DonorsChoose

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers is needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

- How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible
- How to increase the consistency of project vetting across different volunteers to improve the experience for teachers
- How to focus volunteer time on the applications that need the most assistance

The goal of the competition is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved, using the text of project descriptions as well as additional metadata about the project, teacher, and school. DonorsChoose.org can then use this information to identify projects most likely to need further review before approval.

About the DonorsChoose Data Set

The `train.csv` data set provided by DonorsChoose contains the following features:

Feature	Description
`project_id`	A unique identifier for the proposed project. **Example:** `p036502`
`project_title`	Title of the project. **Examples:** • `Art Will Make You Happy!` • `First Grade Fun`
`project_grade_category`	Grade level of students for which the project is targeted. One of the following enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`

Feature	Description
project_subject_categories	One or more (comma-separated) subject categories for the project from the following enumerated list of values: • `Applied Learning` • `Care & Hunger` • `Health & Sports` • `History & Civics` • `Literacy & Language` • `Math & Science` • `Music & The Arts` • `Special Needs` • `Warmth` **Examples:** • `Music & The Arts` • `Literacy & Language, Math & Science`
school_state	State where school is located (Two-letter U.S. postal code (https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations#Postal_codes)). **Example:** `WY`
project_subject_subcategories	One or more (comma-separated) subject subcategories for the project. **Examples:** • `Literacy` • `Literature & Writing, Social Sciences`
project_resource_summary	An explanation of the resources needed for the project. **Example:** • `My students need hands on literacy materials to manage sensory needs!</code`
project_essay_1	First application essay*
project_essay_2	Second application essay*
project_essay_3	Third application essay*
project_essay_4	Fourth application essay*
project_submitted_datetime	Datetime when project application was submitted. **Example:** `2016-04-28 12:43:56.245`
teacher_id	A unique identifier for the teacher of the proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c15c56`
teacher_prefix	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
teacher_number_of_previously_posted_projects	Number of project applications previously submitted by the same teacher. **Example:** `2`

* See the section **Notes on the Essay Data** for more details about these features.

Additionally, the `resources.csv` data set provides more data about the resources required for each project. Each line in this file represents a resource required by a project:

Feature	Description
`id`	A `project_id` value from the `train.csv` file. **Example:** `p036502`
`description`	Desciption of the resource. **Example:** `Tenor Saxophone Reeds, Box of 25`
`quantity`	Quantity of the resource required. **Example:** `3`
`price`	Price of the resource required. **Example:** `9.95`

Note: Many projects require multiple resources. The `id` value corresponds to a `project_id` in train.csv, so you use it as a key to retrieve all resources needed for a project:

The data set contains the following label (the value you will attempt to predict):

Label	Description
`project_is_approved`	A binary flag indicating whether DonorsChoose approved the project. A value of `0` indicates the project was not approved, and a value of `1` indicates the project was approved.

Notes on the Essay Data

Prior to May 17, 2016, the prompts for the essays were as follows:

- **project_essay_1:** "Introduce us to your classroom"
- **project_essay_2:** "Tell us more about your students"
- **project_essay_3:** "Describe how your students will use the materials you're requesting"
- **project_essay_3:** "Close by sharing why your project will make a difference"

Starting on May 17, 2016, the number of essays was reduced from 4 to 2, and the prompts for the first 2 essays were changed to the following:

- **project_essay_1:** "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."
- **project_essay_2:** "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

For all projects with project_submitted_datetime of 2016-05-17 and later, the values of project_essay_3 and project_essay_4 will be NaN.

In [132]:

```
%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
import chart_studio.plotly
# from plotly import plotly
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
from collections import Counter
from scipy.sparse import hstack,vstack
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from sklearn import model_selection
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import RandomizedSearchCV
#from sklearn.impute import SimpleImputer
from sklearn.datasets import load_digits
#from sklearn.feature_selection import SelectKBest, chi2
```

```
from sklearn.model_selection import GridSearchCV
from sklearn.feature_selection import SelectKBest,f_classif
from prettytable import PrettyTable
from sklearn.naive_bayes import MultinomialNB
from sklearn.preprocessing import Normalizer
from sklearn.metrics import confusion_matrix
#import math
import pdb
```

1.1 Reading Data

In [133]:
```
Project_data = pd.read_csv('train_data.csv')
Resource_data = pd.read_csv('resourceS.csv')
print(Project_data.shape)
print(Resource_data.shape)
```

```
(109248, 17)
(1541272, 4)
```

In [134]:
```
# how to replace elements in list python: https://stackoverflow.com/a/2582163/4084039
cols = ['Date' if x=='project_submitted_datetime' else x for x in list(Project_data.columns)]
#sort dataframe based on time pandas python: https://stackoverflow.com/a/49702492/4084039
Project_data['Date'] = pd.to_datetime(Project_data['project_submitted_datetime'])
Project_data.drop('project_submitted_datetime', axis=1, inplace=True)
Project_data.sort_values(by=['Date'], inplace=True)
# how to reorder columns pandas python: https://stackoverflow.com/a/13148611/4084039
Project_data = Project_data[cols]
Project_data.head(2)
```

Out[134]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_subject_categories	project_
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2	Math & Science	Applie
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5	Special Needs	

1.2 preprocessing of `project_subject_categories`

In [135]:
```
y = Project_data['project_is_approved'].values
Project_data.drop(['project_is_approved'], axis=1, inplace=True)
lpd = len(Project_data)
ys = np.zeros(lpd, dtype=np.int32)
X = Project_data
```

In [136]:
```
#Spliting the Dataset into three Train,CV and Test
X1, X_Test, Y1, Y_Test = train_test_split(X, y, test_size=0.33, random_state=0, stratify=ys)
nx1 = len(X1)
ys1 = np.zeros(nx1, dtype=np.int32)
X_Train, X_CV, Y_Train, Y_CV = train_test_split(X1, Y1, test_size=0.33, random_state=0, stratify=ys1)
print('Shape of the X_Train data is {0} and Y_Train data is: {1}'.format(X_Train.shape,Y_Train.shape[0]))
print('Shape of the X_CV data is    {0} and Y_CV data is    : {1}'.format(X_CV.shape,Y_CV.shape[0]))
print('Shape of the X_Test data is  {0} and Y_Test data is : {1}'.format(X_Test.shape,Y_Test.shape[0]))
```

```
Shape of the X_Train data is (49041, 16) and Y_Train data is: 49041
Shape of the X_CV data is    (24155, 16) and Y_CV data is    : 24155
Shape of the X_Test data is  (36052, 16) and Y_Test data is : 36052
```

In [137]:

```
# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python

#************************************************************Train Data**********************************************
catogories = list(X_Train['project_subject_categories'].values)
cat_list = []
for i in catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())

X_Train['clean_categories'] = cat_list
X_Train.drop(['project_subject_categories'], axis=1, inplace=True)

from collections import Counter
my_counter = Counter()
for word in X_Train['clean_categories'].values:
    my_counter.update(word.split())

cat_dict = dict(my_counter)
sorted_cat_dict_Train = dict(sorted(cat_dict.items(), key=lambda kv: kv[1]))
print(len(sorted_cat_dict_Train))
#************************************************************CV Data**********************************************
catogories = list(X_CV['project_subject_categories'].values)
cat_list = []
for i in catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())

```

```
X_CV['clean_categories'] = cat_list
X_CV.drop(['project_subject_categories'], axis=1, inplace=True)

#**************************************************************Test Data*****************************************************
catogories = list(X_Test['project_subject_categories'].values)
cat_list = []
for i in catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())

X_Test['clean_categories'] = cat_list
X_Test.drop(['project_subject_categories'], axis=1, inplace=True)

```

9

1.3 preprocessing of `project_subject_subcategories`

In [138]:

```
# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python
#************************************************************Train Data************************************************
sub_catogories = list(X_Train['project_subject_subcategories'].values)
sub_cat_list = []
for i in sub_catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp +=j.strip()+" "#" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_')
    sub_cat_list.append(temp.strip())

X_Train['clean_subcategories'] = sub_cat_list
X_Train.drop(['project_subject_subcategories'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter = Counter()
for word in X_Train['clean_subcategories'].values:
    my_counter.update(word.split())

sub_cat_dict = dict(my_counter)
sorted_sub_cat_dict_Train = dict(sorted(sub_cat_dict.items(), key=lambda kv: kv[1]))
print(len(sorted_sub_cat_dict_Train))
#************************************************************CV Data************************************************
sub_catogories = list(X_CV['project_subject_subcategories'].values)
sub_cat_list = []
for i in sub_catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp +=j.strip()+" "#" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_')
    sub_cat_list.append(temp.strip())

X_CV['clean_subcategories'] = sub_cat_list
```

```
X_CV.drop(['project_subject_subcategories'], axis=1, inplace=True)

#**************************************************************Test Data*****************************************************
sub_catogories = list(X_Test['project_subject_subcategories'].values)
sub_cat_list = []
for i in sub_catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
        temp +=j.strip()+" "#" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_')
    sub_cat_list.append(temp.strip())

X_Test['clean_subcategories'] = sub_cat_list
X_Test.drop(['project_subject_subcategories'], axis=1, inplace=True)
```

```
30
```

