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

[1]. Reading Data

Applying Decision Tree

[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".

Mounting Google Drive locally

In [1]:
```
from google.colab import drive
drive.mount('/content/gdrive')
```

Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=email%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fpeopleapi.readonly&response_type=code (https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=email%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fpeopleapi.readonly&response_type=code)

Enter your authorization code:
··········
Mounted at /content/gdrive

In [2]:

```
pip install paramiko
```

```
Collecting paramiko
  Downloading https://files.pythonhosted.org/packages/cf/ae/94e70d49044ccc234bfdba20114fa947d7ba6eb68a2e452d89b920e62227/paramiko-2.4.2-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/cf/ae/94e70d49044ccc234bfdba20114fa947d7ba6eb68a2e452d89b920e62227/paramiko-2.4.2-py2.py3-none-any.whl) (193kB)
    100% |████████████████████████████████| 194kB 8.7MB/s
Collecting pynacl>=1.0.1 (from paramiko)
  Downloading https://files.pythonhosted.org/packages/27/15/2cd0a203f318c2240b42cd9dd13c931ddd61067809fee3479f44f086103e/PyNaCl-1.3.0-cp34-abi3-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/27/15/2cd0a203f318c2240b42cd9dd13c931ddd61067809fee3479f44f086103e/PyNaCl-1.3.0-cp34-abi3-manylinux1_x86_64.whl) (759kB)
    100% |████████████████████████████████| 768kB 30.9MB/s
Collecting bcrypt>=3.1.3 (from paramiko)
  Downloading https://files.pythonhosted.org/packages/d0/79/79a4d167a31cc206117d9b396926615fa9c1fdbd52017bcced80937ac501/bcrypt-3.1.6-cp34-abi3-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/d0/79/79a4d167a31cc206117d9b396926615fa9c1fdbd52017bcced80937ac501/bcrypt-3.1.6-cp34-abi3-manylinux1_x86_64.whl) (55kB)
    100% |████████████████████████████████| 61kB 34.6MB/s
Requirement already satisfied: pyasn1>=0.1.7 in /usr/local/lib/python3.6/dist-packages (from paramiko) (0.4.5)
Collecting cryptography>=1.5 (from paramiko)
  Downloading https://files.pythonhosted.org/packages/5b/12/b0409a94dad366d98a8eee2a77678c7a73aafd8c0e4b835abea634ea3896/cryptography-2.6.1-cp34-abi3-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/5b/12/b0409a94dad366d98a8eee2a77678c7a73aafd8c0e4b835abea634ea3896/cryptography-2.6.1-cp34-abi3-manylinux1_x86_64.whl) (2.3MB)
    100% |████████████████████████████████| 2.3MB 15.9MB/s
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from pynacl>=1.0.1->paramiko) (1.12.0)
Requirement already satisfied: cffi>=1.4.1 in /usr/local/lib/python3.6/dist-packages (from pynacl>=1.0.1->paramiko) (1.12.3)
Collecting asn1crypto>=0.21.0 (from cryptography>=1.5->paramiko)
  Downloading https://files.pythonhosted.org/packages/ea/cd/35485615f45f30a510576f1a56d1e0a7ad7bd8ab5ed7cdc600ef7cd06222/asn1crypto-0.24.0-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/ea/cd/35485615f45f30a510576f1a56d1e0a7ad7bd8ab5ed7cdc600ef7cd06222/asn1crypto-0.24.0-py2.py3-none-any.whl) (101kB)
    100% |████████████████████████████████| 102kB 39.1MB/s
Requirement already satisfied: pycparser in /usr/local/lib/python3.6/dist-packages (from cffi>=1.4.1->pynacl>=1.0.1->paramiko) (2.19)
Installing collected packages: pynacl, bcrypt, asn1crypto, cryptography, paramiko
Successfully installed asn1crypto-0.24.0 bcrypt-3.1.6 cryptography-2.6.1 paramiko-2.4.2 pynacl-1.3.0
```

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 [4]:
```
# using SQLite Table to read data.
con = sqlite3.connect("/content/gdrive/My Drive/Dataset/database.sqlite")



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[4]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	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 [6]:
```
print(display.shape)
display.head()
```

(80668, 7)

Out[6]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
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 [7]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[7]:

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

In [0]:
```
display['COUNT(*)'].sum()
```

Out[7]: 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 [9]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND UserId="AR5J8UI46CURR"
ORDER BY ProductID
""", con)
display.head()
```

Out[9]:

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

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

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

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

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

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

In [0]:
```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_pc
```

In [11]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inplace=False)
final.shape
```

Out[11]: (87775, 10)

In [12]:
```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[12]: 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 [13]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)

display.head()
```

Out[13]:

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

In [14]:
```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
final.shape
```

Out[14]: (87773, 10)

In [15]:
```
#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[15]:
```
1    73592
0    14181
Name: Score, dtype: int64
```

[3] Preprocessing

[3.1]. Preprocessing Review Text

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

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

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

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

In [0]:
```
# https://stackoverflow.com/a/47091490/4084039
import re

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

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

In [0]:
```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', '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', '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', 'each', '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', "doesn'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 [18]:
```
# Combining all the above stundents
from bs4 import BeautifulSoup
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 [00:30<00:00, 2896.67it/s]
```

In [0]:
```
final["CleanText"] = [preprocessed_reviews[i] for i in range(len(final))]
```

In [20]:
```
final.head(2)
```

Out[20]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenom
22620	24750	2734888454	A13ISQV0U9GZIC	Sandikaye	1	
22621	24751	2734888454	A1C298ITT645B6	Hugh G. Pritchard	0	

[4] Featurization

In [0]:
```
from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import roc_auc_score
import seaborn as sns

from sklearn.metrics import confusion_matrix

# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
```

In [0]:
```
Total_X = final['CleanText'].values
Total_y = final['Score'].values
```

In [0]:
```
# split the data set into train and test
X_train, X_test, y_train, y_test = train_test_split(Total_X, Total_y, test_size=0.33)

# split the train data set into cross validation train and cross validation test
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.33)
```

In [24]:
```
print(f"Train Data : ({len(X_train)} , {len(y_train)})")
print(f"CV Data : ({len(X_cv)} , {len(y_cv)})")
print(f"Test Data : ({len(X_test)} , {len( y_test)})")
```

```
Train Data : (39400 , 39400)
CV Data : (19407 , 19407)
Test Data : (28966 , 28966)
```

