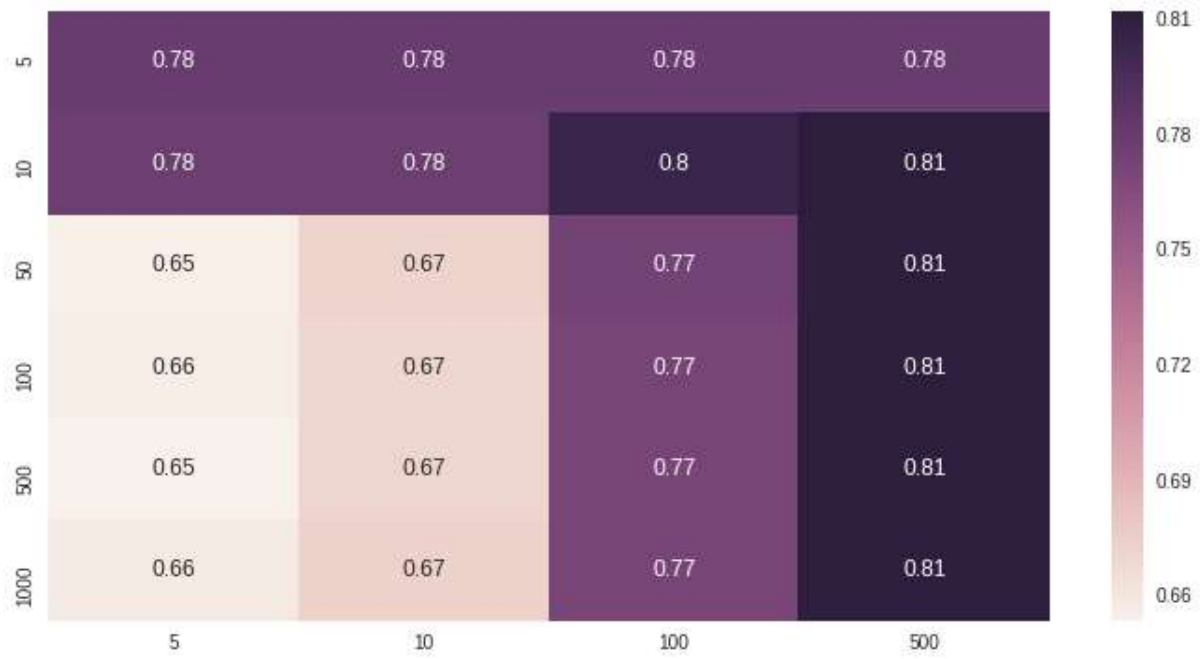
Out[203]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6434844780>



In [204]:
```
print("CV_AUC Heatmap")

#Hyper parameter tuning using seaborn heatmap
scores = cv_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

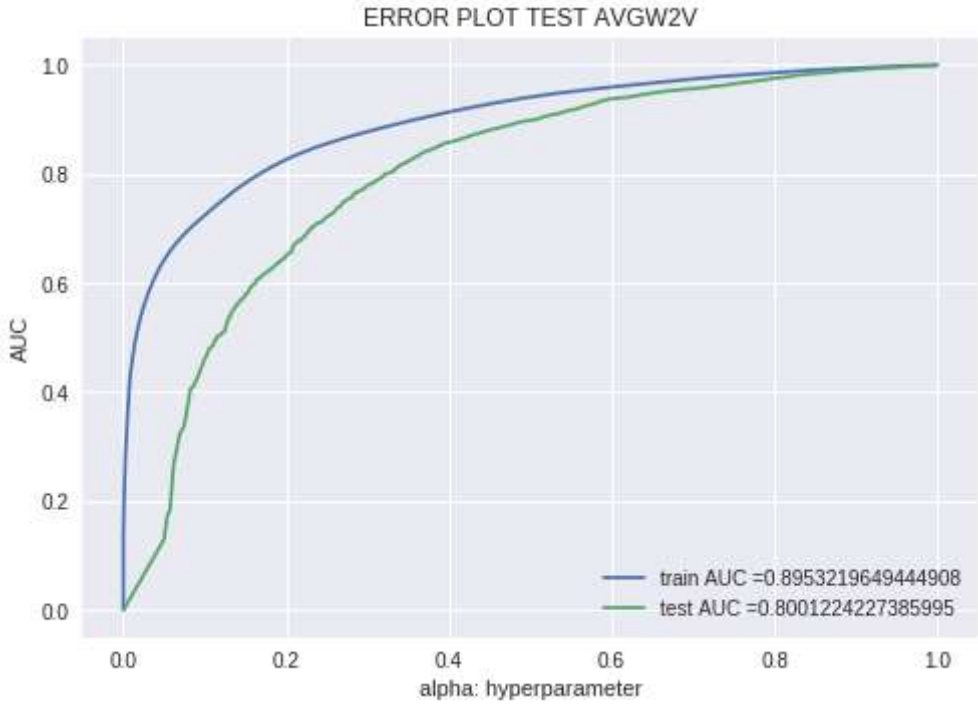
```
CV_AUC Heatmap
```

Out[204]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f64347377f0>
```