1.3 Text preprocessing

In [139]:

```
# merge two column text dataframe:
X_Train["essay"] = X_Train["project_essay_1"].map(str) +\
                        X_Train["project_essay_2"].map(str) + \
                        X_Train["project_essay_3"].map(str) + \
                        X_Train["project_essay_4"].map(str)

X_CV["essay"] = X_CV["project_essay_1"].map(str) +\
                        X_CV["project_essay_2"].map(str) + \
                        X_CV["project_essay_3"].map(str) + \
                        X_CV["project_essay_4"].map(str)

X_Test["essay"] = X_Test["project_essay_1"].map(str) +\
                        X_Test["project_essay_2"].map(str) + \
                        X_Test["project_essay_3"].map(str) + \
                        X_Test["project_essay_4"].map(str)
```

In [140]:

```python
# 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 [141]:

```python
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords = ['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 [142]:

```
# Combining all the above stundents
# tqdm is for printing the status bar

#----------------------------------PreProcessing of Essays in Train data set-------------------------------------------------
preprocessed_essays_Train = []
for sentance in tqdm(X_Train['essay'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    preprocessed_essays_Train.append(sent.lower().strip())
# pdb.set_trace()

#----------------------------------PreProcessing of Essays in CV data set-------------------------------------------------
preprocessed_essays_CV = []
for sentance in tqdm(X_CV['essay'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    preprocessed_essays_CV.append(sent.lower().strip())
# pdb.set_trace()

#----------------------------------PreProcessing of Essays in Test data set-------------------------------------------------
preprocessed_essays_Test = []
for sentance in tqdm(X_Test['essay'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    preprocessed_essays_Test.append(sent.lower().strip())
# pdb.set_trace()
```

```
100%|██████████████████████████████████████████████████████████| 49041/49041 [00:44<00:00, 1111.52it/s]
100%|██████████████████████████████████████████████████████████| 24155/24155 [00:20<00:00, 1158.57it/s]
```

```
100%|██████████████████████████████████████████████████████████| 36052/36052 [00:30<00:00, 1192.20it/s]
```

In [143]:

```
word_count_essay_Train = []
for a in tqdm(X_Train["essay"]) :
    b = len(a.split())
    word_count_essay_Train.append(b)

X_Train["word_count_essay_Train"] = word_count_essay_Train

word_count_essay_CV = []
for a in tqdm(X_CV["essay"]) :
    b = len(a.split())
    word_count_essay_CV.append(b)

X_CV["word_count_essay_CV"] = word_count_essay_CV

word_count_essay_Test = []
for a in tqdm(X_Test["essay"]) :
    b = len(a.split())
    word_count_essay_Test.append(b)

X_Test["word_count_essay_Test"] = word_count_essay_Test
```

```
100%|██████████████████████████████████████████████████████████| 49041/49041 [00:01<00:00, 31680.06it/s]
100%|██████████████████████████████████████████████████████████| 24155/24155 [00:01<00:00, 23588.93it/s]
100%|██████████████████████████████████████████████████████████| 36052/36052 [00:00<00:00, 36052.63it/s]
```

1.4 Preprocessing of `project_title`

In [144]:

```
# Combining all the above stundents
# tqdm is for printing the status bar

#----------------------------------PreProcessing of Project Title in Train data set-------------------------------------
preprocessed_titles_Train = []
for sentance in tqdm(X_Train['project_title'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_titles_Train.append(sent.lower().strip())
# pdb.set_trace()

#----------------------------------PreProcessing of Project Title in CV data set-------------------------------------
preprocessed_titles_CV = []
for sentance in tqdm(X_CV['project_title'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_titles_CV.append(sent.lower().strip())
# pdb.set_trace()

#----------------------------------PreProcessing of Project Title in Test data set-------------------------------------
preprocessed_titles_Test = []
for sentance in tqdm(X_Test['project_title'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_titles_Test.append(sent.lower().strip())
# pdb.set_trace()
```

```
100%|██████████████████████████████████████████████████████| 49041/49041 [00:01<00:00, 24818.33it/s]
100%|██████████████████████████████████████████████████████| 24155/24155 [00:00<00:00, 24648.09it/s]
100%|██████████████████████████████████████████████████████| 36052/36052 [00:01<00:00, 24359.48it/s]
```

In [145]:

```
word_count_title_Train = []
for a in tqdm(X_Train["project_title"]) :
    b = len(a.split())
    word_count_title_Train.append(b)

X_Train["word_count_title_Train"] = word_count_title_Train

word_count_title_CV = []
for a in tqdm(X_CV["project_title"]) :
    b = len(a.split())
    word_count_title_CV.append(b)

X_CV["word_count_title_CV"] = word_count_title_CV

word_count_title_Test = []
for a in tqdm(X_Test["project_title"]) :
    b = len(a.split())
    word_count_title_Test.append(b)

X_Test["word_count_title_Test"] = word_count_title_Test
```

```
100%|██████████| 49041/49041 [00:00<00:00, 371504.70it/s]
100%|██████████| 24155/24155 [00:00<00:00, 215648.55it/s]
100%|██████████| 36052/36052 [00:00<00:00, 450599.40it/s]
```

1.5 Preparing data for models

1.5.1 Vectorizing Categorical data

- https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/ (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/)

In [146]:

```
#------------------------Vectorizing categorical data for Train,CV and Test-----------------------------------------------

# we use count vectorizer to convert the values into one hot encoding
vectorizer_cat = CountVectorizer(vocabulary=list(sorted_cat_dict_Train.keys()), lowercase=False, binary=True)
vectorizer_cat.fit(X_Train['clean_categories'].values)
categories_one_hot_Train = vectorizer_cat.transform(X_Train['clean_categories'].values)
categories_one_hot_CV = vectorizer_cat.transform(X_CV['clean_categories'].values)
categories_one_hot_Test = vectorizer_cat.transform(X_Test['clean_categories'].values)
print(vectorizer_cat.get_feature_names())
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(categories_one_hot_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(categories_one_hot_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(categories_one_hot_Test.shape))
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds', 'Health_Sports', 'Math_Science', 'Li
teracy_Language']
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 9)
Shape of CV dataset matrix after one hot encoding is: (24155, 9)
Shape of Test dataset matrix after one hot encoding is: (36052, 9)
```

In [147]:

```
#------------------------Vectorizing Sub categorical data for Train,CV and Test-----------------------------------------

# we use count vectorizer to convert the values into one hot encoding
vectorizer_sub_cat = CountVectorizer(vocabulary=list(sorted_sub_cat_dict_Train.keys()), lowercase=False, binary=True)
vectorizer_sub_cat.fit(X_Train['clean_categories'].values)
sub_categories_one_hot_Train = vectorizer_sub_cat.transform(X_Train['clean_subcategories'].values)
sub_categories_one_hot_CV = vectorizer_sub_cat.transform(X_CV['clean_subcategories'].values)
sub_categories_one_hot_Test = vectorizer_sub_cat.transform(X_Test['clean_subcategories'].values)
print(vectorizer_sub_cat.get_feature_names())
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(sub_categories_one_hot_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(sub_categories_one_hot_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(sub_categories_one_hot_Test.shape))
```

```
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular', 'Civics_Government', 'ForeignLangu
ages', 'Warmth', 'Care_Hunger', 'NutritionEducation', 'SocialSciences', 'PerformingArts', 'CharacterEducation', 'TeamSports', 'Ot
her', 'College_CareerPrep', 'Music', 'History_Geography', 'EarlyDevelopment', 'Health_LifeScience', 'ESL', 'Gym_Fitness', 'Enviro
nmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedSciences', 'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Litera
cy']
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 30)
Shape of CV dataset matrix after one hot encoding is: (24155, 30)
Shape of Test dataset matrix after one hot encoding is: (36052, 30)
```