Decision Tree Training and Tesing Function

In [0]:
```
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
```

In [0]:
```
def DT(X_train_reg,X_cv_reg, y_train=y_train,  y_cv=y_cv,  y_test=y_test):
    train_auc = []
    cv_auc = []
    max_d=0
    min_s = 0
    max_roc_auc=-1

    tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples_split': [5, 10, 100, 500]}]

    #Using GridSearchCV
    model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, n_jobs=2 ,scoring = 'roc_auc', cv=5)
    model.fit(X_train_reg, y_train)

    print(model.best_estimator_)
    print("_"*10)
    print("Best HyperParameter: ",model.best_params_)
    print(f"Best Accuracy: {model.best_score_*100}")



    y_train_pred =  model.predict_proba(X_train_reg)[:,1]
    y_cv_pred =  model.predict_proba(X_cv_reg)[:,1]

    proba2 = roc_auc_score(y_cv, y_cv_pred) * float(100)


    max_depth = [1, 5, 10, 50, 100, 500, 1000]
    for i in max_depth:
        if(max_roc_auc<proba2):
            max_roc_auc=proba2
            max_d=i

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

    plt.figure(1)
    plt.plot(max_depth, train_auc, label='Train AUC')
    plt.plot(max_depth, cv_auc, label='CV AUC')
    plt.legend()
    plt.xlabel("max_depth: hyperparameter")
    plt.ylabel("Area Under ROC Curve")
    plt.title("ERROR PLOTS")
    plt.show()
```

Testing the max_depth with Test datapoints and Confusion Matrix

In [0]:
```
def testing_DT(X_train_reg,X_test_reg, y_train=y_train,  y_test=y_test):

    clf= GridSearchCV(DecisionTreeClassifier(), tuned_parameters, n_jobs=2 ,scoring = 'roc_auc', cv=5)
    clf.fit(X_test_reg, y_test)

    train_fpr, train_tpr, thresholds = roc_curve(y_train, clf.predict_proba(X_train_reg)[:,1])
    test_fpr, test_tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test_reg)[:,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.xscale(value = 'log')
    plt.legend()
    plt.xlabel("False Positive Rate")
    plt.ylabel("True Positive Rate")
    plt.title("ROC curves")
    plt.show()

    print(f"Precision on test data: {precision_score(y_test, clf.predict(X_test_reg))}")
    print(f"Recall on test data: {recall_score(y_test, clf.predict(X_test_reg))}")
    print(f"F1-Score on test data: {f1_score(y_test, clf.predict(X_test_reg))}")

    print("\nConfusion Matrix of Train and Test set:\n [ [TN  FP]\n   [FN TP] ]\n")
    confusionMatrix_train=confusion_matrix(y_train, clf.predict(X_train_reg))
    confusionMatrix_test=confusion_matrix(y_test, clf.predict(X_test_reg))
    df_cm_tr = pd.DataFrame(confusionMatrix_train, range(2),range(2))
    df_cm_te = pd.DataFrame(confusionMatrix_test, range(2),range(2))
    plt.figure(figsize = (7,5))
    plt.ylabel("Predicted label")
    plt.xlabel("Actual label")
    plt.title("Confusion Matrix of Train Set")
    sns.set(font_scale=1.4)#for label size
    sns.heatmap(df_cm_tr, annot=True,annot_kws={"size": 12},fmt="d")

    plt.figure(figsize = (7,6))
    plt.ylabel("Predicted label")
    plt.xlabel("Actual label")
    plt.title("Confusion Matrix of Test Set")
    sns.heatmap(df_cm_te, annot=True,annot_kws={"size": 12},fmt="d")
```

[4.1] BAG OF WORDS

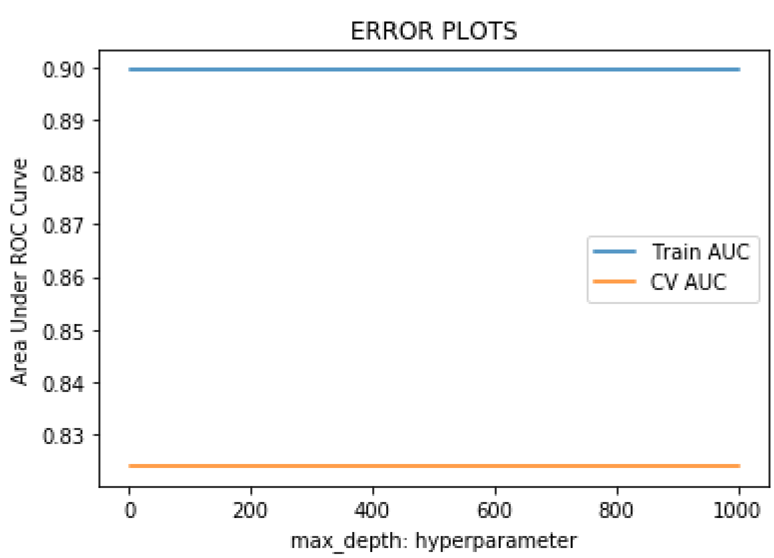
[5.1] Decision Tree on BOW, SET 1

In [27]:
```
#BoW
count_vect = CountVectorizer() #in scikit-learn
count_vect.fit(X_train)
print("some feature names ", count_vect.get_feature_names()[1000:1010])
print('='*50)

# we use the fitted CountVectorizer to convert the text to vector
X_train_bow = count_vect.transform(X_train)
X_cv_bow = count_vect.transform(X_cv)
X_test_bow = count_vect.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)
```

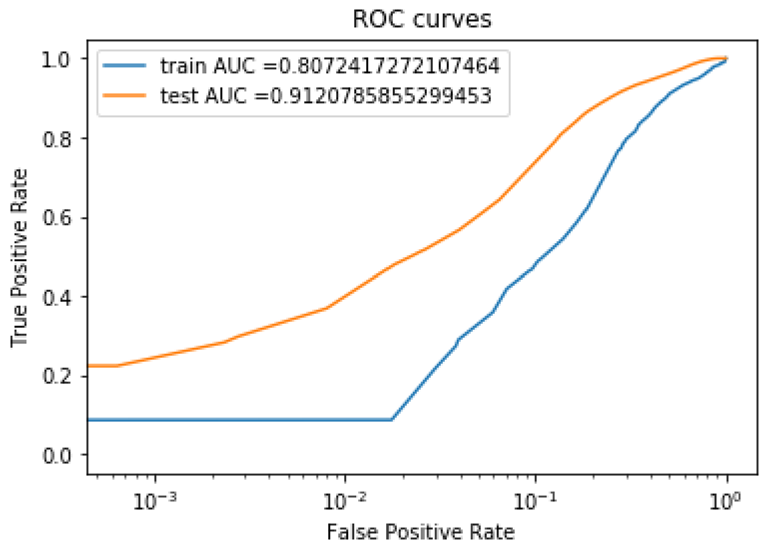
```
some feature names  ['aluminum', 'alvacado', 'alvita', 'alwadi', 'alwasy', 'alway', 'always', 'alwaysgood', 'alwsay
s', 'alyssia']
==================================================
After vectorizations
(39400, 37461) (39400,)
(19407, 37461) (19407,)
(28966, 37461) (28966,)
================================================================================================
============
```