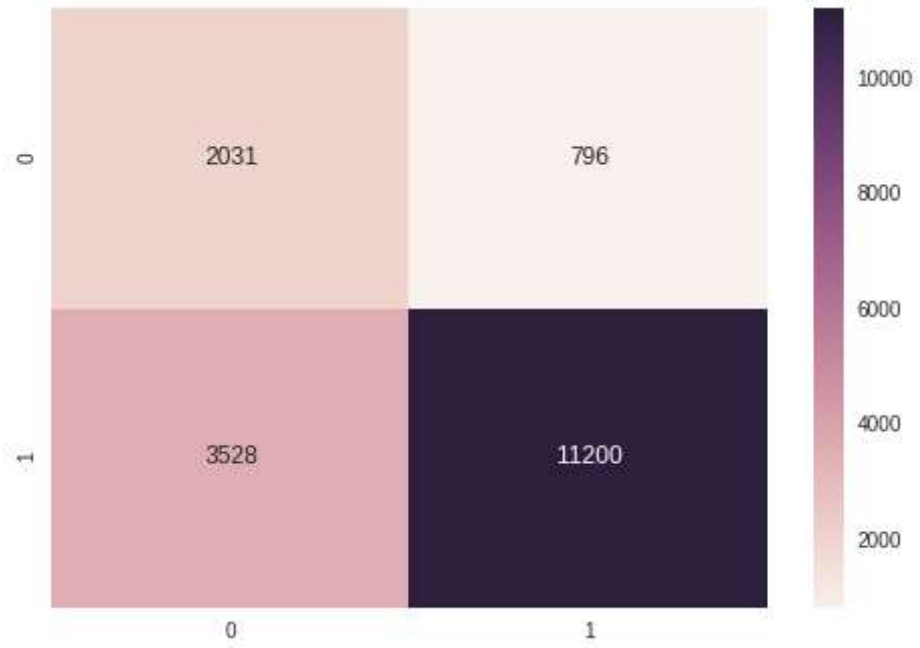
Testing with Test Data

In [206]:
```
#Best hyperparameters are max_depth=50,min_samples_split=100

from sklearn.metrics import roc_curve, auc


clf = DecisionTreeClassifier(max_depth=10,min_samples_split=100,class_weight=
'balanced')
clf.fit(sent_vectors_train,y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estim
ates of the positive class
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(sent_v
ectors_train)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(sent_vect
ors_test)[:,1])

#plotting the results
plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tp
r)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOT TEST AVGW2V")
plt.show()

print("="*100)

#printing confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns;sns.set()
print("Train confusion matrix")

#printing train_confusion matrix using pandas dataframe and seaborn

df_train=pd.DataFrame(confusion_matrix(y_train, clf.predict(sent_vectors_train
)))
sns.heatmap(df_train,annot=True,fmt="d")
plt.show()

#printing train_confusion matrix using pandas dataframe and seaborn
print("Test confusion matrix")

df_test=pd.DataFrame(confusion_matrix(y_test,clf.predict(sent_vectors_test)))
sns.heatmap(df_test,annot=True,fmt="d")
plt.show()
```




```
==============================================================================
========================
Train confusion matrix
```

```
Test confusion matrix
```