School State

In [148]:

```
#------------------------Vectorizing categorical data of School state for Train dataset-----------------------------------

school_catogories_Train = list(X_Train['school_state'].values)
school_list_Train = []
for sent in school_catogories_Train:
    school_list_Train.append(sent.lower().strip())
X_Train['school_categories'] = school_list_Train
X_Train.drop(['school_state'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_school_Train = Counter()
for word in X_Train['school_categories'].values:
    my_counter_school_Train.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
school_dict_Train = dict(my_counter_school_Train)
sorted_school_dict_Train = dict(sorted(school_dict_Train.items(), key=lambda kv: kv[1]))

vectorizer_school = CountVectorizer(vocabulary=list(sorted_school_dict_Train.keys()), lowercase=False, binary=True)
vectorizer_school.fit(X_Train['school_categories'].values)
#print(vectorizer.get_feature_names())

school_one_hot_Train = vectorizer_school.transform(X_Train['school_categories'].values)

#------------------------Vectorizing categorical data of School state for CV dataset-----------------------------------

school_catogories_CV = list(X_CV['school_state'].values)
school_list_CV = []
for sent in school_catogories_CV:
    school_list_CV.append(sent.lower().strip())
X_CV['school_categories'] = school_list_CV
X_CV.drop(['school_state'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_school_CV = Counter()
for word in X_CV['school_categories'].values:
    my_counter_school_CV.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
school_dict_CV = dict(my_counter_school_CV)
sorted_school_dict_CV = dict(sorted(school_dict_CV.items(), key=lambda kv: kv[1]))
school_one_hot_CV = vectorizer_school.transform(X_CV['school_categories'].values)

#------------------------Vectorizing categorical data of School state for Test dataset-----------------------------------

```

```
school_catogories_Test = list(X_Test['school_state'].values)
school_list_Test = []
for sent in school_catogories_Test:
    school_list_Test.append(sent.lower().strip())
X_Test['school_categories'] = school_list_Test
X_Test.drop(['school_state'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_school_Test = Counter()
for word in X_Test['school_categories'].values:
    my_counter_school_Test.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
school_dict_Test = dict(my_counter_school_Test)
sorted_school_dict_Test = dict(sorted(school_dict_Test.items(), key=lambda kv: kv[1]))
school_one_hot_Test = vectorizer_school.transform(X_Test['school_categories'].values)
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(school_one_hot_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(school_one_hot_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(school_one_hot_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 51)
Shape of CV dataset matrix after one hot encoding is: (24155, 51)
Shape of Test dataset matrix after one hot encoding is: (36052, 51)
```

Prefix

In [149]:

```
#------------------------Vectorizing categorical data of Teacher Prefix for Train dataset-----------------------------------

# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039
# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python
prefix_catogories_Train = list(X_Train['teacher_prefix'].values)
prefix_list_Train = []
for sent in prefix_catogories_Train:
    sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
    prefix_list_Train.append(sent.lower().strip())
X_Train['prefix_catogories'] = prefix_list_Train
X_Train.drop(['teacher_prefix'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_prefix_Train = Counter()
for word in X_Train['prefix_catogories'].values:
    my_counter_prefix_Train.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
prefix_dict_Train = dict(my_counter_prefix_Train)
sorted_prefix_dict_Train = dict(sorted(prefix_dict_Train.items(), key=lambda kv: kv[1]))


vectorizer_prefix = CountVectorizer(vocabulary=list(sorted_prefix_dict_Train.keys()), lowercase=False, binary=True)
vectorizer_prefix.fit(X_Train['prefix_catogories'].values)
#print(vectorizer.get_feature_names())

prefix_one_hot_Train = vectorizer_prefix.transform(X_Train['prefix_catogories'].values)
#print("Shape of matrix after one hot encodig ",prefix_one_hot.shape)

#------------------------Vectorizing categorical data of Teacher Prefix for CV dataset-----------------------------------

prefix_catogories_CV = list(X_CV['teacher_prefix'].values)
prefix_list_CV = []
for sent in prefix_catogories_CV:
    sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
    prefix_list_CV.append(sent.lower().strip())
X_CV['prefix_catogories'] = prefix_list_CV
X_CV.drop(['teacher_prefix'], axis=1, inplace=True)

```

```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_prefix_CV = Counter()
for word in X_CV['prefix_catogories'].values:
    my_counter_prefix_CV.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
prefix_dict_CV = dict(my_counter_prefix_CV)
sorted_prefix_dict_CV = dict(sorted(prefix_dict_CV.items(), key=lambda kv: kv[1]))
prefix_one_hot_CV = vectorizer_prefix.transform(X_CV['prefix_catogories'].values)

#------------------------Vectorizing categorical data of Teacher Prefix for Test dataset-----------------------------------

prefix_catogories_Test = list(X_Test['teacher_prefix'].values)
prefix_list_Test = []
for sent in prefix_catogories_Test:
    sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
    prefix_list_Test.append(sent.lower().strip())
X_Test['prefix_catogories'] = prefix_list_Test
X_Test.drop(['teacher_prefix'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_prefix_Test = Counter()
for word in X_Test['prefix_catogories'].values:
    my_counter_prefix_Test.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
prefix_dict_Test = dict(my_counter_prefix_Test)
sorted_prefix_dict_Test = dict(sorted(prefix_dict_Test.items(), key=lambda kv: kv[1]))
prefix_one_hot_Test = vectorizer_prefix.transform(X_Test['prefix_catogories'].values)
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(prefix_one_hot_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(prefix_one_hot_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(prefix_one_hot_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 6)
Shape of CV dataset matrix after one hot encoding is: (24155, 6)
Shape of Test dataset matrix after one hot encoding is: (36052, 6)
```

project_grade_category

In [150]:

```
#------------------------Vectorizing categorical data of Project Grade for Train dataset----------------------------------

# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039
# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python
grade_catogories_Train = list(X_Train['project_grade_category'].values)
grade_list_Train = []
for sent in grade_catogories_Train:
    sent = sent.replace('-','_')
    sent = sent.replace(' ','_')
    # sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
    grade_list_Train.append(sent.lower().strip())

# temp = temp.replace('-','_')
X_Train['new_grade_category'] = grade_list_Train
X_Train.drop(['project_grade_category'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_grade_Train = Counter()
for word in X_Train['new_grade_category'].values:
    my_counter_grade_Train.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
grade_dict_Train = dict(my_counter_grade_Train)
sorted_grade_dict_Train = dict(sorted(grade_dict_Train.items(), key=lambda kv: kv[1]))

vectorizer_grade = CountVectorizer(vocabulary=list(sorted_grade_dict_Train.keys()), lowercase=False, binary=True)
vectorizer_grade.fit(X_Train['new_grade_category'].values)
#print(vectorizer.get_feature_names())

grade_one_hot_Train = vectorizer_grade.transform(X_Train['new_grade_category'].values)

#------------------------Vectorizing categorical data of Project Grade for CV dataset----------------------------------

grade_catogories_CV = list(X_CV['project_grade_category'].values)
grade_list_CV = []
for sent in grade_catogories_CV:
    sent = sent.replace('-','_')
    sent = sent.replace(' ','_')
    # sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
```

```
    grade_list_CV.append(sent.lower().strip())

# temp = temp.replace('-','_')
X_CV['new_grade_category'] = grade_list_CV
X_CV.drop(['project_grade_category'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_grade_CV = Counter()
for word in X_CV['new_grade_category'].values:
    my_counter_grade_CV.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
grade_dict_CV = dict(my_counter_grade_CV)
sorted_grade_dict_CV = dict(sorted(grade_dict_CV.items(), key=lambda kv: kv[1]))

grade_one_hot_CV = vectorizer_grade.transform(X_CV['new_grade_category'].values)

#-------------------------Vectorizing categorical data of Project Grade for Train dataset----------------------------------

grade_catogories_Test = list(X_Test['project_grade_category'].values)
grade_list_Test = []
for sent in grade_catogories_Test:
    sent = sent.replace('-','_')
    sent = sent.replace(' ','_')
    # sent = re.sub('[^A-Za-z0-9]+', ' ', str(sent))
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split())
    grade_list_Test.append(sent.lower().strip())

# temp = temp.replace('-','_')
X_Test['new_grade_category'] = grade_list_Test
X_Test.drop(['project_grade_category'], axis=1, inplace=True)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter_grade_Test = Counter()
for word in X_Test['new_grade_category'].values:
    my_counter_grade_Test.update(word.split())

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
grade_dict_Test = dict(my_counter_grade_Test)
sorted_grade_dict_Test = dict(sorted(grade_dict_Test.items(), key=lambda kv: kv[1]))

grade_one_hot_Test = vectorizer_grade.transform(X_Test['new_grade_category'].values)
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(grade_one_hot_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(grade_one_hot_CV.shape))
```

```
92 print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(grade_one_hot_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 4)
Shape of CV dataset matrix after one hot encoding is: (24155, 4)
Shape of Test dataset matrix after one hot encoding is: (36052, 4)
```

1.5.2 Vectorizing Numerical features

In [151]:
```
1  price_data = Resource_data.groupby('id').agg({'price':'sum', 'quantity':'sum'}).reset_index()
2  X_Train = pd.merge(X_Train, price_data, on='id', how='left')
3  X_CV = pd.merge(X_CV, price_data, on='id', how='left')
4  X_Test = pd.merge(X_Test, price_data, on='id', how='left')
```