In [82]:
```
DT(X_train_bow,X_cv_bow)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')

____________
Best HyperParameter:  {'max_depth': 50, 'min_samples_split': 500}
Best Accuracy: 80.78750941730853
```

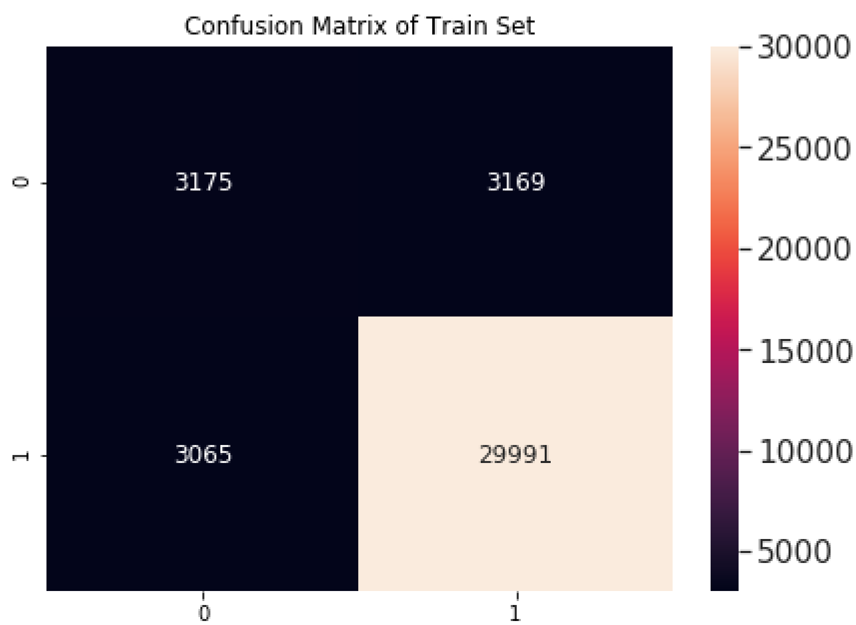


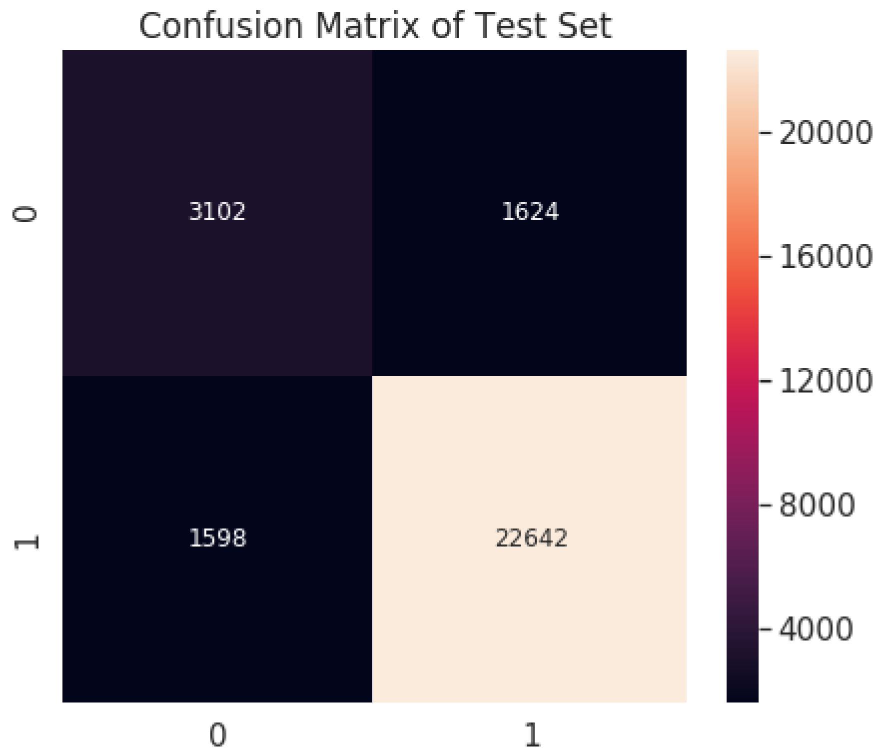
In [86]:
```
testing_DT(X_train_bow,X_test_bow)
```



Precision on test data: 0.9330750844803428
Recall on test data: 0.9340759075907591
F1-Score on test data: 0.9335752278068693

Confusion Matrix of Train and Test set:
[[TN FP]
[FN TP]]



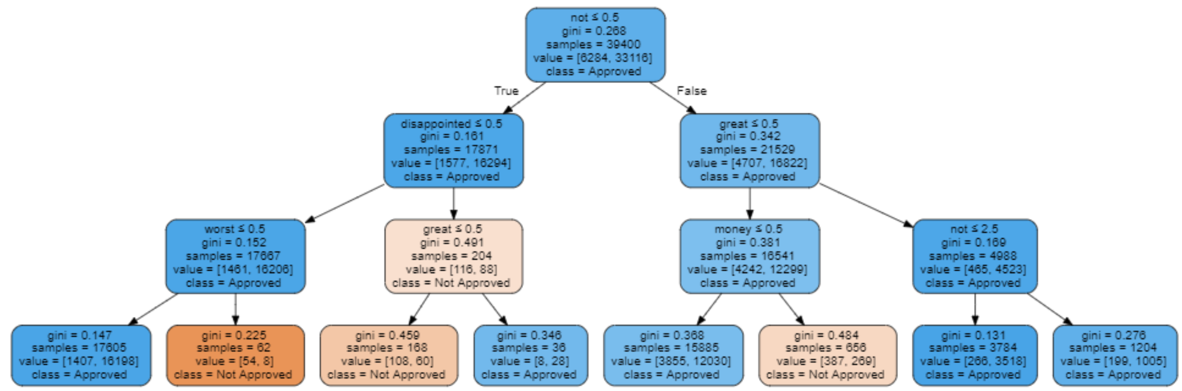


Tree Visualization

In [0]:
```
from sklearn import tree
import graphviz
import pydotplus
from sklearn.tree import DecisionTreeClassifier, export_graphviz
```

In [104]:
```
# Create and fit the decision tree
clf_dt = DecisionTreeClassifier(criterion = 'gini', max_depth = 3)
clf_dt.fit(X_train_bow, y_train)


dot_data = export_graphviz(clf_dt, out_file=None,  feature_names=count_vect.get_feature_names(), class_nam
                filled=True, rounded=True, \
                special_characters=True)
graph = graphviz.Source(dot_data)

gvz_graph = graphviz.Source(pydot_graph.to_string())
gvz_graph
```