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

In [208]:
```
#preparing reviews for genism model
i=0
list_of_sentance_tf_train=[]
for sentance in X_train:
    list_of_sentance_tf_train.append(sentance.split())

#Training w2v models

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

# this line of code trains w2v model on the give list of sentances
w2v_model=Word2Vec(list_of_sentance_tf_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])
```

```
number of words that occured minimum 5 times  13833
sample words  ['love', 'tea', 'glad', 'amazon', 'carries', 'hard', 'find', 's
tores', 'one', 'place', 'know', 'way', 'expensive', 'buying', 'online', 'grea
t', 'herbal', 'strongly', 'cinnamon', 'naturally', 'sweet', 'not', 'need', 'a
dd', 'sweetner', 'also', 'bags', 'go', 'long', 'use', 'large', 'saucepan', 'f
ull', 'water', 'taste', 'somewhat', 'addicting', 'regular', 'boring', 'drinki
ng', 'reluctant', 'try', 'product', 'reviewing', 'reviews', 'people', 'though
t', 'heck', 'much', 'satisfy']
```

In [209]:
```
#Train Data

from tqdm import tqdm
import numpy as np

model = TfidfVectorizer()
model.fit(X_train)
tf_idf_matrix = model.transform(X_train)
print(type(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_)))

# 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_train= []; # the tfidf-w2v for each sentence/review is stored in th
is list
row=0;
for sent in tqdm(list_of_sentance_tf_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_train.append(sent_vec)
    row += 1
```

```
  0%|          | 5/52663 [00:00<19:38, 44.66it/s]

<class 'list'>

100%|██████████| 52663/52663 [19:47<00:00, 44.36it/s]
```

In [210]:
```
#CV Data
i=0
list_of_sentance_tf_cv=[]
for sentance in X_cv:
    list_of_sentance_tf_cv.append(sentance.split())

# 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_cv= []; # the tfidf-w2v for each sentence/review is stored in this
 list
row=0;
for sent in tqdm(list_of_sentance_tf_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_cv.append(sent_vec)
    row += 1
```

```
100%|██████████| 17555/17555 [06:32<00:00, 44.68it/s]
```

In [211]:
```
#Test Data
i=0
list_of_sentance_tf_test=[]
for sentance in X_test:
    list_of_sentance_tf_test.append(sentance.split())

# 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_test= []; # the tfidf-w2v for each sentence/review is stored in thi
s list
row=0;
for sent in tqdm(list_of_sentance_tf_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_test.append(sent_vec)
    row += 1
```

```
100%|██████████| 17555/17555 [06:30<00:00, 44.96it/s]
```

Hyper parameter Tuning

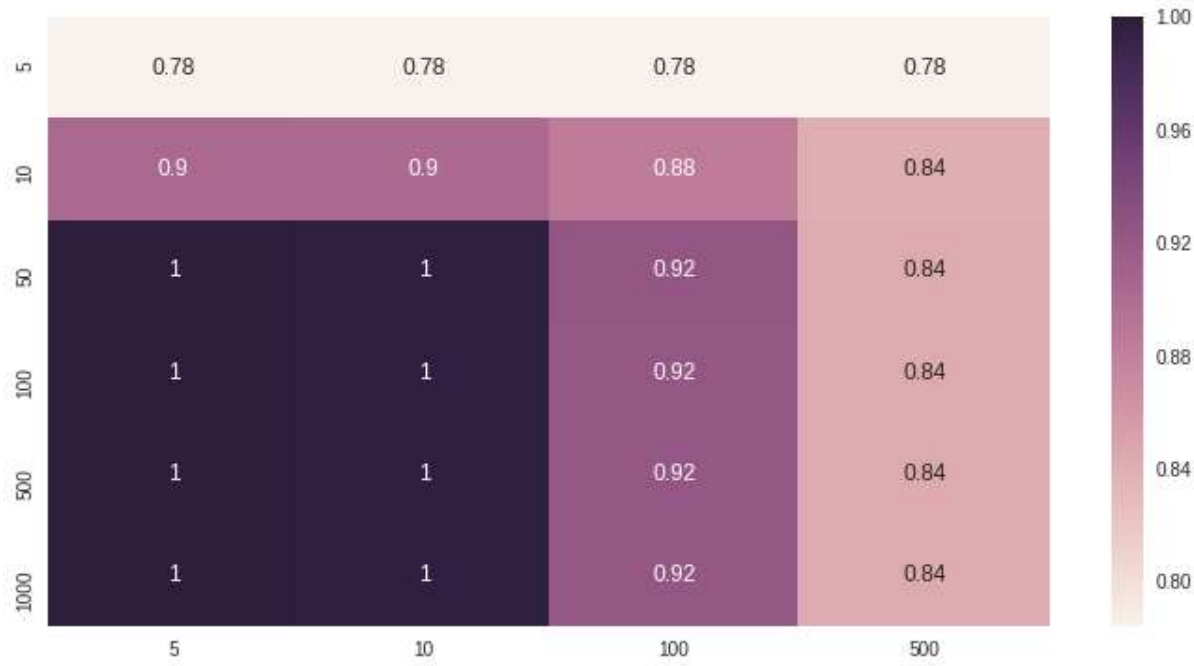
In [212]:
```
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt


#Hyper parameter tuning
train_auc = []
cv_auc = []
C=[ 5, 10, 50, 100, 500, 1000]#best_depth
D=[5, 10, 100, 500]#min_sample_splits
#For loop to find train_auc,cv_auc
for i in C:
  for j in D:
    clf = DecisionTreeClassifier(max_depth=i,min_samples_split=j,class_weight=
'balanced')
    clf.fit(tfidf_sent_train,y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability e
stimates of the positive class
    # not the predicted outputs
    y_train_pred =  clf.predict_proba(tfidf_sent_train)[:,1]
    y_cv_pred =  clf.predict_proba(tfidf_sent_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(cv_auc)
```

```
[0.7694034634517829, 0.7694034634517829, 0.7694034634517829, 0.76940346345178
29, 0.766954610964308, 0.7672378732940683, 0.7810709650502963, 0.791533567696
837, 0.6515203205588433, 0.6602073142975418, 0.7537239564104699, 0.7902479084
809961, 0.6492484533304845, 0.6590262261224548, 0.7548863641719388, 0.7902479
084809961, 0.6535891702302643, 0.6594877992597352, 0.7540882367496218, 0.7902
072092355572, 0.6472237026906865, 0.662528644119121, 0.7555996687700011, 0.79
02072092355572]
```

In [213]:
```
import numpy as np
train_auc=np.array(train_auc)
cv_auc = np.array(cv_auc)

#Train_auc Heatmap
print("Train_AUC Heatmap")

scores = train_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

```
Train_AUC Heatmap
```

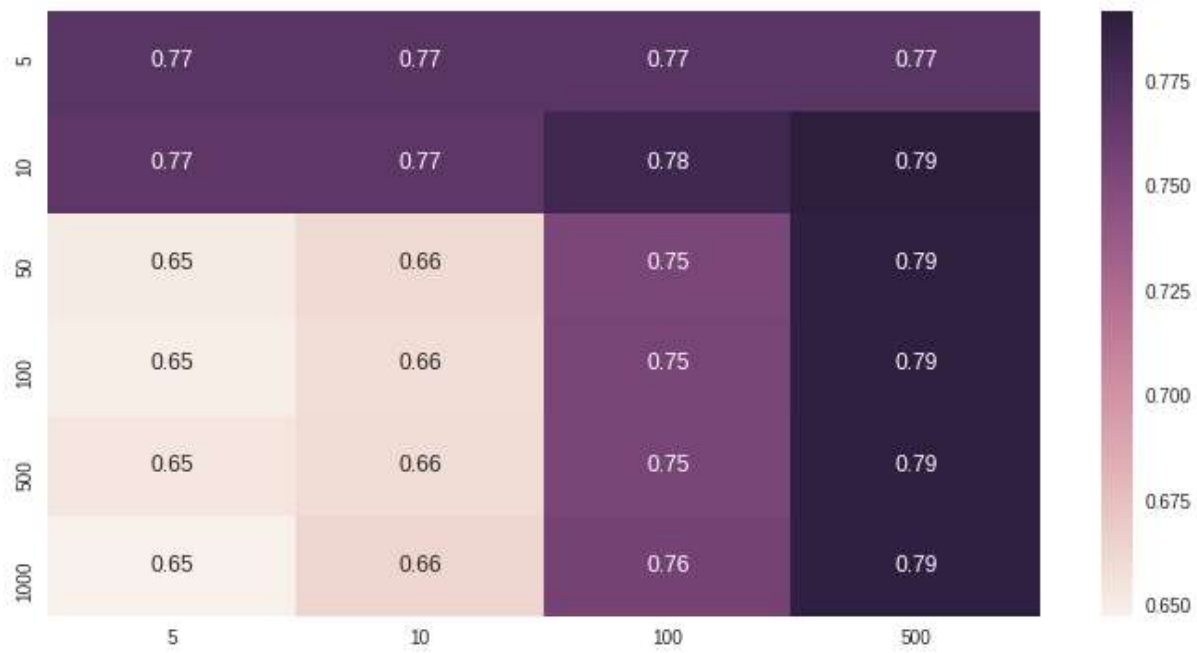
Out[213]: <matplotlib.axes._subplots.AxesSubplot at 0x7f64266068d0>



In [214]:
```
print("CV_AUC Heatmap")

#Hyper parameter tuning using seaborn heatmap
scores = cv_auc.reshape(len(C),len(D))
plt.figure(figsize = (12,6))
df = pd.DataFrame(scores, index=C, columns= D)
sns.heatmap(df, annot=True)
```

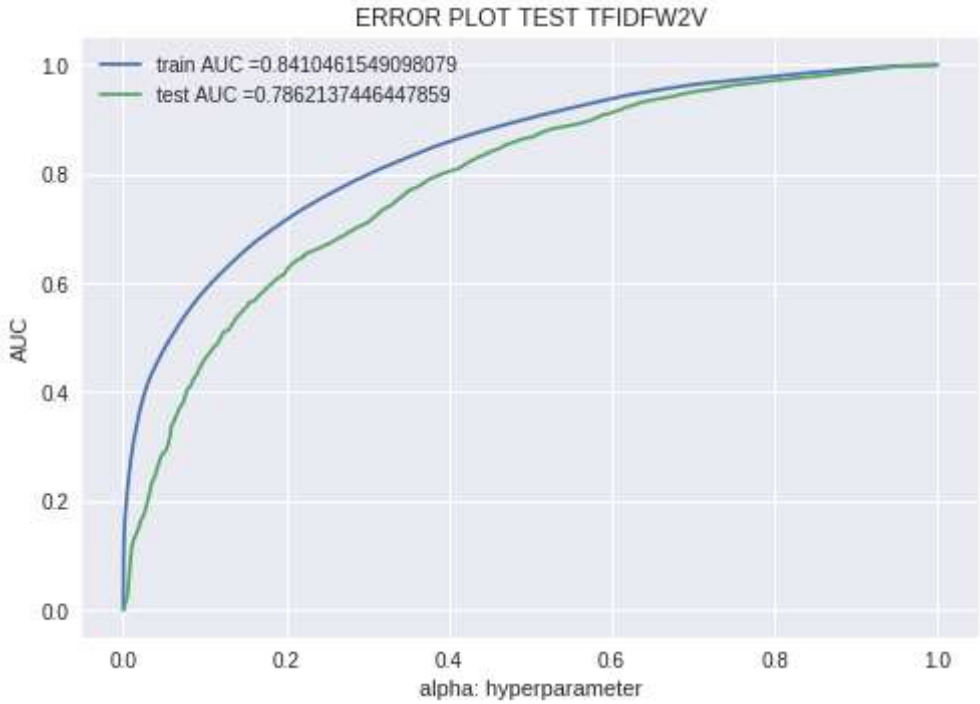
```
CV_AUC Heatmap
```

Out[214]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6426606e80>



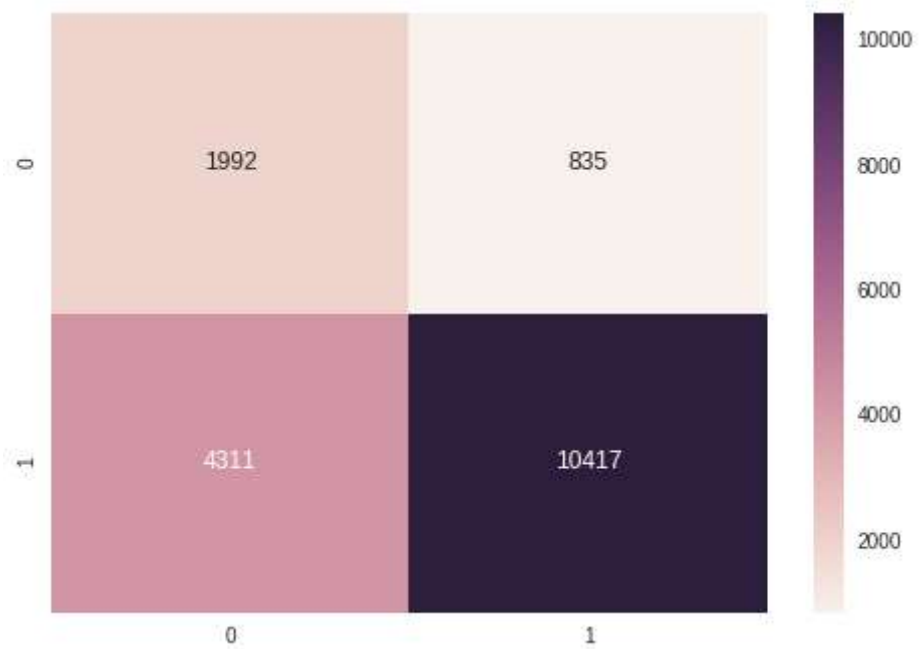
Testing with Test Data

In [215]:
```
#Best hyperparameters are max_depth=50,min_samples_split=100

from sklearn.metrics import roc_curve, auc


clf = DecisionTreeClassifier(max_depth=10,min_samples_split=500,class_weight=
'balanced')
clf.fit(tfidf_sent_train,y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estim
ates of the positive class
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(tfidf_
sent_train)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(tfidf_sen
t_test)[:,1])

#plotting the results
plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tp
r)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOT TEST TFIDFW2V")
plt.show()

print("="*100)

#printing confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns;sns.set()
print("Train confusion matrix")

#printing train_confusion matrix using pandas dataframe and seaborn

df_train=pd.DataFrame(confusion_matrix(y_train, clf.predict(tfidf_sent_train
)))
sns.heatmap(df_train,annot=True,fmt="d")
plt.show()

#printing train_confusion matrix using pandas dataframe and seaborn
print("Test confusion matrix")

df_test=pd.DataFrame(confusion_matrix(y_test,clf.predict(tfidf_sent_test)))
sns.heatmap(df_test,annot=True,fmt="d")
plt.show()
```




```
==========================================================================
========================
Train confusion matrix
```

```
Test confusion matrix
```



[6] Conclusions

In [216]:
```
from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["Vectorizer", "Best_depth", "Best_Min_samples_split", "AUC"]

x.add_row(["BOW",50,100,0.8175 ])
x.add_row(["TFIDF",100,100,0.8187 ])
x.add_row(["AVGW2V",50,100,0.8001 ])
x.add_row(["TFIDFW2V",50,100,0.7862 ])



print(x)
```

```
+------------+------------+------------------------+--------+
| Vectorizer | Best_depth | Best_Min_samples_split |  AUC   |
+------------+------------+------------------------+--------+
|    BOW     |     50     |          100           | 0.8175 |
|   TFIDF    |    100     |          100           | 0.8187 |
|   AVGW2V   |     50     |          100           | 0.8001 |
|  TFIDFW2V  |     50     |          100           | 0.7862 |
+------------+------------+------------------------+--------+
```

1.TFIDF gives us best AUC with best_depth of 50 and min_samples_split=100

2.Average min_samples_split is 100.