In [152]:
```
1  price_norm = Normalizer(norm='l2', copy=False)
2  price_norm.fit(X_Train['price'].values.reshape(1,-1))
3
4  p=price_norm.transform(X_Train['price'].values.reshape(1,-1))
5  price_norm.transform(X_CV['price'].values.reshape(1,-1))
6  price_norm.transform(X_Test['price'].values.reshape(1,-1))
7  price_norm_Train = (X_Train['price'].values.reshape(-1,1))
8  price_norm_CV = (X_CV['price'].values.reshape(-1,1))
9  price_norm_Test = (X_Test['price'].values.reshape(-1,1))
10 print("-"*120)
11 print('Shape of Train normalized price dataset matrix after one hot encoding is: {0}'.format(price_norm_Train.shape))
12 print('Shape of CV normalized price dataset matrix after one hot encoding is: {0}'.format(price_norm_CV.shape))
13 print('Shape of Test normalized price dataset matrix after one hot encoding is: {0}'.format(price_norm_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train normalized price dataset matrix after one hot encoding is: (49041, 1)
Shape of CV normalized price dataset matrix after one hot encoding is: (24155, 1)
Shape of Test normalized price dataset matrix after one hot encoding is: (36052, 1)
```

In [153]:

```
quantity_norm = Normalizer(norm='l2', copy=False)
quantity_norm.fit(X_Train['quantity'].values.reshape(1,-1))

quantity_norm.transform(X_Train['quantity'].values.reshape(1,-1))
quantity_norm.transform(X_CV['quantity'].values.reshape(1,-1))
quantity_norm.transform(X_Test['quantity'].values.reshape(1,-1))
quantity_norm_Train = quantity_norm.transform(X_Train['quantity'].values.reshape(-1,1))
quantity_norm_CV = quantity_norm.transform(X_CV['quantity'].values.reshape(-1,1))
quantity_norm_Test = quantity_norm.transform(X_Test['quantity'].values.reshape(-1,1))
print("-"*120)
print('Shape of Train normalized quantity dataset matrix after one hot encoding is: {0}'.format(quantity_norm_Train.shape))
print('Shape of CV normalized quantity dataset matrix after one hot encoding is: {0}'.format(quantity_norm_CV.shape))
print('Shape of Test normalized quantity dataset matrix after one hot encoding is: {0}'.format(quantity_norm_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train normalized quantity dataset matrix after one hot encoding is: (49041, 1)
Shape of CV normalized quantity dataset matrix after one hot encoding is: (24155, 1)
Shape of Test normalized quantity dataset matrix after one hot encoding is: (36052, 1)
```

In [154]:

```
teacher_prev_post_norm = Normalizer(norm='l2', copy=False)
teacher_prev_post_norm.fit(X_Train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))

teacher_prev_post_norm.transform(X_Train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
teacher_prev_post_norm.transform(X_CV['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
teacher_prev_post_norm.transform(X_Test['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
teacher_prev_post_norm_Train = teacher_prev_post_norm.transform(X_Train['teacher_number_of_previously_posted_projects'].values
teacher_prev_post_norm_CV = teacher_prev_post_norm.transform(X_CV['teacher_number_of_previously_posted_projects'].values.reshap
teacher_prev_post_norm_Test = teacher_prev_post_norm.transform(X_Test['teacher_number_of_previously_posted_projects'].values.re
print("-"*120)
print('Shape of Train normalized previously posted project dataset matrix after one hot encoding is: {0}'.format(teacher_prev_p
print('Shape of CV normalized previously posted project dataset matrix after one hot encoding is: {0}'.format(teacher_prev_post
print('Shape of Test normalized previously posted project dataset matrix after one hot encoding is: {0}'.format(teacher_prev_p
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train normalized previously posted project dataset matrix after one hot encoding is: (49041, 1)
Shape of CV normalized previously posted project dataset matrix after one hot encoding is: (24155, 1)
Shape of Test normalized previously posted project dataset matrix after one hot encoding is: (36052, 1)
```

In [155]:
```
title_norm = Normalizer(norm='l2', copy=False)
title_norm.fit(X_Train['word_count_title_Train'].values.reshape(1,-1))
title_norm.transform(X_Train['word_count_title_Train'].values.reshape(1,-1))
title_norm.transform(X_CV['word_count_title_CV'].values.reshape(1,-1))
title_norm.transform(X_Test['word_count_title_Test'].values.reshape(1,-1))
word_count_title_Train = title_norm.transform(X_Train['word_count_title_Train'].values.reshape(-1,1))
word_count_title_CV = title_norm.transform(X_CV['word_count_title_CV'].values.reshape(-1,1))
word_count_title_Test = title_norm.transform(X_Test['word_count_title_Test'].values.reshape(-1,1))
print("-"*120)
print('Shape of Train normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_title_Train.shape))
print('Shape of CV normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_title_CV.shape))
print('Shape of Test normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_title_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train normalized title dataset matrix after one hot encoding is: (49041, 1)
Shape of CV normalized title dataset matrix after one hot encoding is: (24155, 1)
Shape of Test normalized title dataset matrix after one hot encoding is: (36052, 1)
```

In [156]:
```
essay_norm = Normalizer(norm='l2', copy=False)
essay_norm.fit(X_Train['word_count_essay_Train'].values.reshape(1,-1))
essay_norm.transform(X_Train['word_count_essay_Train'].values.reshape(1,-1))
essay_norm.transform(X_CV['word_count_essay_CV'].values.reshape(1,-1))
essay_norm.transform(X_Test['word_count_essay_Test'].values.reshape(1,-1))
word_count_essay_Train = essay_norm.transform(X_Train['word_count_essay_Train'].values.reshape(-1,1))
word_count_essay_CV = essay_norm.transform(X_CV['word_count_essay_CV'].values.reshape(-1,1))
word_count_essay_Test = essay_norm.transform(X_Test['word_count_essay_Test'].values.reshape(-1,1))
print("-"*120)
print('Shape of Train normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_essay_Train.shape))
print('Shape of CV normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_essay_CV.shape))
print('Shape of Test normalized title dataset matrix after one hot encoding is: {0}'.format(word_count_essay_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Shape of Train normalized title dataset matrix after one hot encoding is: (49041, 1)
Shape of CV normalized title dataset matrix after one hot encoding is: (24155, 1)
Shape of Test normalized title dataset matrix after one hot encoding is: (36052, 1)
```

1.5.3 Vectorizing Text data

1.5.3.1 Bag of words

In [157]:
```
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer_essays_bow = CountVectorizer(min_df=10)
text_bow_Train = vectorizer_essays_bow.fit_transform(preprocessed_essays_Train)
text_bow_CV = vectorizer_essays_bow.transform(preprocessed_essays_CV)
text_bow_Test = vectorizer_essays_bow.transform(preprocessed_essays_Test)
print("-"*120)
print("Applying Bag Of Words for Text Data")
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(text_bow_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(text_bow_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(text_bow_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Applying Bag Of Words for Text Data
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 12097)
Shape of CV dataset matrix after one hot encoding is: (24155, 12097)
Shape of Test dataset matrix after one hot encoding is: (36052, 12097)
```

Bag of Words for Project Title

In [158]:
```
# you can vectorize the title also
# before you vectorize the title make sure you preprocess it
vectorizer_titles_bow = CountVectorizer(min_df=10)
title_bow_Train = vectorizer_titles_bow.fit_transform(preprocessed_titles_Train)
title_bow_CV = vectorizer_titles_bow.transform(preprocessed_titles_CV)
title_bow_Test = vectorizer_titles_bow.transform(preprocessed_titles_Test)
print("-"*120)
print("Applying Bag Of Words for Project Title Data")
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(title_bow_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(title_bow_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(title_bow_Test.shape))

```

```
------------------------------------------------------------------------------------------------------------------------
Applying Bag Of Words for Project Title Data
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 2083)
Shape of CV dataset matrix after one hot encoding is: (24155, 2083)
Shape of Test dataset matrix after one hot encoding is: (36052, 2083)
```