Out[104]: <graphviz.files.Source at 0x7f6ccc828f28>



[5.1.1] Top 10 important features of positive and negative class from SET 1

Code reference: https://www.datacamp.com/community/tutorials/wordcloud-python (https://www.datacamp.com/community/tutorials/wordcloud-python)

https://python-graph-gallery.com/wordcloud/ (https://python-graph-gallery.com/wordcloud/)

https://stackoverflow.com/questions/11116697/how-to-get-most-informative-features-for-scikit-learn-classifiers (https://stackoverflow.com/questions/11116697/how-to-get-most-informative-features-for-scikit-learn-classifiers)

In [0]:

```
from wordcloud import WordCloud, STOPWORDS
```

In [0]:

```
def important_features(vect,max_depth,min_samples_split,X_train_reg, n):
    clf = DecisionTreeClassifier(max_depth = max_depth, min_samples_split=min_samples_split)
    clf.fit(X_train_reg, y_train)

    features =vect.get_feature_names()
    coef = clf.feature_importances_
    coef_df = pd.DataFrame({'word': features, 'coeficient': coef}, index = None)
    df = coef_df.sort_values("coeficient", ascending = False)[:n]
    cloud = " ".join(word for word in df.word)
    stopwords = set(STOPWORDS)
    wordcloud = WordCloud(width = 1000, height = 600, background_color ='white', stopwords = stopwords).gene

    # plot the WordCloud image
    plt.figure(figsize = (10, 8))
    plt.imshow(wordcloud, interpolation = 'bilinear')
    plt.axis("off")
    plt.title(f"Top {n} most important features")
    plt.tight_layout(pad = 0)

    plt.show()
```

In [79]:
```
important_features(count_vect,50,500,X_train_bow,20)
```

[4.2] Bi-Grams and n-Grams.

In [45]:
```
#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.fe

# 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

[5.2] Decision Tree on TFIDF, SET 2

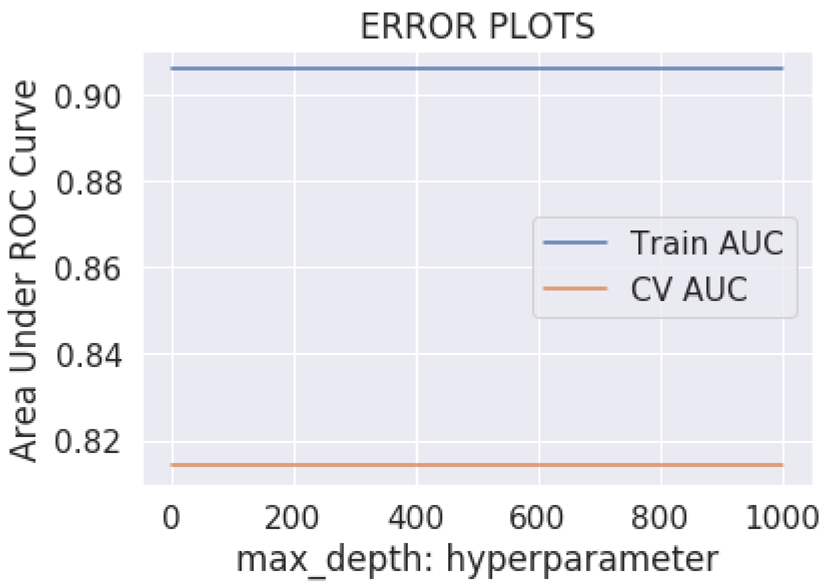
In [87]:

```
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)


# we use the fitted CountVectorizer to convert the text to vector
X_train_tf_idf = tf_idf_vect.transform(X_train)
X_cv_tf_idf = tf_idf_vect.transform(X_cv)
X_test_tf_idf = tf_idf_vect.transform(X_test)


print("After vectorizations")
print(X_train_tf_idf.shape, y_train.shape)
print(X_cv_tf_idf.shape, y_cv.shape)
print(X_test_tf_idf.shape, y_test.shape)
print("="*100)
```

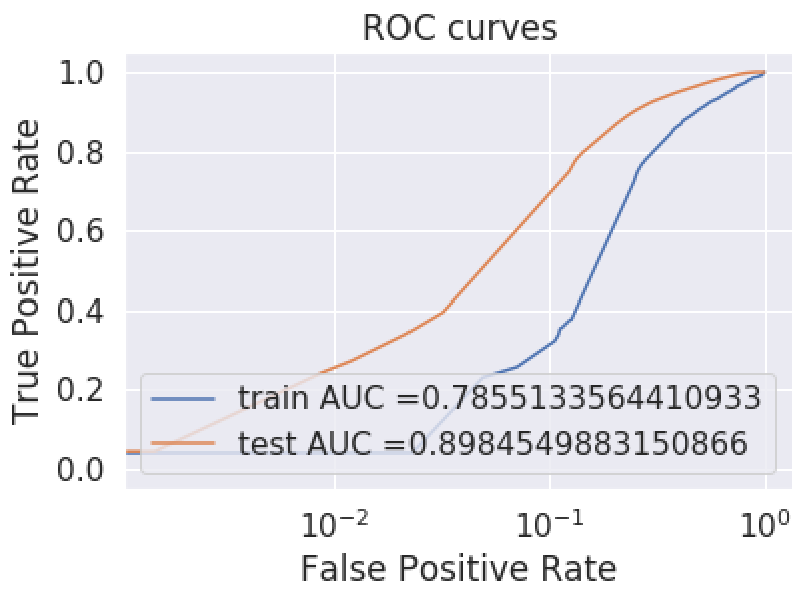
```
some sample features(unique words in the corpus) ['ability', 'able', 'able buy', 'able chew', 'able drink', 'able e
at', 'able enjoy', 'able find', 'able finish', 'able get']
==================================================
After vectorizations
(39400, 23412) (39400,)
(19407, 23412) (19407,)
(28966, 23412) (28966,)
==========================================================================================
============
```

In [88]:
```
DT(X_train_tf_idf,X_cv_tf_idf)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
          max_features=None, max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=500,
          min_weight_fraction_leaf=0.0, presort=False, random_state=None,
          splitter='best')
____________
Best HyperParameter:  {'max_depth': 50, 'min_samples_split': 500}
Best Accuracy: 78.54995967401271
```

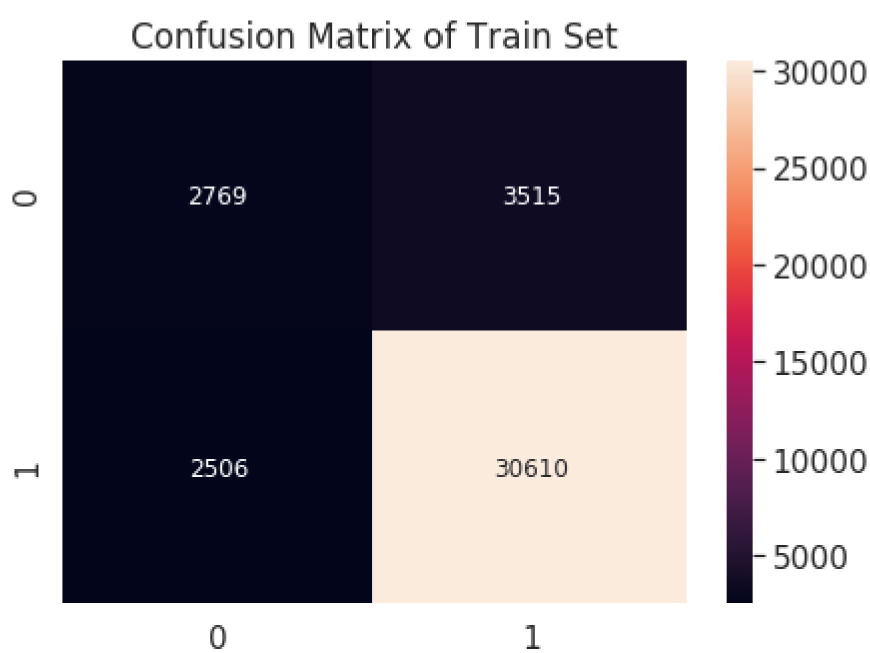


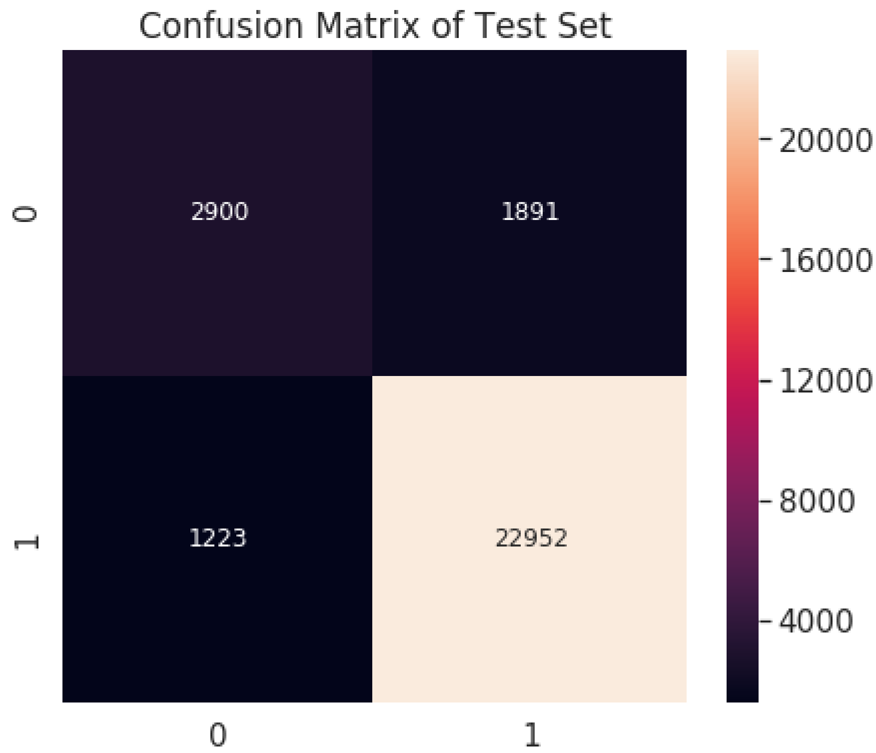
In [54]:
```
testing_DT(X_train_tf_idf,X_test_tf_idf)
```



Precision on test data: 0.9238819788270338
Recall on test data: 0.9494105480868666
F1-Score on test data: 0.9364723162919745

Confusion Matrix of Train and Test set:
[[TN FP]
[FN TP]]





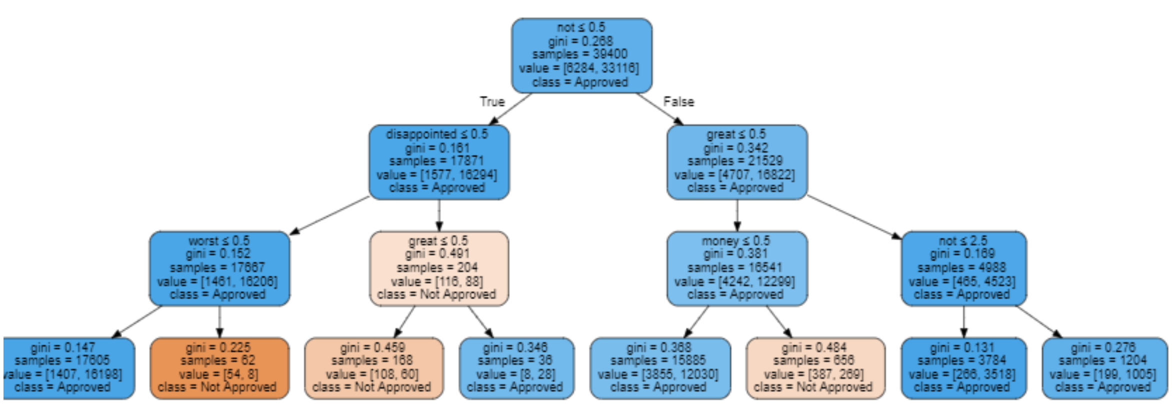
Graphviz Reference: https://stackoverflow.com/questions/51346827/how-can-i-specify-the-figsize-of-a-graphviz-representation-of-decision-tree (https://stackoverflow.com/questions/51346827/how-can-i-specify-the-figsize-of-a-graphviz-representation-of-decision-tree)

In [109]:

```
# Create and fit the decision tree
clf_dt = DecisionTreeClassifier(criterion = 'gini', max_depth = 3)
clf_dt.fit(X_train_tf_idf, y_train)


dot_data = export_graphviz(clf_dt, out_file=None,  feature_names=tf_idf_vect.get_feature_names(), class_name
                filled=True, rounded=True, \
                special_characters=True)
graph = graphviz.Source(dot_data)

gvz_graph = graphviz.Source(pydot_graph.to_string())
gvz_graph
```

Out[109]:

```
<graphviz.files.Source at 0x7f6cc86c3898>
```



[5.2.2] Top 10 important features of Positive and Negative class from SET 2

In [91]:
```
important_features(tf_idf_vect,50,500,X_train_tf_idf,20)
```

[4.4] Word2Vec

In [0]:
```
i=0

w2v_train=[]
w2v_cv=[]
w2v_test=[]

for sentance in X_train:
    w2v_train.append(sentance.split())

for sentance in X_cv:
    w2v_cv.append(sentance.split())


for sentance in X_test:
    w2v_test.append(sentance.split())
```

In [93]:
```
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)
    w2v_model_train = Word2Vec(w2v_train,min_count=5,size=50, workers=4)
    print(w2v_model_train.wv.most_similar('great'))
    print('='*50)
else:
    pass
```