1.5.2.2 TFIDF vectorizer

In [159]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer_essays_tfidf = TfidfVectorizer(min_df=10)
text_tfidf_Train = vectorizer_essays_tfidf.fit_transform(preprocessed_essays_Train)
text_tfidf_CV = vectorizer_essays_tfidf.transform(preprocessed_essays_CV)
text_tfidf_Test = vectorizer_essays_tfidf.transform(preprocessed_essays_Test)
print("-"*120)
print("Applying TFIDF for Text Data")
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(text_tfidf_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(text_tfidf_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(text_tfidf_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Applying TFIDF for Text Data
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 12097)
Shape of CV dataset matrix after one hot encoding is: (24155, 12097)
Shape of Test dataset matrix after one hot encoding is: (36052, 12097)
```

TFIDF vectorizer for Project Title

In [160]:
```
vectorizer_titles_tfidf = TfidfVectorizer(min_df=10)
title_tfidf_Train = vectorizer_titles_tfidf.fit_transform(preprocessed_titles_Train)
title_tfidf_CV = vectorizer_titles_tfidf.transform(preprocessed_titles_CV)
title_tfidf_Test = vectorizer_titles_tfidf.transform(preprocessed_titles_Test)
print("-"*120)
print("Applying TFIDF for Project Title")
print("-"*120)
print('Shape of Train dataset matrix after one hot encoding is: {0}'.format(title_tfidf_Train.shape))
print('Shape of CV dataset matrix after one hot encoding is: {0}'.format(title_tfidf_CV.shape))
print('Shape of Test dataset matrix after one hot encoding is: {0}'.format(title_tfidf_Test.shape))
```

```
------------------------------------------------------------------------------------------------------------------------
Applying TFIDF for Project Title
------------------------------------------------------------------------------------------------------------------------
Shape of Train dataset matrix after one hot encoding is: (49041, 2083)
Shape of CV dataset matrix after one hot encoding is: (24155, 2083)
Shape of Test dataset matrix after one hot encoding is: (36052, 2083)
```

Assignment 4: Naive Bayes

1. **Apply Multinomial NaiveBayes on these feature sets**

- Set 1: categorical, numerical features + project_title(BOW) + preprocessed_eassay (BOW)
- Set 2: categorical, numerical features + project_title(TFIDF)+ preprocessed_eassay (TFIDF)

2. **The hyper paramter tuning(find best Alpha)**

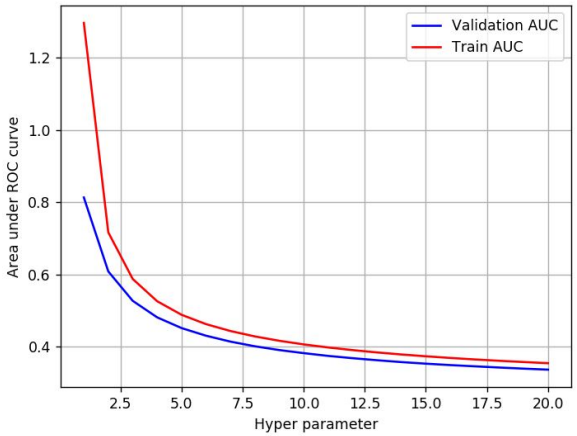
- Find the best hyper parameter which will give the maximum AUC (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/receiver-operating-characteristic-curve-roc-curve-and-auc-1/) value
- Consider a wide range of alpha values for hyperparameter tuning, start as low as 0.00001
- Find 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. **Feature importance**

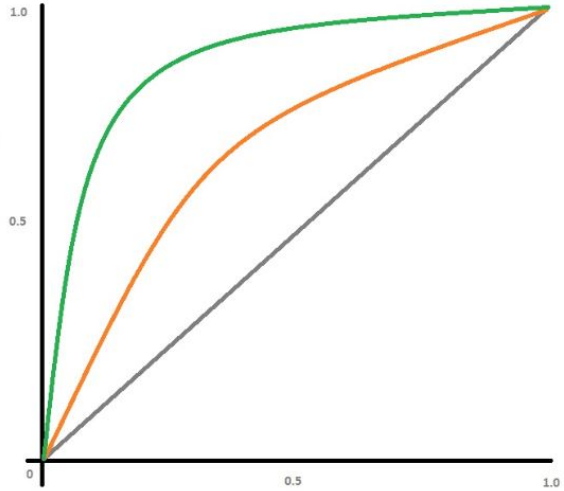
- Find the top 10 features of positive class and top 10 features of negative class for both feature sets Set 1 and Set 2 using values of `feature_log_prob_` parameter of MultinomialNB (https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html) and print their corresponding feature names

4. **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. Here on X-axis you will have alpha values, since they have a wide range, just to represent those alpha values on the graph, apply log function on those alpha values.



- 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.

	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

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

5. **Conclusion**

- You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link (http://zetcode.com/python/prettytable/)

Vectorizer	Model	Hyper parameter	AUC
BOW	Brute	7	0.78
TFIDF	Brute	12	0.79
W2V	Brute	10	0.78
TFIDFW2V	Brute	6	0.78

1.5.4 Merging all the above features

- we need to merge all the numerical vectors i.e catogorical, text, numerical vectors

In [161]:
```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
BOW_Train = hstack((categories_one_hot_Train,sub_categories_one_hot_Train,school_one_hot_Train,grade_one_hot_Train,prefix_one_h
print(BOW_Train.shape)
TFIDF_Train = hstack((categories_one_hot_Train,sub_categories_one_hot_Train,school_one_hot_Train,grade_one_hot_Train,prefix_one
print(TFIDF_Train.shape)
```

```
(49041, 14285)
(49041, 14285)
```

In [162]:
```
BOW_CV = hstack((categories_one_hot_CV,sub_categories_one_hot_CV,school_one_hot_CV,grade_one_hot_CV,prefix_one_hot_CV,text_bow_
print(BOW_CV.shape)
TFIDF_CV = hstack((categories_one_hot_CV,sub_categories_one_hot_CV,school_one_hot_CV,grade_one_hot_CV,prefix_one_hot_CV,text_tf
print(TFIDF_CV.shape)
```

```
(24155, 14285)
(24155, 14285)
```

In [163]:
```
BOW_Test = hstack((categories_one_hot_Test,sub_categories_one_hot_Test,school_one_hot_Test,grade_one_hot_Test,prefix_one_hot_Te
print(BOW_Test.shape)
TFIDF_Test = hstack((categories_one_hot_Test,sub_categories_one_hot_Test,school_one_hot_Test,grade_one_hot_Test,prefix_one_hot_
print(TFIDF_Test.shape)
```

```
(36052, 14285)
(36052, 14285)
```

Loading Pickle files

2. Naive Bayes

2.4 Appling NB() on different kind of featurization as mentioned in the instructions

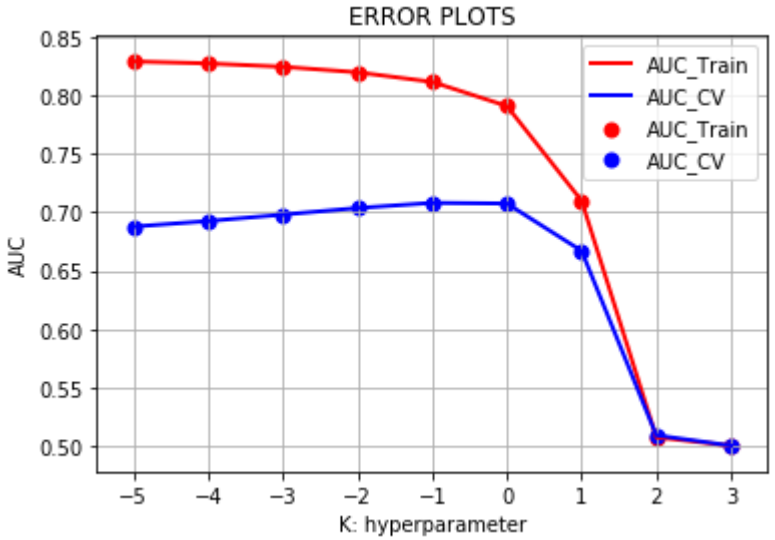
Apply Naive Bayes on different kind of featurization as mentioned in the instructions
For Every model that you work on make sure you do the step 2 and step 3 of instrucations

2.4.1 Applying Naive Bayes on BOW, SET 1

In [164]:

```
%%time

BOW_TR_CSR = BOW_Train.tocsr()
BOW_CV_CSR = BOW_CV.tocsr()
BOW_TS_CSR = BOW_Test.tocsr()

alpha = [0.00001, 0.0001, 0.001, 0.01,0.1, 1, 10,100,1000]
L_ALPHA= []
ACCV = []
AUC_TR = []
AUC_CV = []

for i in tqdm(alpha):
    nb = MultinomialNB(alpha = i,fit_prior=True,class_prior=[0.5,0.5])
    # pdb.set_trace()
    nb.fit(BOW_Train, Y_Train)
    pred = nb.predict(BOW_CV)
    acc = accuracy_score(Y_CV, pred, normalize=True) * float(100)
    ACCV.append(acc)
    a_fpr_train, a_tpr_train, thresholds = roc_curve(Y_Train, nb.predict_proba(BOW_TR_CSR) [:,1])
    AUC_TR.append(auc(a_fpr_train, a_tpr_train))

    a_fpr_cv, a_tpr_cv, thresholds = roc_curve(Y_CV, nb.predict_proba(BOW_CV_CSR) [:,1])
    AUC_CV.append(auc(a_fpr_cv, a_tpr_cv))


for av in tqdm(alpha):
    b = np.log10(av)
    L_ALPHA.append(b)

# Performance of model on Train data and Test data for each hyper parameter.
plt.plot(L_ALPHA, AUC_TR, label='AUC_Train',color='red',linewidth=2)
plt.scatter(L_ALPHA, AUC_TR, label='AUC_Train',color='red',linewidth=2)
plt.gca()
plt.plot(L_ALPHA, AUC_CV, label='AUC_CV',color='blue',linewidth=2)
plt.scatter(L_ALPHA, AUC_CV, label='AUC_CV',color='blue',linewidth=2)
plt.gca()
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```

```
100%|██████████| 9/9 [00:06<00:00,  1.43it/s]
100%|██████████| 9/9 [00:00<00:00, 2249.36it/s]
```




```
Wall time: 21.4 s
Compiler : 376 ms
Parser   : 291 ms
```

In [165]:

```
print(AUC_TR)
print(AUC_CV)
```

```
[0.8294012820005231, 0.8276958244116213, 0.8248167281856434, 0.8200770452977412, 0.811853158302271, 0.7911488255143547, 0.7099457
736235792, 0.5071686477789238, 0.5]
[0.6878082577138354, 0.692571394489643, 0.6980880457054459, 0.7037202471954482, 0.7081541941967424, 0.7075945397845345, 0.6669896
607278614, 0.5088766907673291, 0.5]
```

OBSERVATION:

1. Optimal alpha has been choose based on the alpha which maximize the AUC CV score.
2. In this case **α(alpha)=0.1** is having the highest AUC CV score.
3. Hence **α(alpha)=0.1** is choosen as the optimal alpha value.

In [166]:

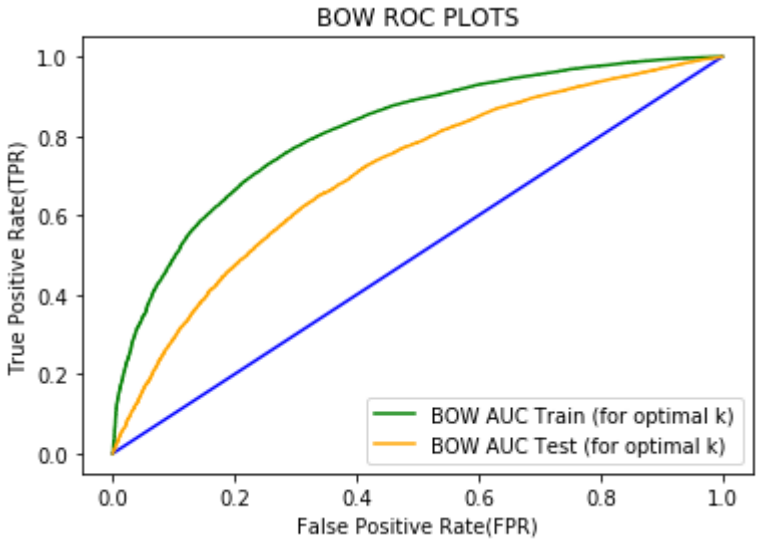
```
A_OPT=0.1
BOW_opt = MultinomialNB(alpha=A_OPT,fit_prior=True,class_prior=[0.5,0.5])
BOW_opt.fit(BOW_Train, Y_Train)
pred = BOW_opt.predict(BOW_Test)
acc = accuracy_score(Y_Test, pred, normalize=True) * float(100)
print('\nTest accuracy for alpha(a) = {0} is {1}%'.format(A_OPT,acc))

a_fpr_train, a_tpr_train, thresholds = roc_curve(Y_Train, BOW_opt.predict_proba(BOW_TR_CSR) [:,1])
a_fpr_Test, a_tpr_Test, thresholds = roc_curve(Y_Test, BOW_opt.predict_proba(BOW_TS_CSR)[:,1])
```

```
Test accuracy for alpha(a) = 0.1 is 69.73815599689338%
```

BOW ROC PLOT

In [167]:

```
%%time

#https://stackoverflow.com/questions/52910061/implementing-roc-curves-for-k-nn-machine-learning-algorithm-using-python-and-sci

plt.plot([0,1],[0,1],'k-', color='blue')
plt.plot(a_fpr_train, a_tpr_train, label="BOW AUC Train (for optimal k)", color='green')
plt.plot(a_fpr_Test, a_tpr_Test, label="BOW AUC Test (for optimal k)", color='orange')
plt.legend()
plt.ylabel("True Positive Rate(TPR)")
plt.xlabel("False Positive Rate(FPR)")
plt.title("BOW ROC PLOTS")
plt.show()
print("-"*120)
print("AUC Train (for optimal k) =", auc(a_fpr_train, a_tpr_train))
print("AUC Test (for optimal k) =", auc(a_fpr_Test, a_tpr_Test))
BOW_AOPT=A_OPT
BOW_AUC=round(auc(a_fpr_Test, a_tpr_Test)*100)
pred1 = BOW_opt.predict(BOW_Train)
pred2 = BOW_opt.predict(BOW_Test)

```



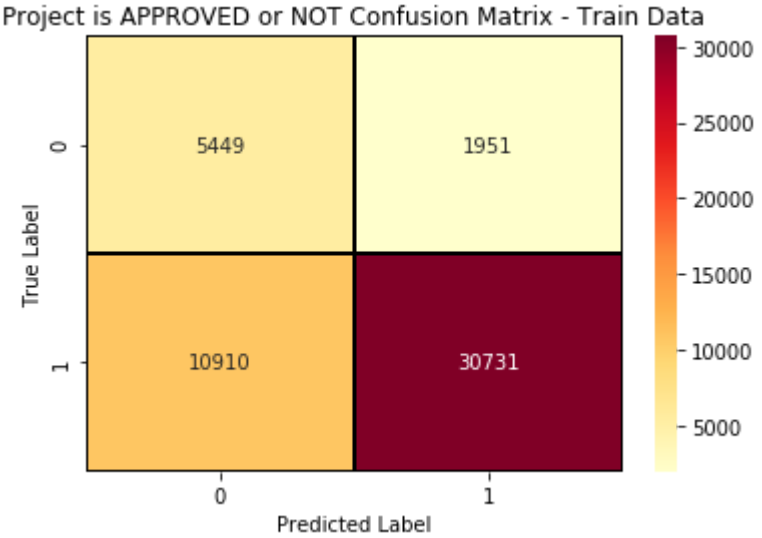

```
------------------------------------------------------------------------------------------------------------------------
AUC Train (for optimal k) = 0.811853158302271
AUC Test (for optimal k) = 0.7047236730122236
Wall time: 1.21 s
```

BOW CONFUSION MATRIX

In [168]:
```
%%time
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
#https://getaravind.com/blog/confusion-matrix-seaborn-heatmap/
%matplotlib inline
from sklearn.metrics import confusion_matrix
Train = confusion_matrix(Y_Train, pred1)
sns.heatmap(Train,annot=True,cbar=True,fmt='d',cmap='YlOrRd',linewidths=1,linecolor='black')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Project is APPROVED or NOT Confusion Matrix - Train Data')
```

```
Wall time: 1.24 s
```

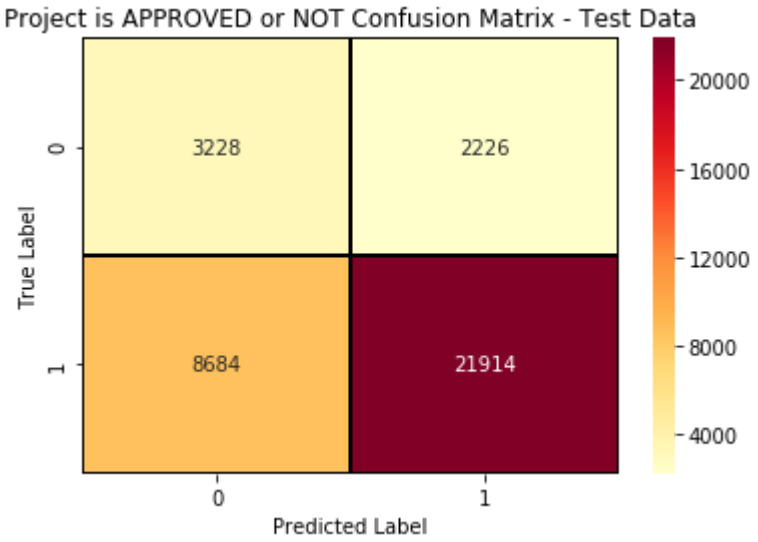
Out[168]:
```
Text(0.5, 1, 'Project is APPROVED or NOT Confusion Matrix - Train Data')
```



In [169]:
```
%%time
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
#https://getaravind.com/blog/confusion-matrix-seaborn-heatmap/
Test = confusion_matrix(Y_Test, pred2)
sns.heatmap(Test,annot=True,cbar=True,fmt='d',cmap='YlOrRd',linewidths=1,linecolor='black')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Project is APPROVED or NOT Confusion Matrix - Test Data')
```

```
Wall time: 240 ms
```

Out[169]:
```
Text(0.5, 1, 'Project is APPROVED or NOT Confusion Matrix - Test Data')
```



2.4.1.1 Top 10 important features of positive class from SET 1

In [170]:
```
B_NB = MultinomialNB(alpha = A_OPT, class_prior=[0.5,0.5])
B_NB.fit(BOW_TR_CSR, Y_Train)

BOW_F = []

for BOW in vectorizer_cat.get_feature_names() :
    BOW_F.append(BOW)
for BOW1 in vectorizer_sub_cat.get_feature_names() :
    BOW_F.append(BOW1)
for BOW2 in vectorizer_school.get_feature_names() :
    BOW_F.append(BOW2)
for BOW3 in vectorizer_grade.get_feature_names() :
    BOW_F.append(BOW3)
for BOW4 in vectorizer_prefix.get_feature_names() :
    BOW_F.append(BOW4)
for BOW5 in vectorizer_essays_bow.get_feature_names() :
    BOW_F.append(BOW5)
for BOW6 in vectorizer_titles_bow.get_feature_names() :
    BOW_F.append(BOW6)
BOW_F.append("price")
BOW_F.append("quantity")
BOW_F.append("prev_proposed_projects")
BOW_F.append("title_word_count")
BOW_F.append("essay_word_count")
len(BOW_F)
```

Out[170]: 14285

In [171]:
```
POS_F_BOW = B_NB.feature_log_prob_[1, :].argsort()[::-1]
for i in POS_F_BOW[:10]:
    print(BOW_F[i])
```

```
students
school
learning
classroom
not
learn
help
essay_word_count
title_word_count
quantity
```

2.4.1.2 Top 10 important features of negative class from SET 1

In [172]:

```
NEG_F_BOW = B_NB.feature_log_prob_[0, :].argsort()[::-1]
for i in NEG_F_BOW[:10]:
    print(BOW_F[i])
```

```
students
school
learning
classroom
not
learn
help
essay_word_count
title_word_count
quantity
```

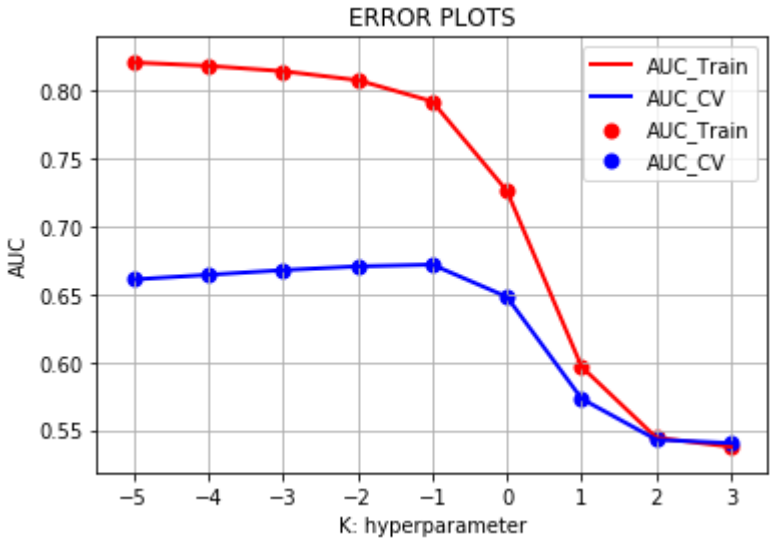
2.4.2 Applying Naive Bayes on TFIDF, SET 2

In [173]:

```
%%time

TFIDF_TR_CSR = TFIDF_Train.tocsr()
TFIDF_CV_CSR = TFIDF_CV.tocsr()
TFIDF_TS_CSR = TFIDF_Test.tocsr()

alpha = [0.00001, 0.0001, 0.001, 0.01,0.1,1,10,100,1000]
L_ALPHA= []
ACCV = []
AUC_TR = []
AUC_CV = []

for i in tqdm(alpha):
    nb = MultinomialNB(alpha = i,fit_prior=True,class_prior=[0.5,0.5])
    # pdb.set_trace()
    nb.fit(TFIDF_Train, Y_Train)
    pred = nb.predict(TFIDF_CV)
    acc = accuracy_score(Y_CV, pred, normalize=True) * float(100)
    ACCV.append(acc)
    a_fpr_train, a_tpr_train, thresholds = roc_curve(Y_Train, nb.predict_proba(TFIDF_TR_CSR) [:,1])
    AUC_TR.append(auc(a_fpr_train, a_tpr_train))

    a_fpr_cv, a_tpr_cv, thresholds = roc_curve(Y_CV, nb.predict_proba(TFIDF_CV_CSR) [:,1])
    AUC_CV.append(auc(a_fpr_cv, a_tpr_cv))


for av in tqdm(alpha):
    b = np.log10(av)
    L_ALPHA.append(b)

# Performance of model on Train data and Test data for each hyper parameter.
plt.plot(L_ALPHA, AUC_TR, label='AUC_Train',color='red',linewidth=2)
plt.scatter(L_ALPHA, AUC_TR, label='AUC_Train',color='red',linewidth=2)
plt.gca()
plt.plot(L_ALPHA, AUC_CV, label='AUC_CV',color='blue',linewidth=2)
plt.scatter(L_ALPHA, AUC_CV, label='AUC_CV',color='blue',linewidth=2)
plt.gca()
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 9/9 [00:05<00:00,  1.59it/s]
```

```
100%|██████████| 9/9 [00:00<?, ?it/s]
```




```
Wall time: 7.12 s
```

In [174]:
```
1 print(AUC_TR)
2 print(AUC_CV)
```

```
[0.8209130343210337, 0.8184105792952243, 0.8144660748859134, 0.8078640480373748, 0.792222775175454, 0.7261600491848925, 0.5965382
854216575, 0.5442793176813133, 0.5375339793096331]
[0.6609192836423523, 0.6643669795100033, 0.6677887752645999, 0.6705764620620442, 0.6718734284394317, 0.6480679522520089, 0.573004
0988684286, 0.5428300498437408, 0.540232275128462]
```

OBSERVATION:

1. Optimal alpha has been choose based on the alpha which maximize the AUC CV score.
2. In this case **α(alpha)=0.1** is having the highest AUC CV score.
3. Hence **α(alpha)=0.1** is choosen as the optimal alpha value.

In [175]:

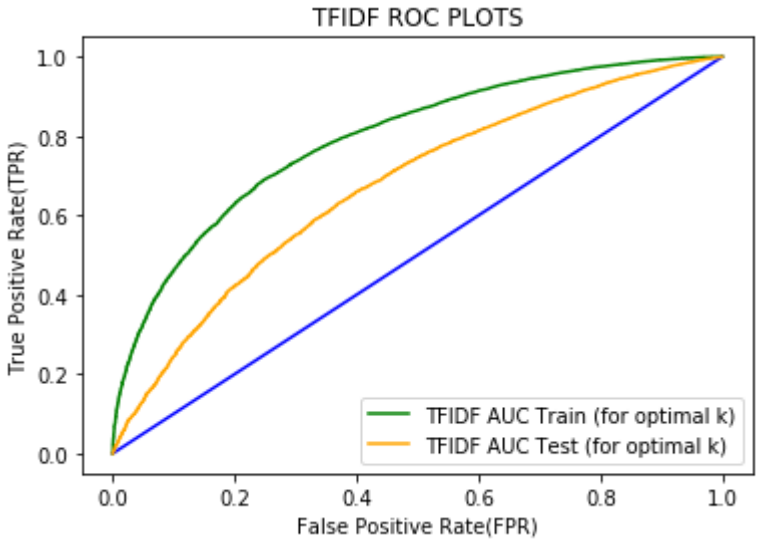
```
A_OPT=0.1
TFIDF_opt = MultinomialNB(alpha=A_OPT,fit_prior=True,class_prior=[0.5,0.5])
TFIDF_opt.fit(TFIDF_Train, Y_Train)
pred = TFIDF_opt.predict(TFIDF_Test)
acc = accuracy_score(Y_Test, pred, normalize=True) * float(100)
print('\nTest accuracy for alpha(a) = {0} is {1}%'.format(A_OPT,acc))

a_fpr_train, a_tpr_train, thresholds = roc_curve(Y_Train, TFIDF_opt.predict_proba(TFIDF_TR_CSR) [:,1])
a_fpr_Test, a_tpr_Test, thresholds = roc_curve(Y_Test, TFIDF_opt.predict_proba(TFIDF_TS_CSR)[:,1])
```

```
Test accuracy for alpha(a) = 0.1 is 67.11139465216908%
```

TFIDF ROC PLOT

In [176]:

```
%%time

#https://stackoverflow.com/questions/52910061/implementing-roc-curves-for-k-nn-machine-learning-algorithm-using-python-and-sci

plt.plot([0,1],[0,1],'k-', color='blue')
plt.plot(a_fpr_train, a_tpr_train, label="TFIDF AUC Train (for optimal k)", color='green')
plt.plot(a_fpr_Test, a_tpr_Test, label="TFIDF AUC Test (for optimal k)", color='orange')
plt.legend()
plt.ylabel("True Positive Rate(TPR)")
plt.xlabel("False Positive Rate(FPR)")
plt.title("TFIDF ROC PLOTS")
plt.show()
print("-"*120)
print("AUC Train (for optimal k) =", auc(a_fpr_train, a_tpr_train))
print("AUC Test (for optimal k) =", auc(a_fpr_Test, a_tpr_Test))
TFIDF_AOPT=A_OPT
TFIDF_AUC=round(auc(a_fpr_Test, a_tpr_Test)*100)
pred3 = TFIDF_opt.predict(TFIDF_Train)
pred4 = TFIDF_opt.predict(TFIDF_Test)

```



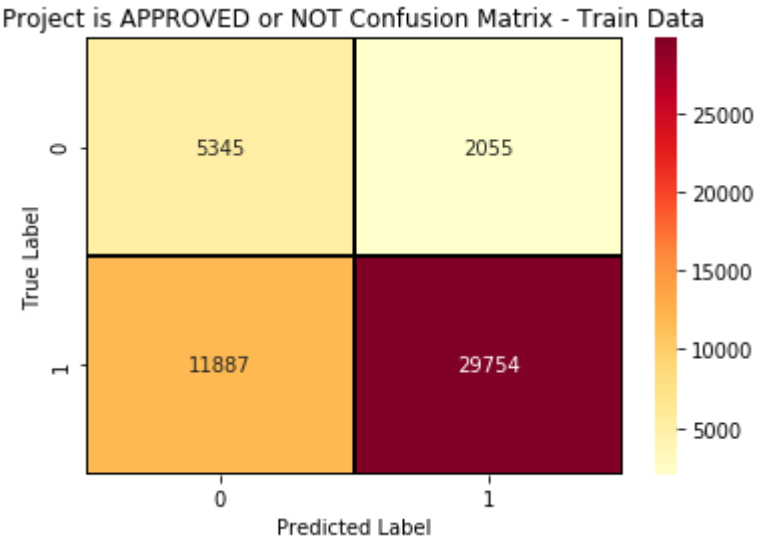

```
------------------------------------------------------------------------------------------------------------------------
AUC Train (for optimal k) = 0.792222775175454
AUC Test (for optimal k) = 0.6727470054018932
Wall time: 1.28 s
```

TFIDF CONFUSION MATRIX

In [177]:
```
%%time
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
#https://getaravind.com/blog/confusion-matrix-seaborn-heatmap/
%matplotlib inline
from sklearn.metrics import confusion_matrix
Train = confusion_matrix(Y_Train, pred3)
sns.heatmap(Train,annot=True,cbar=True,fmt='d',cmap='YlOrRd',linewidths=1,linecolor='black')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Project is APPROVED or NOT Confusion Matrix - Train Data')
```

```
Wall time: 388 ms
```

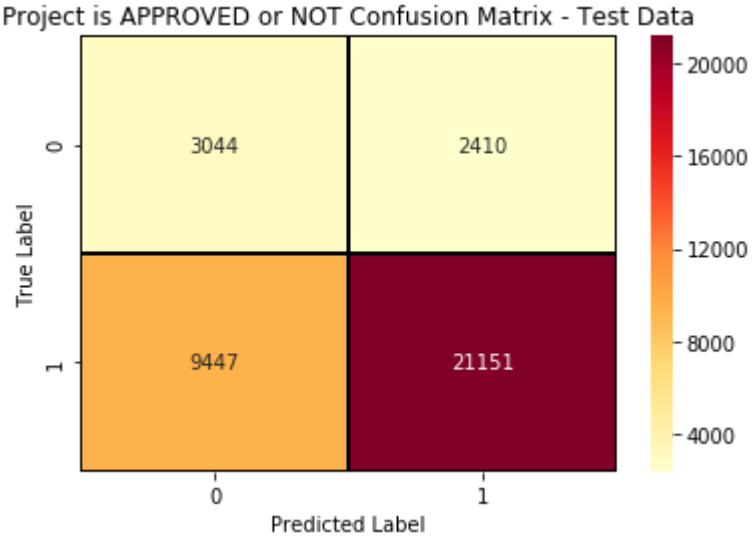
Out[177]:
```
Text(0.5, 1, 'Project is APPROVED or NOT Confusion Matrix - Train Data')
```



In [178]:
```
%%time
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
#https://getaravind.com/blog/confusion-matrix-seaborn-heatmap/
Test = confusion_matrix(Y_Test, pred4)
sns.heatmap(Test,annot=True,cbar=True,fmt='d',cmap='YlOrRd',linewidths=1,linecolor='black')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Project is APPROVED or NOT Confusion Matrix - Test Data')
```

```
Wall time: 156 ms
```

Out[178]:
```
Text(0.5, 1, 'Project is APPROVED or NOT Confusion Matrix - Test Data')
```



2.4.2.1 Top 10 important features of positive class from SET 2

In [179]:
```
T_NB = MultinomialNB(alpha = A_OPT, class_prior=[0.5,0.5])
T_NB.fit(TFIDF_TR_CSR, Y_Train)

TFIDF_F = []

for TFIDF in vectorizer_cat.get_feature_names() :
    TFIDF_F.append(TFIDF)
for TFIDF1 in vectorizer_sub_cat.get_feature_names() :
    TFIDF_F.append(TFIDF1)
for TFIDF2 in vectorizer_school.get_feature_names() :
    TFIDF_F.append(TFIDF2)
for TFIDF3 in vectorizer_grade.get_feature_names() :
    TFIDF_F.append(TFIDF3)
for TFIDF4 in vectorizer_prefix.get_feature_names() :
    TFIDF_F.append(TFIDF4)
for TFIDF5 in vectorizer_essays_tfidf.get_feature_names() :
    TFIDF_F.append(TFIDF5)
for TFIDF6 in vectorizer_titles_tfidf.get_feature_names() :
    TFIDF_F.append(TFIDF6)
TFIDF_F.append("price")
TFIDF_F.append("quantity")
TFIDF_F.append("prev_proposed_projects")
TFIDF_F.append("title_word_count")
TFIDF_F.append("essay_word_count")
len(TFIDF_F)
```

Out[179]: 14285

In [180]:
```
POS_F_TFIDF = T_NB.feature_log_prob_[1, :].argsort()[::-1]
for i in POS_F_TFIDF[:10]:
    print(TFIDF_F[i])
```

```
essay_word_count
title_word_count
quantity
prev_proposed_projects
mrs
Literacy_Language
grades_prek_2
Math_Science
ms
grades_3_5
```

2.4.2.2 Top 10 important features of negative class from SET 2

In [181]:
```
NEG_F_TFIDF = T_NB.feature_log_prob_[0, :].argsort()[::-1]
for i in NEG_F_TFIDF[:10]:
    print(TFIDF_F[i])
```

```
essay_word_count
quantity
title_word_count
prev_proposed_projects
mrs
Literacy_Language
Math_Science
grades_prek_2
ms
grades_3_5
```

3. Conclusions

In [182]:
```
# Please compare all your models using Prettytable library

pt = PrettyTable()
pt.field_names= ("Vectorizer", "Model", "HyperParameter", "AUC")
pt.add_row(["BOW", "Naive Bayes",BOW_AOPT, BOW_AUC])
pt.add_row(["Tf-Idf", "Naive Bayes", TFIDF_AOPT, TFIDF_AUC])
print(pt)
```

```
+------------+-------------+----------------+------+
| Vectorizer |    Model    | HyperParameter | AUC  |
+------------+-------------+----------------+------+
|    BOW     | Naive Bayes |      0.1       | 70.0 |
|   Tf-Idf   | Naive Bayes |      0.1       | 67.0 |
+------------+-------------+----------------+------+
```

SUMMARY:

1. Compare to KNN "Naive Bayes" is giving better AUC value and good interpretable.
2. Also latency time is very low when compare to KNN.
3. Feature importance can be find very easily from the given data.
4. Alpha value needs to be choose correctly else there is a chance for high variance and low bias model(i.e.,Overfitting).