```
[('awesome', 0.8482818603515625), ('good', 0.7975322604179382), ('fantastic', 0.779043436050415), ('won
derful', 0.7715604901313782), ('perfect', 0.7698479294776917), ('excellent', 0.7695501446723938), ('terrifi
c', 0.7188437581062317), ('fabulous', 0.701779305934906), ('amazing', 0.7003705501556396), ('delicious',
0.678266167640686)]
==================================================
```

In [94]:
```
w2v_words_train = list(w2v_model_train.wv.vocab)

print("number of words that occured minimum 5 times ",len(w2v_words_train ))
print("sample words ", w2v_words_train[0:50])
```

```
number of words that occured minimum 5 times  12017
sample words  ['dont', 'know', 'say', 'besides', 'fact', 'best', 'sauce', 'tried', 'ghirardelli', 'torani', 'ect', 'one', 'sm
ooth', 'tasty', 'mmmmmm', 'make', 'caramel', 'would', 'delicious', 'ice', 'cream', 'dessert', 'matter', 'ordered', 'e
ner', 'g', 'brand', 'bread', 'disney', 'kids', 'adult', 'celiac', 'excited', 'everything', 'brownies', 'not', 'measure', 'ma
kes', 'never', 'met', 'brownie', 'like', 'could', 'eat', 'even', 'clear', 'sign', 'throw', 'away', 'desperate']
```

#Converting text into vectors using Avg W2V, TFIDF-W2V

[5.1.3] Applying Decision Tree on AVG W2V, SET 3

[4.4.1.1] Avg W2v

In [95]:

```
train_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(w2v_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 us
    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_train:
            vec = w2v_model_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    train_vectors.append(sent_vec)
print()
print(len(train_vectors))
print(len(train_vectors[0]))
```

```
100%|██████████| 39400/39400 [00:55<00:00, 716.35it/s]


39400
50
```

In [96]:

```
# compute average word2vec for each review.
cv_vectors = [] # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(w2v_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 us
    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_train:
            vec = w2v_model_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    cv_vectors.append(sent_vec)
print()
print(len(cv_vectors))
print(len(cv_vectors[0]))
```

```
100%|██████████| 19407/19407 [00:26<00:00, 731.23it/s]


19407
50
```

In [97]:

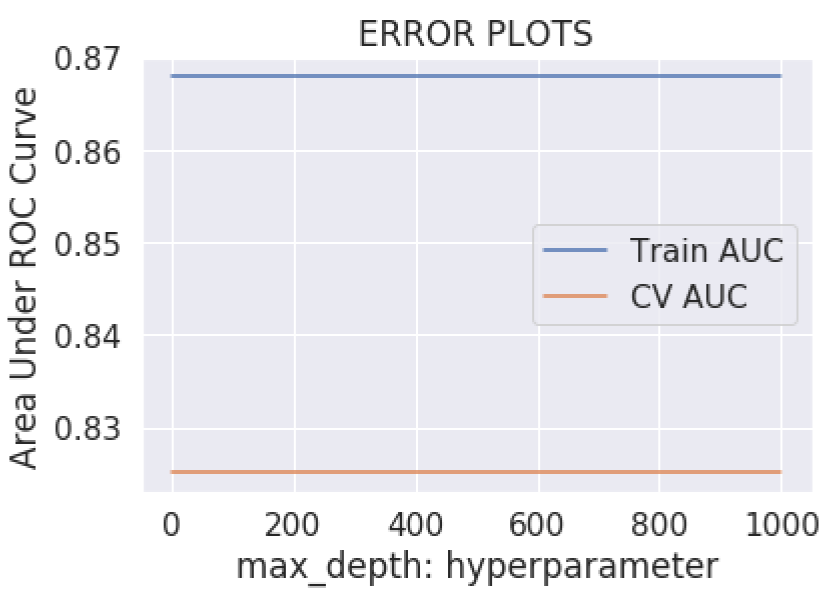
```
test_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(w2v_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change this to 300 if you us
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train:
            vec =  w2v_model_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    test_vectors.append(sent_vec)
print()
print(len(test_vectors))
print(len(test_vectors[0]))
```

```
100%|██████████| 28966/28966 [00:41<00:00, 701.65it/s]


28966
50
```

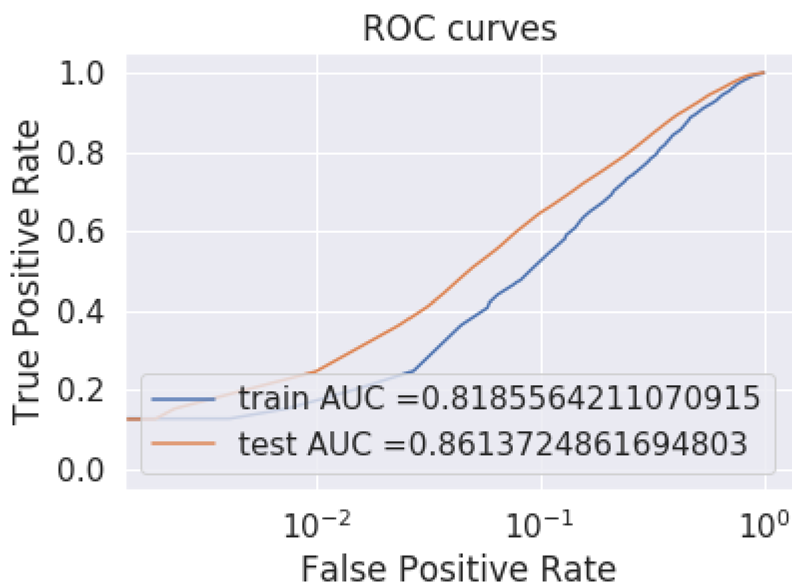
In [98]:

```
DT(train_vectors,cv_vectors)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=10,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
____________
Best HyperParameter:  {'max_depth': 10, 'min_samples_split': 500}
Best Accuracy: 82.10689904165885
```

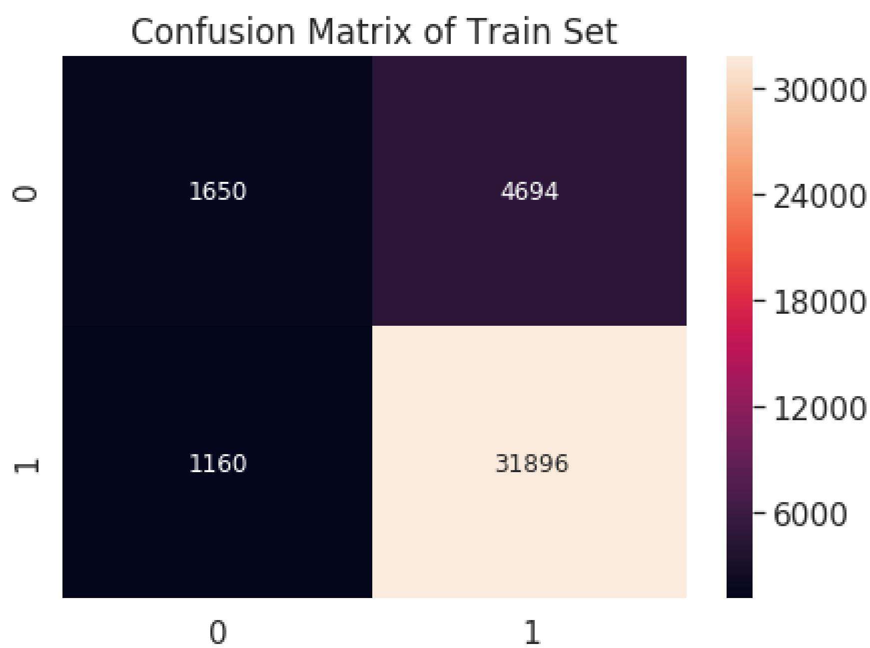


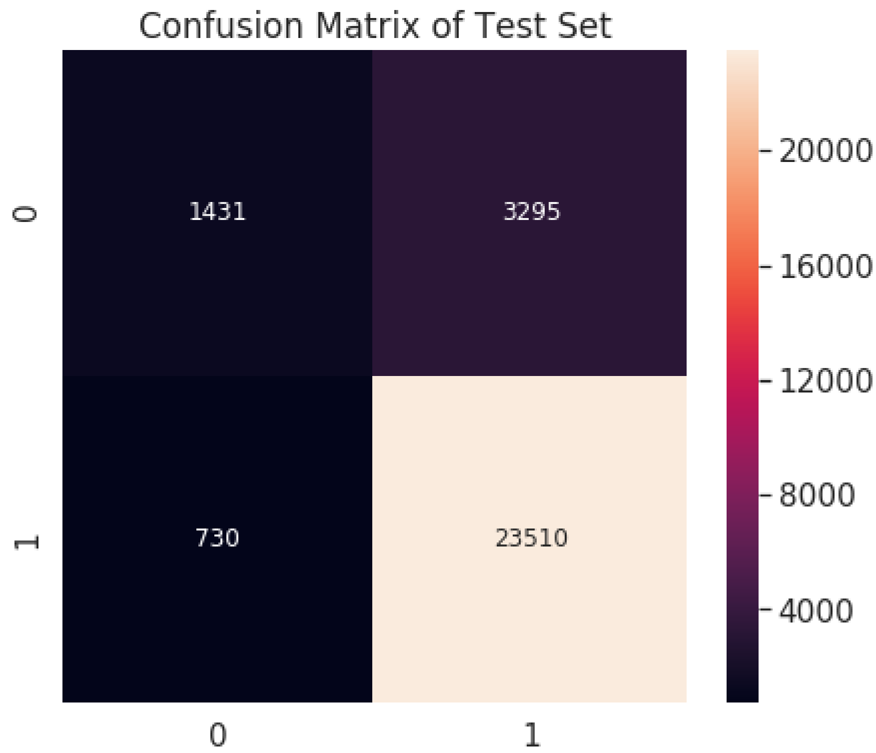
In [100]:
```
testing_DT(train_vectors, test_vectors)
```



Precision on test data: 0.8770751725424362
Recall on test data: 0.9698844884488449
F1-Score on test data: 0.9211480066607896

Confusion Matrix of Train and Test set:
[[TN FP]
[FN TP]]





[5.4] Decision Tree on TFIDF W2V

In [0]:

```
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 [102]:

```
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

train_tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(w2v_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_train and word in tfidf_feat:
            vec = w2v_model_train.wv[word]
            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
    train_tfidf_sent_vectors.append(sent_vec)
    row += 1
```

```
100%|██████████| 39400/39400 [10:52<00:00, 60.40it/s]
```

In [103]:
```
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

cv_tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(w2v_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_train and word in tfidf_feat:
            vec = w2v_model_train.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
    cv_tfidf_sent_vectors.append(sent_vec)
    row += 1
```

```
100%|██████████| 19407/19407 [05:20<00:00, 60.55it/s]
```

In [104]:
```
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

test_tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(w2v_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_train and word in tfidf_feat:
            vec = w2v_model_train.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
    test_tfidf_sent_vectors.append(sent_vec)
    row += 1
```

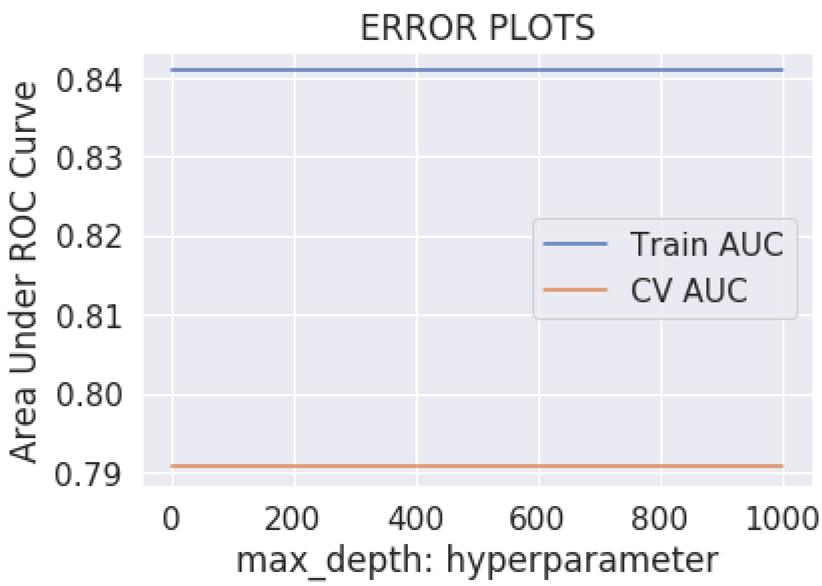
```
100%|██████████| 28966/28966 [08:00<00:00, 60.29it/s]
```

In [105]:
```
DT(train_tfidf_sent_vectors, cv_tfidf_sent_vectors)
```

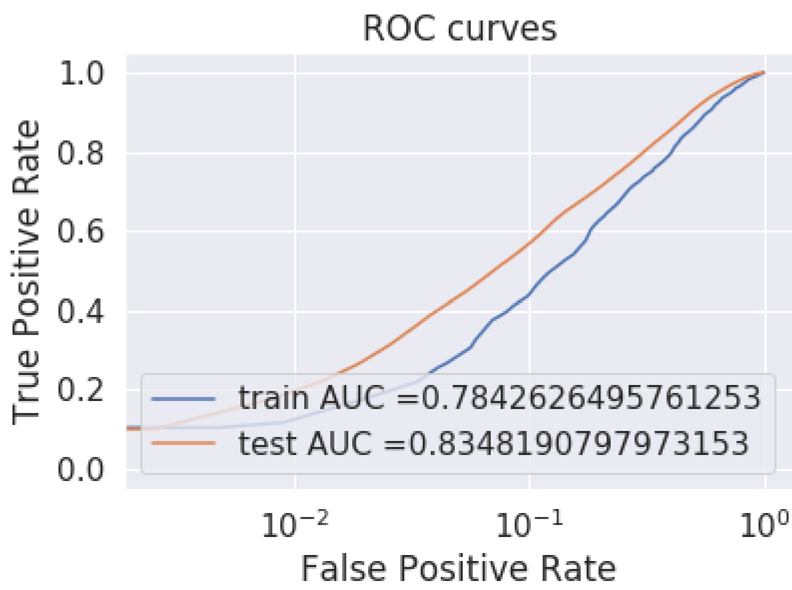
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=10,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

____

Best HyperParameter: {'max_depth': 10, 'min_samples_split': 500}
Best Accuracy: 78.43620062891476

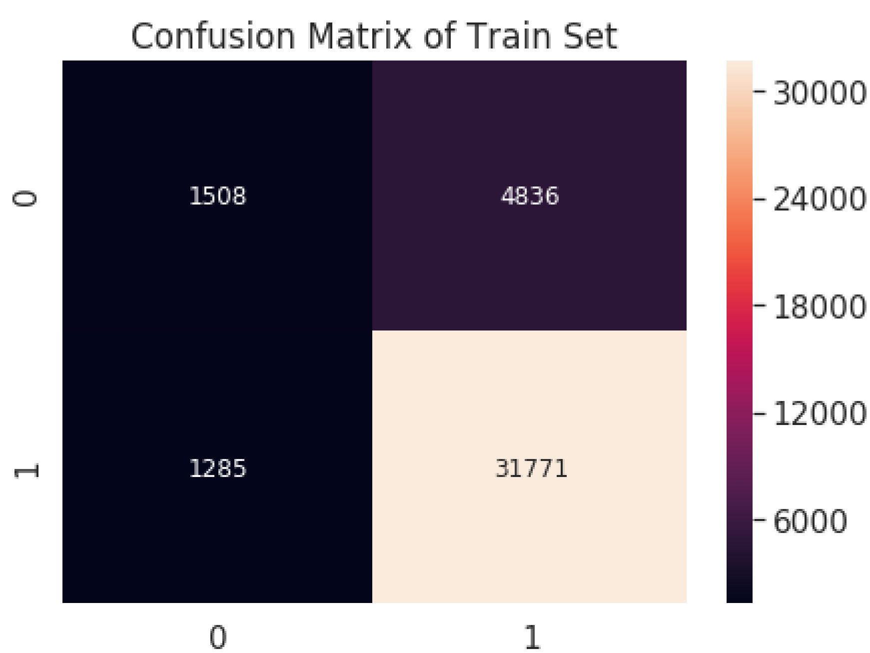


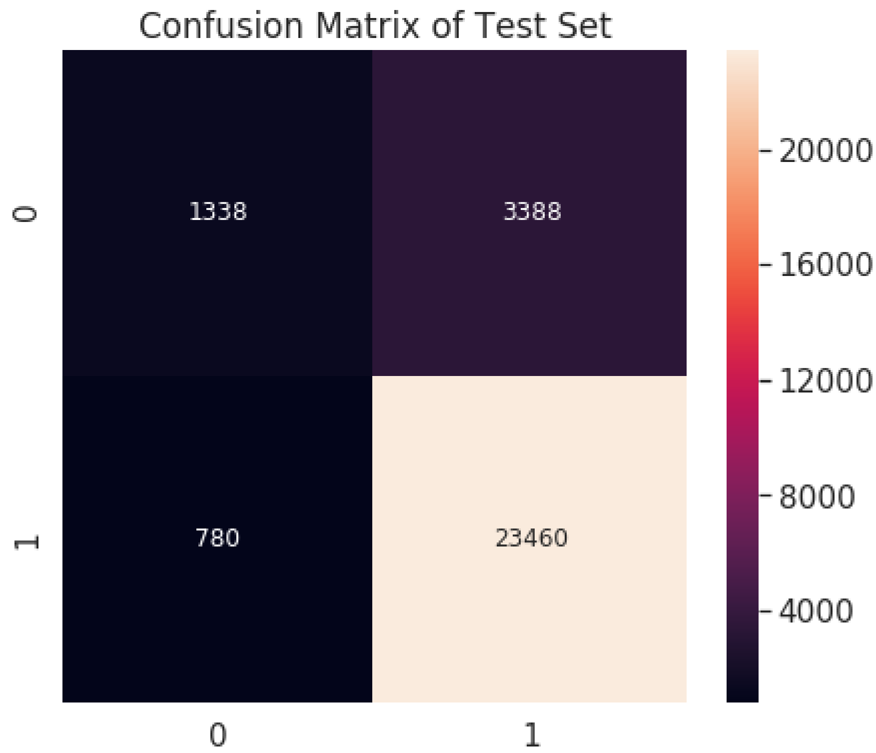
In [106]:
```
testing_DT(train_tfidf_sent_vectors, test_tfidf_sent_vectors)
```



Precision on test data: 0.87380810488677
Recall on test data: 0.9678217821782178
F1-Score on test data: 0.9184152834325086

Confusion Matrix of Train and Test set:
[[TN FP]
[FN TP]]





In [0]:

```

```

[6] Conclusions

In [2]:

```
pip install ptable
```

Requirement already satisfied: ptable in /usr/local/lib/python3.6/dist-packages (0.9.2)

In [2]:
```
from prettytable import PrettyTable

x = PrettyTable(["Vectorizer", "Model)", "Max_Depth", "Min_samples_split", "AUC"])



x.add_row(["BoW","Decision Tree" ,"50", "500", 80.78])
x.add_row(["Tf-Idf","Decision Tree" ,"50", "500", 78.54])
x.add_row(["AVG_W2V","Decision Tree" ,"10", "500", 82.10])
x.add_row(["TFIDF_W2V","Decision Tree" ,"10", "500", 78.43])

print(x)
```

```
+------------+---------------+-----------+-------------------+-------+
| Vectorizer |     Model)    | Max_Depth | Min_samples_split |  AUC  |
+------------+---------------+-----------+-------------------+-------+
|    BoW     | Decision Tree |     50    |        500        | 80.78 |
|   Tf-Idf   | Decision Tree |     50    |        500        | 78.54 |
|  AVG_W2V   | Decision Tree |     10    |        500        |  82.1 |
| TFIDF_W2V  | Decision Tree |     10    |        500        | 78.43 |
+------------+---------------+-----------+-------------------+-------+
```

Test Prob.(unseen data) using:

Word2Vec has predicted highest AUC 82.10%

In [0]: