# Donors choose data analysis

## Table of contents

## Little History about Data Set

Founded in 2000 by a high school teacher in the Bronx, DonorsChoose.org empowers public school teachers from across the country to request much-needed materials and experiences for their students. At any given time, there are thousands of classroom requests that can be brought to life with a gift of any amount.

## Answers to What and Why Questions on Data Set

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 p<br>**Example:** `p036502` |

| Feature | Description |
| --- | --- |
| `project_title` | Title of the project. **Examples:**<br>• `Art Will Make You Happy`<br>• `First Grade Fun` |
| `project_grade_category` | Grade level of students for which the project is targeted. One of the follow enumerated values:<br>• `Grades PreK-2`<br>• `Grades 3-5`<br>• `Grades 6-8`<br>• `Grades 9-12` |
| `project_subject_categories` | One or more (comma-separated) su categories for the project from the fo enumerated list of values:<br>• `Applied Learning`<br>• `Care & Hunger`<br>• `Health & Sports`<br>• `History & Civics`<br>• `Literacy & Language`<br>• `Math & Science`<br>• `Music & The Arts`<br>• `Special Needs`<br>• `Warmth`<br>**Examples:**<br>• `Music & The Arts`<br>• `Literacy & Language, Ma & Science` |

| Feature | Description |
| --- | --- |
| **`school_state`** | State where school is located (Two-l U.S. postal code). **Example:** `WY` |
| **`project_subject_subcategories`** | One or more (comma-separated) su subcategories for the project. **Exam**<br>• `Literacy`<br>• `Literature & Writing, Social Sciences` |
| **`project_resource_summary`** | An explanation of the resources nee the project. **Example:**<br>• `My students need hands literacy materials to manage sensory needs!` |
| **`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 w submitted. **Example:** `2016-04-28 12:43:56.245` |
| **`teacher_id`** | A unique identifier for the teacher of proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c` |

| Feature | Description |
| --- | --- |
| **`teacher_prefix`** | Teacher's title. One of the following enumerated values:<br>• `nan`<br>• `Dr.`<br>• `Mr.`<br>• `Mrs.`<br>• `Ms.`<br>• `Teacher.` |
| **`teacher_number_of_previously_posted_projects`** | Number of project applications previ submitted by the same teacher. **Exa** 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_4:__ "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.

## Importing required libraries

In [83]:
```
# numpy for easy numerical computations
import numpy as np
# pandas for dataframes and filterings
import pandas as pd
```

```
# sqlite3 library for performing operations on sqlite file
import sqlite3
# matplotlib for plotting graphs
import matplotlib.pyplot as plt
# seaborn library for easy plotting
import seaborn as sbrn
# warnings library for specific settings
import warnings
# regularlanguage for regex operations
import re
# For loading precomputed models
import pickle
# For loading natural language processing tool-kit
import nltk

# For loading files from google drive
from google.colab import drive
# For working with files in google drive
drive.mount('/content/drive')
# tqdm for tracking progress of loops
from tqdm import tqdm_notebook as tqdm
# For creating dictionary of words
from collections import Counter
# For creating BagOfWords Model
from sklearn.feature_extraction.text import CountVectorizer
# For creating TfidfModel
from sklearn.feature_extraction.text import TfidfVectorizer
# For standardizing values
from sklearn.preprocessing import StandardScaler
# For merging sparse matrices along row direction
from scipy.sparse import hstack
# For merging sparse matrices along column direction
from scipy.sparse import vstack
# For calculating TSNE values
from sklearn.manifold import TSNE
# For calculating the accuracy score on cross validate data
from sklearn.metrics import accuracy_score
# For performing the k-fold cross validation
from sklearn.model_selection import cross_val_score
```

```
# For splitting the data set into test and train data
from sklearn import model_selection
# Logistic Regression classifier for classification
from sklearn.linear_model import LogisticRegression
# For creating samples for making dataset balanced
from sklearn.utils import resample
# For shuffling the dataframes
from sklearn.utils import shuffle
# For calculating roc_curve parameters
from sklearn.metrics import roc_curve
# For calculating auc value
from sklearn.metrics import auc
# For displaying results in table format
from prettytable import PrettyTable
# For generating confusion matrix
from sklearn.metrics import confusion_matrix
# For using gridsearch cv to find best parameter
from sklearn.model_selection import GridSearchCV
# For performing min-max standardization to features
from sklearn.preprocessing import MinMaxScaler
# For calculating sentiment score of the text
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')

warnings.filterwarnings('ignore')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remoun
t, call drive.mount("/content/drive", force_remount=True).
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

## Reading and Storing Data

In [0]:
```
projectsData = pd.read_csv('drive/My Drive/train_data.csv');
resourcesData = pd.read_csv('drive/My Drive/resources.csv');
```

In [85]:
```
projectsData.head(3)
```

Out[85]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |

In [86]:
```
projectsData.tail(3)
```

Out[86]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_ |
|---|---|---|---|---|---|
| **109245** | 143653 | p155633 | cdbfd04aa041dc6739e9e576b1fb1478 | Mrs. | NJ |

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_ |
|---|---|---|---|---|---|
| **109246** | 164599 | p206114 | 6d5675dbfafa1371f0e2f6f1b716fe2d | Mrs. | NY |
| **109247** | 128381 | p191189 | ca25d5573f2bd2660f7850a886395927 | Ms. | VA |

In [87]:
```
resourcesData.head(3)
```

Out[87]:

| | id | description | quantity | price |
|---|---|---|---|---|
| **0** | p233245 | LC652 - Lakeshore Double-Space Mobile Drying Rack | 1 | 149.00 |
| **1** | p069063 | Bouncy Bands for Desks (Blue support pipes) | 3 | 14.95 |
| **2** | p069063 | Cory Stories: A Kid's Book About Living With Adhd | 1 | 8.45 |

In [88]:
```
resourcesData.tail(3)
```

Out[88]:

| | id | description | quantity | price |
|---|---|---|---|---|
| **1541269** | p031981 | Black Electrical Tape (GIANT 3 PACK) Each Roll... | 6 | 8.99 |
| **1541270** | p031981 | Flormoon DC Motor Mini Electric Motor 0.5-3V 1... | 2 | 8.14 |
| **1541271** | p031981 | WAYLLSHINE 6PCS 2 x 1.5V AAA Battery Spring Cl... | 2 | 7.39 |

## Helper functions and classes

In [0]:
```
def equalsBorder(numberOfEqualSigns):
    """
    This function prints passed number of equal signs
    """
    print("="* numberOfEqualSigns);
```

In [0]:
```
# Citation link: https://stackoverflow.com/questions/8924173/how-do-i-print-bold-text-in-python
class color:
   PURPLE = '\033[95m'
   CYAN = '\033[96m'
   DARKCYAN = '\033[36m'
   BLUE = '\033[94m'
   GREEN = '\033[92m'
   YELLOW = '\033[93m'
   RED = '\033[91m'
   BOLD = '\033[1m'
   UNDERLINE = '\033[4m'
   END = '\033[0m'
```

In [0]:
```
def printStyle(text, style):
    "This function prints text with the style passed to it"
    print(style + text + color.END);
```

## Shapes of projects data and resources data

In [92]:
```
printStyle("Number of data points in projects data: {}".format(projectsData.shape[0]), color.BOLD);
printStyle("Number of attributes in projects data:{}".format(projectsData.shape[1]), color.BOLD);
equalsBorder(60);
printStyle("Number of data points in resources data: {}".format(resourcesData.shape[0]), color.BOLD);
```

```
printStyle("Number of attributes in resources data: {}".format(resource
sData.shape[1]), color.BOLD);
```

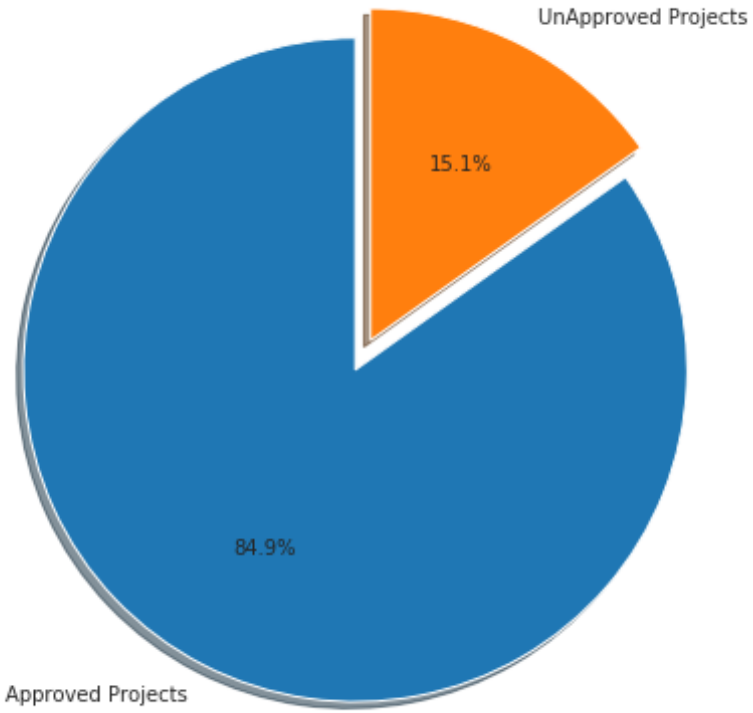
```
Number of data points in projects data: 109248
Number of attributes in projects data:17
============================================================
Number of data points in resources data: 1541272
Number of attributes in resources data: 4
```

# Univariate data analysis

In [93]:

```
approvedProjects = projectsData[projectsData.project_is_approved == 1].
shape[0];
unApprovedProjects = projectsData[projectsData.project_is_approved == 0
].shape[0];
totalProjects = projectsData.shape[0];
print("Number of projects approved for funding: {}, ({})".format(approv
edProjects, (approvedProjects / totalProjects) * 100));
print("Number of projects not approved for funding: {}, ({})".format(un
ApprovedProjects, (unApprovedProjects / totalProjects) * 100));
# Pie chart representation
# Citation: https://matplotlib.org/gallery/pie_and_polar_charts/pie_fea
tures.html
labels = ["Approved Projects", "UnApproved Projects"];
explode = (0, 0.1);
sizes = [approvedProjects, unApprovedProjects];
figure, ax = plt.subplots();
ax.pie(sizes, labels = labels, explode = explode, autopct = '%1.1f%%',
shadow = True, startangle = 90);
ax.axis('equal');
plt.rcParams['figure.figsize'] = (7, 7);
plt.show();
```

```
Number of projects approved for funding: 92706, (84.85830404217927)
Number of projects not approved for funding: 16542, (15.14169595782073
9)
```




### Observation:

1. There are more number of approved projects compared to rejected projects. So this is a imbalanced dataset.

## Univariate Analysis : 'school_state'

### Project proposal percentage in different states

In [94]:
```
groupedByStatesData = pd.DataFrame(projectsData.groupby(['school_state'
```

```
])['project_is_approved'].apply(np.mean)).reset_index();
groupedByStatesData.columns = ['state_code', 'number_of_proposals'];
groupedByStatesData = groupedByStatesData.sort_values(by=['number_of_pr
oposals'], ascending = True);
printStyle("5 States with lowest percentage of project approvals:", col
or.BOLD);
equalsBorder(60);
groupedByStatesData.head(5)
```

**5 States with lowest percentage of project approvals:**
**============================================================**

Out[94]:

| | state_code | number_of_proposals |
|---|---|---|
| **46** | VT | 0.800000 |
| **7** | DC | 0.802326 |
| **43** | TX | 0.813142 |
| **26** | MT | 0.816327 |
| **18** | LA | 0.831245 |

In [95]:
```
printStyle("5 states with highest percentage of project approvals: ", c
olor.BOLD);
equalsBorder(60);
groupedByStatesData.tail(5).iloc[::-1]
```

**5 states with highest percentage of project approvals:**
**============================================================**

Out[95]:

| | state_code | number_of_proposals |
|---|---|---|
| **8** | DE | 0.897959 |
| **28** | ND | 0.888112 |
| **47** | WA | 0.876178 |

| | state_code | number_of_proposals |
|---|---|---|
| **35** | OH | 0.875152 |
| **30** | NH | 0.873563 |

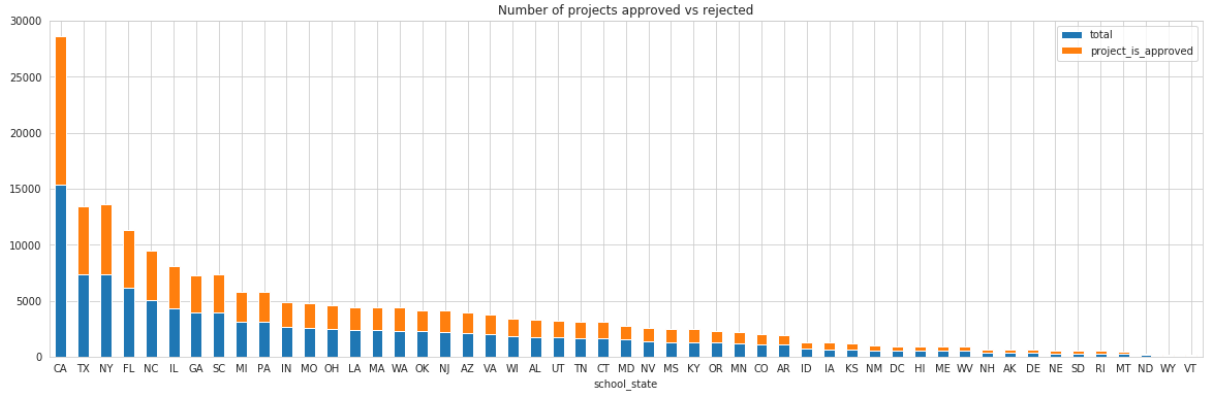
In [0]:
```
def univariateBarPlots(data, col1, col2 = 'project_is_approved', orient
ation = 'vertical', plot = True):
    groupedData = data.groupby(col1);
    # Count number of zeros in dataframe python: https://stackoverflow.
com/a/51540521/4084039
    tempData = pd.DataFrame(groupedData[col2].agg(lambda x: x.eq(1).sum
())).reset_index();
    tempData['total'] = pd.DataFrame(groupedData[col2].agg({'total': 'c
ount'})).reset_index()['total'];
    tempData['approval_rate'] = pd.DataFrame(groupedData[col2].agg({'ap
proval_rate': 'mean'})).reset_index()['approval_rate'];
    tempData.sort_values(by=['total'], inplace = True, ascending = Fals
e);
    tempDataWithTotalAndCol2 = tempData[['total', col2, col1]]
    if plot:
        if(orientation == 'vertical'):
            tempDataWithTotalAndCol2.plot(x = col1, align= 'center', ki
nd = 'bar', title = "Number of projects approved vs rejected", figsize
= (20, 6), stacked = True, rot = 0);
        else:
            tempDataWithTotalAndCol2.plot(x = col1, align= 'center', ki
nd = 'barh', title = "Number of projects approved vs rejected", width =
 0.8, figsize = (23, 20), stacked = True);
    return tempData;
```

In [97]:
```
statesCharacteristicsData = univariateBarPlots(projectsData, 'school_st
ate', 'project_is_approved', orientation = 'vertical');
printStyle("Top 5 states with high project proposals", color.BOLD)
equalsBorder(60);
statesCharacteristicsData.head(5)
```

**Top 5 states with high project proposals**
==========

Out[97]:

| | school_state | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **4** | CA | 13205 | 15388 | 0.858136 |
| **43** | TX | 6014 | 7396 | 0.813142 |
| **34** | NY | 6291 | 7318 | 0.859661 |
| **9** | FL | 5144 | 6185 | 0.831690 |
| **27** | NC | 4353 | 5091 | 0.855038 |




In [98]:
```
printStyle("Top 5 states with least project proposals", color.BOLD)
equalsBorder(60);
statesCharacteristicsData.tail(5)
```

**Top 5 states with least project proposals**

==========

Out[98]:

| | school_state | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **39** | RI | 243 | 285 | 0.852632 |
| **26** | MT | 200 | 245 | 0.816327 |
| **28** | ND | 127 | 143 | 0.888112 |

| | school_state | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **50** | WY | 82 | 98 | 0.836735 |
| **46** | VT | 64 | 80 | 0.800000 |

### Observation:

1. Highest number of project proposals are from CA(California) and it was almost about 16000 projects
2. Every state has more than 80% approval rate.

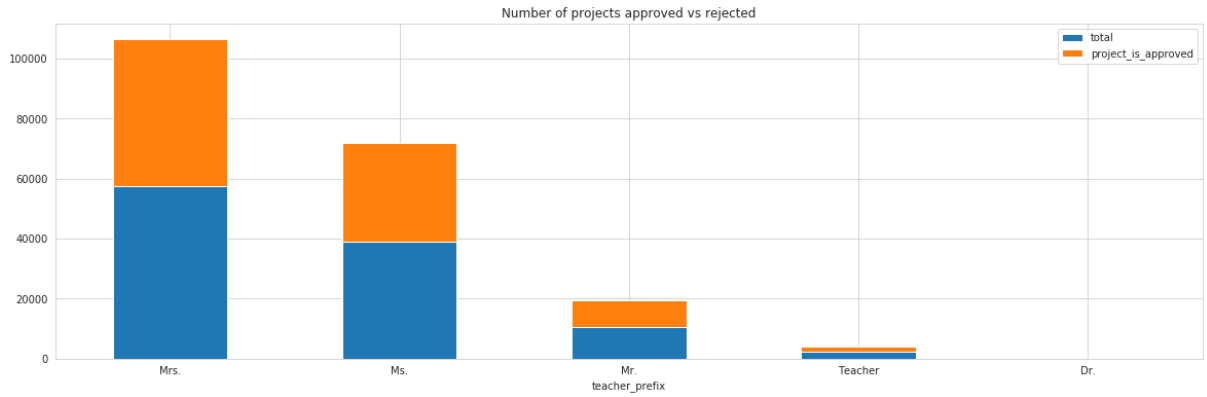
### Univariate Analysis: teacher_prefix

In [99]:
```
teacherPrefixCharacteristicsData = univariateBarPlots(projectsData, 'teacher_prefix', 'project_is_approved', orientation = 'vertical', plot = True);
printStyle("Project proposals characteristics based on types of persons", color.BOLD);
equalsBorder(60);
teacherPrefixCharacteristicsData
```

**Project proposals characteristics based on types of persons**
==========

Out[99]:

| | teacher_prefix | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **2** | Mrs. | 48997 | 57269 | 0.855559 |
| **3** | Ms. | 32860 | 38955 | 0.843537 |
| **1** | Mr. | 8960 | 10648 | 0.841473 |
| **4** | Teacher | 1877 | 2360 | 0.795339 |
| **0** | Dr. | 9 | 13 | 0.692308 |




## Observataion:

1. When compared to others Dr.'s have proposed very less number of projects.
2. Women have proposed more number of projects than men.

## Univariate Analysis: project_grade

In [100]:
```
gradeCharacteristicsData = univariateBarPlots(projectsData, 'project_grade_category', 'project_is_approved', orientation = 'vertical', plot = True);
printStyle("Project proposal characteristics based on grades", color.BOLD);
equalsBorder(60);
gradeCharacteristicsData
```

**Project proposal characteristics based on grades**

==========

Out[100]:

| | project_grade_category | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **3** | Grades PreK-2 | 37536 | 44225 | 0.848751 |
| **0** | Grades 3-5 | 31729 | 37137 | 0.854377 |

| | project_grade_category | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **1** | Grades 6-8 | 14258 | 16923 | 0.842522 |
| **2** | Grades 9-12 | 9183 | 10963 | 0.837636 |


## Observation:

1. Most number of projects proposed are for students less than grade-5 (for primary school students) which means that children are being taught with project oriented teaching which is great.

## Univariate Analysis: project_subject_categories

In [0]:
```
# 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
def cleanCategories(subjectCategories):
```

```
    cleanedCategories = []
    for subjectCategory in tqdm(subjectCategories):
        tempCategory = ""
        for category in subjectCategory.split(","):
            if 'The' in category.split(): # this will split each of the
 catogory based on space "Math & Science"=> "Math","&", "Science"
                category = category.replace('The','') # if we have the
 words "The" we are going to replace it with ''(i.e removing 'The')
            category = category.replace(' ','') # we are placeing all t
he ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Science"
            tempCategory += category.strip()+" "# abc ".strip() will r
eturn "abc", remove the trailing spaces
            tempCategory = tempCategory.replace('&','_')
        cleanedCategories.append(tempCategory)
    return cleanedCategories
```

In [102]:

```
# projectDataWithCleanedCategories = pd.DataFrame(projectsData);
subjectCategories = list(projectsData.project_subject_categories);
cleanedCategories = cleanCategories(subjectCategories);
printStyle("Sample categories: ", color.BOLD);
equalsBorder(60);
print(subjectCategories[0:5]);
equalsBorder(60);
printStyle("Sample cleaned categories: ", color.BOLD);
equalsBorder(60);
print(cleanedCategories[0:5]);
projectsData['cleaned_categories'] = cleanedCategories;
projectsData.head(5)
```

```
Sample categories:
============================================================
['Literacy & Language', 'History & Civics, Health & Sports', 'Health &
 Sports', 'Literacy & Language, Math & Science', 'Math & Science']
============================================================
Sample cleaned categories:
============================================================
['Literacy_Language ', 'History_Civics Health_Sports ', 'Health_Sports
 ', 'Literacy_Language Math_Science ', 'Math_Science ']
```

Out[102]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY |
| **4** | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX |

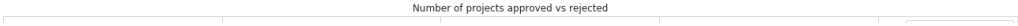
In [103]:

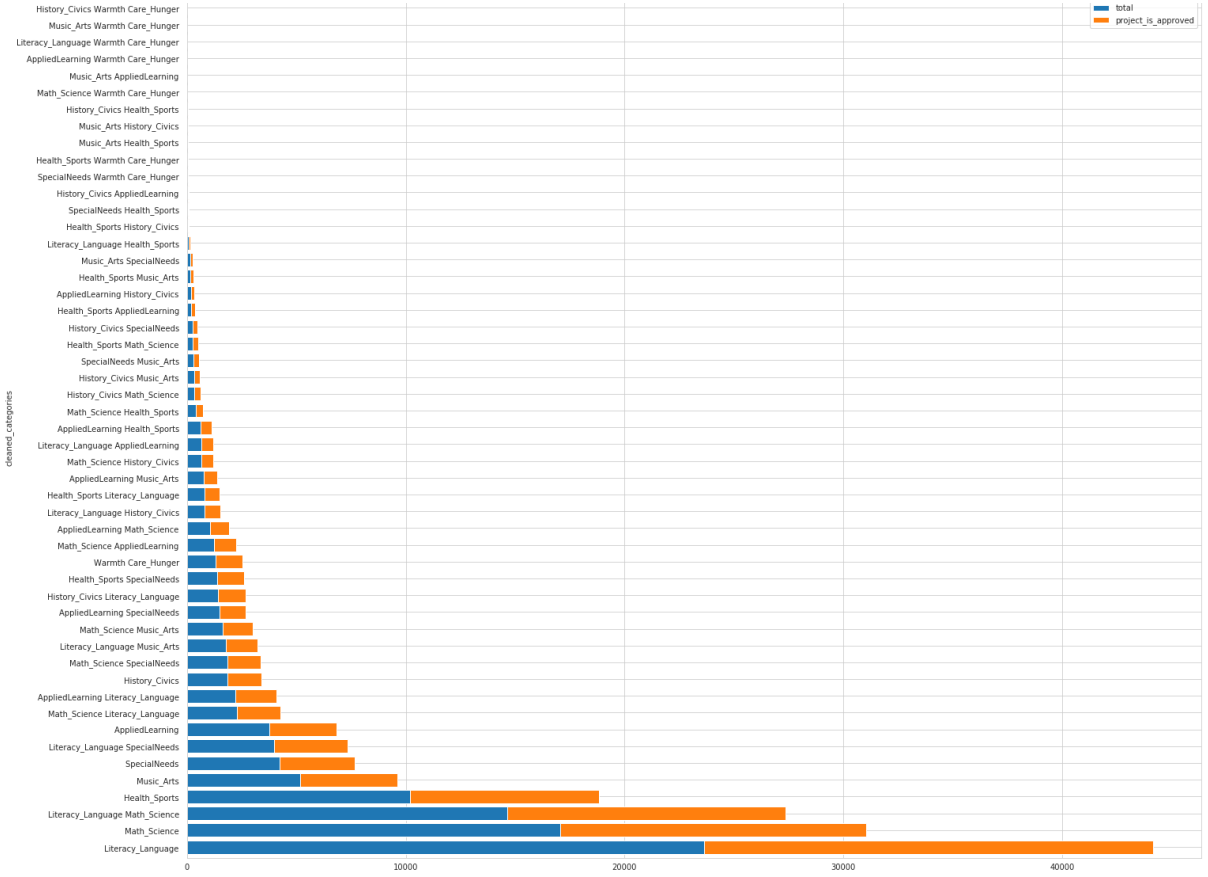
```
categoriesCharacteristicsData = univariateBarPlots(projectsData, 'clean
ed_categories', 'project_is_approved', orientation = 'horizontal', plot
 = True);
print("Project proposals characteristics based on subject categories");
equalsBorder(60);
categoriesCharacteristicsData.head(5)
```

```
Project proposals characteristics based on subject categories
============================================================
```

Out[103]:

| | cleaned_categories | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **24** | Literacy_Language | 20520 | 23655 | 0.867470 |
| **32** | Math_Science | 13991 | 17072 | 0.819529 |
| **28** | Literacy_Language Math_Science | 12725 | 14636 | 0.869432 |
| **8** | Health_Sports | 8640 | 10177 | 0.848973 |
| **40** | Music_Arts | 4429 | 5180 | 0.855019 |






In [104]:
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
categoriesCounter = Counter()
for subjectCategory in projectsData.cleaned_categories.values:
    categoriesCounter.update(subjectCategory.split());
categoriesCounter
```

Out[104]:
```
Counter({'AppliedLearning': 12135,
         'Care_Hunger': 1388,
         'Health_Sports': 14223,
         'History_Civics': 5914,
         'Literacy_Language': 52239,
         'Math_Science': 41421,
```

```
                'Music_Arts': 10293,
                'SpecialNeeds': 13642,
                'Warmth': 1388})
```

In [105]:
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
categoriesDictionary = dict(categoriesCounter);
sortedCategoriesDictionary = dict(sorted(categoriesDictionary.items(), key = lambda keyValue: keyValue[1]));
sortedCategoriesData = pd.DataFrame.from_dict(sortedCategoriesDictionary, orient='index');
sortedCategoriesData.columns = ['subject_categories'];
printStyle("Number of projects by Subject Categories: ", color.BOLD);
equalsBorder(60);
sortedCategoriesData
```

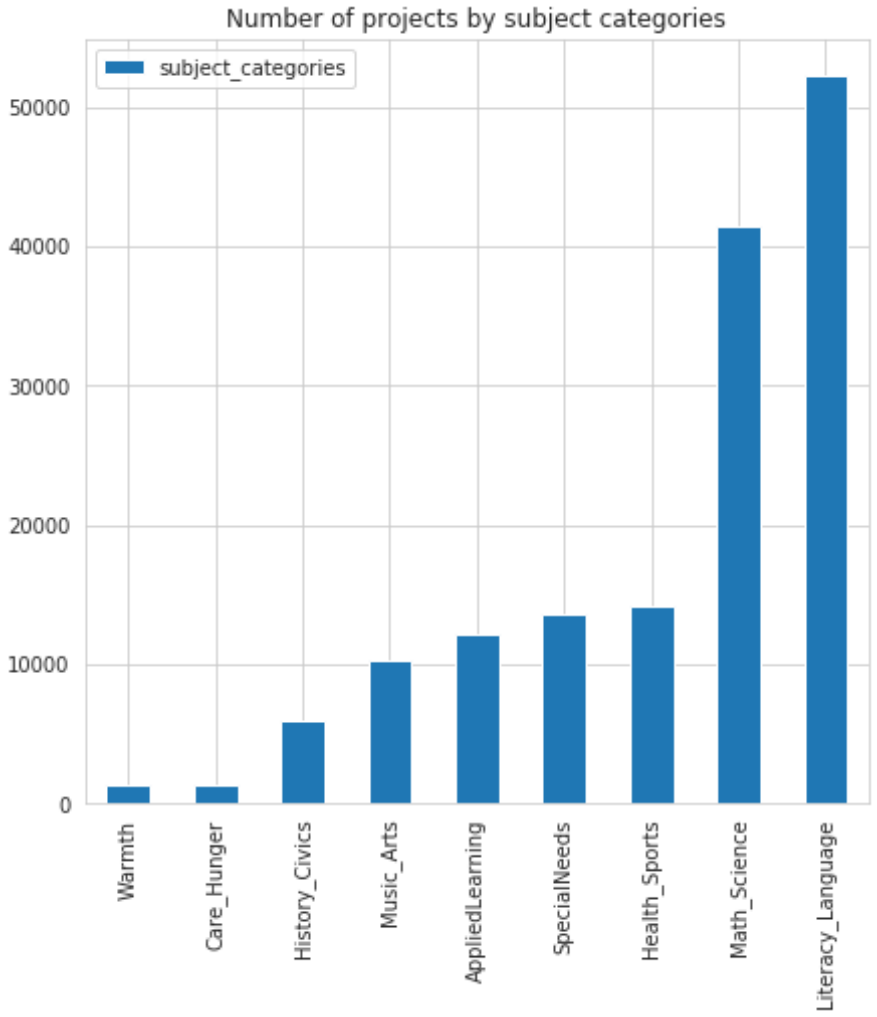
**Number of projects by Subject Categories:**
==========

Out[105]:

| | subject_categories |
|---|---|
| **Warmth** | 1388 |
| **Care_Hunger** | 1388 |
| **History_Civics** | 5914 |
| **Music_Arts** | 10293 |
| **AppliedLearning** | 12135 |
| **SpecialNeeds** | 13642 |
| **Health_Sports** | 14223 |
| **Math_Science** | 41421 |
| **Literacy_Language** | 52239 |

In [106]:
```
sortedCategoriesData.plot(kind = 'bar', title = 'Number of projects by
```

```
subject categories');
```




## Observation:

1. Many number of projects proposed belong to multiple subject categories.
2. When compared to others literacy_language & math_science have large number of project proposals.

## Univariate Analysis: project_subject_subcategories

In [107]:
```
subjectSubCategories = projectsData.project_subject_subcategories;
cleanedSubCategories = cleanCategories(subjectSubCategories);
printStyle("Sample subject sub categories: ", color.BOLD);
equalsBorder(70);
print(subjectSubCategories[0:5]);
equalsBorder(70);
printStyle("Sample cleaned subject sub categories: ", color.BOLD);
equalsBorder(70);
print(cleanedSubCategories[0:5]);
projectsData['cleaned_sub_categories'] = cleanedSubCategories;
```

```
Sample subject sub categories:
======================================================================
0                       ESL, Literacy
1    Civics & Government, Team Sports
2      Health & Wellness, Team Sports
3                Literacy, Mathematics
4                         Mathematics
Name: project_subject_subcategories, dtype: object
======================================================================
Sample cleaned subject sub categories:
======================================================================
['ESL Literacy ', 'Civics_Government TeamSports ', 'Health_Wellness Tea
mSports ', 'Literacy Mathematics ', 'Mathematics ']
```

In [108]:
```
projectsData.head(5)
```

Out[108]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY |
| **4** | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX |

In [109]:
```
subCategoriesCharacteristicsData = univariateBarPlots(projectsData, 'cleaned_sub_categories', 'project_is_approved', plot = False);
print("Project proposals characteristics based on subject sub categories");
equalsBorder(60);
subCategoriesCharacteristicsData.head(5)
```

```
Project proposals characteristics based on subject sub categories
============================================================
```

Out[109]:

| | cleaned_sub_categories | project_is_approved | total | approval_rate |
|---|---|---|---|---|
| **317** | Literacy | 8371 | 9486 | 0.882458 |
| **319** | Literacy Mathematics | 7260 | 8325 | 0.872072 |
| **331** | Literature_Writing Mathematics | 5140 | 5923 | 0.867803 |
| **318** | Literacy Literature_Writing | 4823 | 5571 | 0.865733 |
| **342** | Mathematics | 4385 | 5379 | 0.815207 |

In [110]:
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
subjectsSubCategoriesCounter = Counter();
for subCategory in projectsData.cleaned_sub_categories:
    subjectsSubCategoriesCounter.update(subCategory.split());
subjectsSubCategoriesCounter
```

Out[110]:
```
Counter({'AppliedSciences': 10816,
         'Care_Hunger': 1388,
         'CharacterEducation': 2065,
         'Civics_Government': 815,
         'College_CareerPrep': 2568,
         'CommunityService': 441,
         'ESL': 4367,
         'EarlyDevelopment': 4254,
         'Economics': 269,
         'EnvironmentalScience': 5591,
         'Extracurricular': 810,
         'FinancialLiteracy': 568,
         'ForeignLanguages': 890,
         'Gym_Fitness': 4509,
         'Health_LifeScience': 4235,
         'Health_Wellness': 10234,
         'History_Geography': 3171,
```

```
'Literacy': 33700,
'Literature_Writing': 22179,
'Mathematics': 28074,
'Music': 3145,
'NutritionEducation': 1355,
'Other': 2372,
'ParentInvolvement': 677,
'PerformingArts': 1961,
'SocialSciences': 1920,
'SpecialNeeds': 13642,
'TeamSports': 2192,
'VisualArts': 6278,
'Warmth': 1388})
```

In [111]:
```
# dict sort by value python: https://stackoverflow.com/a/613218/4084039
dictionarySubCategories = dict(subjectsSubCategoriesCounter);
sortedDictionarySubCategories = dict(sorted(dictionarySubCategories.ite
ms(), key = lambda keyValue: keyValue[1]));
sortedSubCategoriesData = pd.DataFrame.from_dict(sortedDictionarySubCat
egories, orient = 'index');
sortedSubCategoriesData.columns = ['subject_sub_categories']
sortedSubCategoriesData.plot(kind = 'bar', title = "Number of projects
 by subject sub categories");
printStyle("Number of projects sorted by subject sub categories: ", col
or.BOLD);
equalsBorder(70);
sortedSubCategoriesData
```

**Number of projects sorted by subject sub categories:**

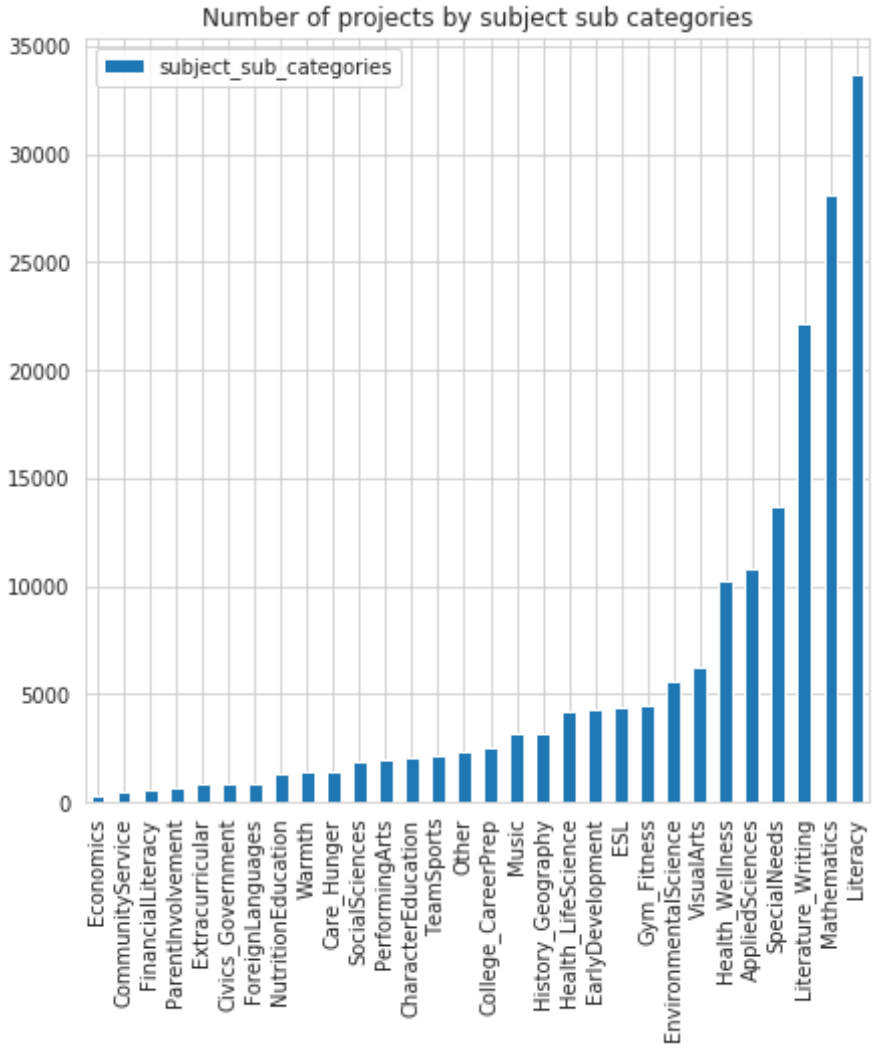
==========

Out[111]:

| | subject_sub_categories |
|---|---|
| **Economics** | 269 |
| **CommunityService** | 441 |
| **FinancialLiteracy** | 568 |

| | subject_sub_categories |
|---|---|
| **ParentInvolvement** | 677 |
| **Extracurricular** | 810 |
| **Civics_Government** | 815 |
| **ForeignLanguages** | 890 |
| **NutritionEducation** | 1355 |
| **Warmth** | 1388 |
| **Care_Hunger** | 1388 |
| **SocialSciences** | 1920 |
| **PerformingArts** | 1961 |
| **CharacterEducation** | 2065 |
| **TeamSports** | 2192 |
| **Other** | 2372 |
| **College_CareerPrep** | 2568 |
| **Music** | 3145 |
| **History_Geography** | 3171 |
| **Health_LifeScience** | 4235 |
| **EarlyDevelopment** | 4254 |
| **ESL** | 4367 |
| **Gym_Fitness** | 4509 |
| **EnvironmentalScience** | 5591 |
| **VisualArts** | 6278 |
| **Health_Wellness** | 10234 |

| | subject_sub_categories |
|---|---|
| **AppliedSciences** | 10816 |
| **SpecialNeeds** | 13642 |
| **Literature_Writing** | 22179 |
| **Mathematics** | 28074 |
| **Literacy** | 33700 |




## Observation:

1. There are more number of subject subcategories than subject categories.
2. Even more number of projects proposed belong to multiple subject sub categories.

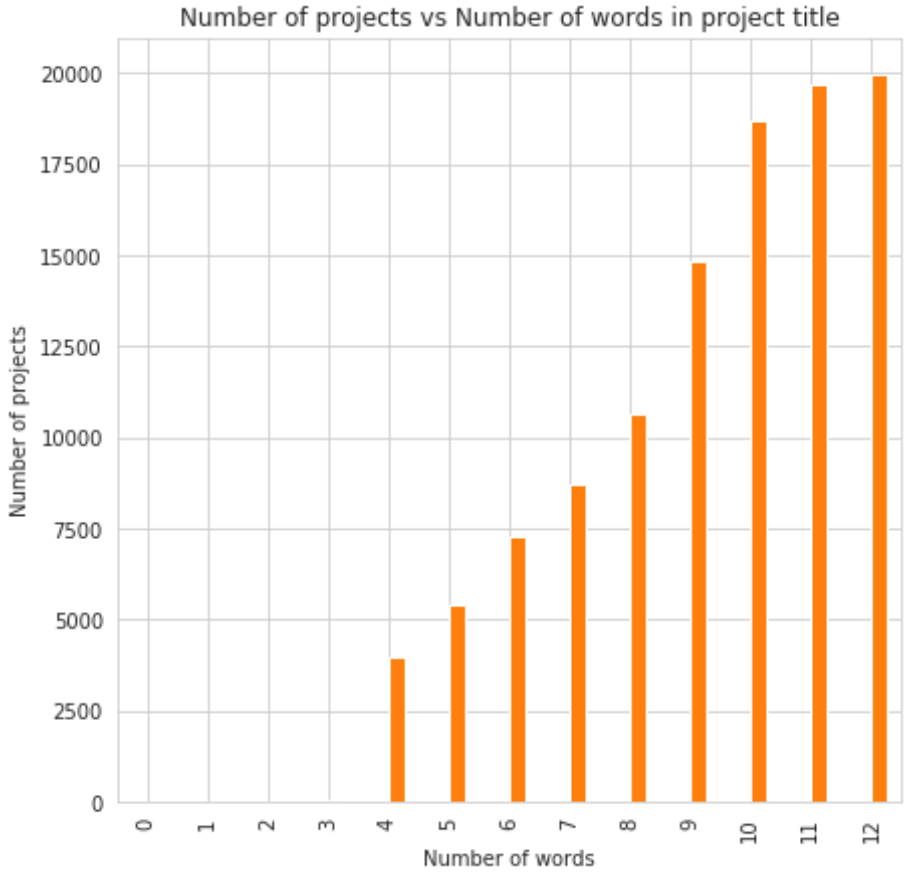
## Univariate Analysis : project_title

In [112]:
```
#How to calculate number of words in a string in DataFrame: https://sta
ckoverflow.com/a/37483537/4084039
wordCounts = projectsData['project_title'].str.split().apply(len).value
_counts();
dictionaryWordCounts = dict(wordCounts);
dictionaryWordCounts = dict(sorted(dictionaryWordCounts.items(), key =
lambda kv: kv[1]));
wordCountsData = pd.DataFrame.from_dict({'number_of_words': list(dictio
naryWordCounts.keys()), 'number_of_projects': list(dictionaryWordCounts
.values())}).sort_values(by = ['number_of_projects']);
wordCountsData.plot(kind = 'bar', title = "Number of projects vs Number
 of words in project title", legend = False);
plt.xlabel('Number of words');
plt.ylabel('Number of projects');
wordCountsData
```

Out[112]:

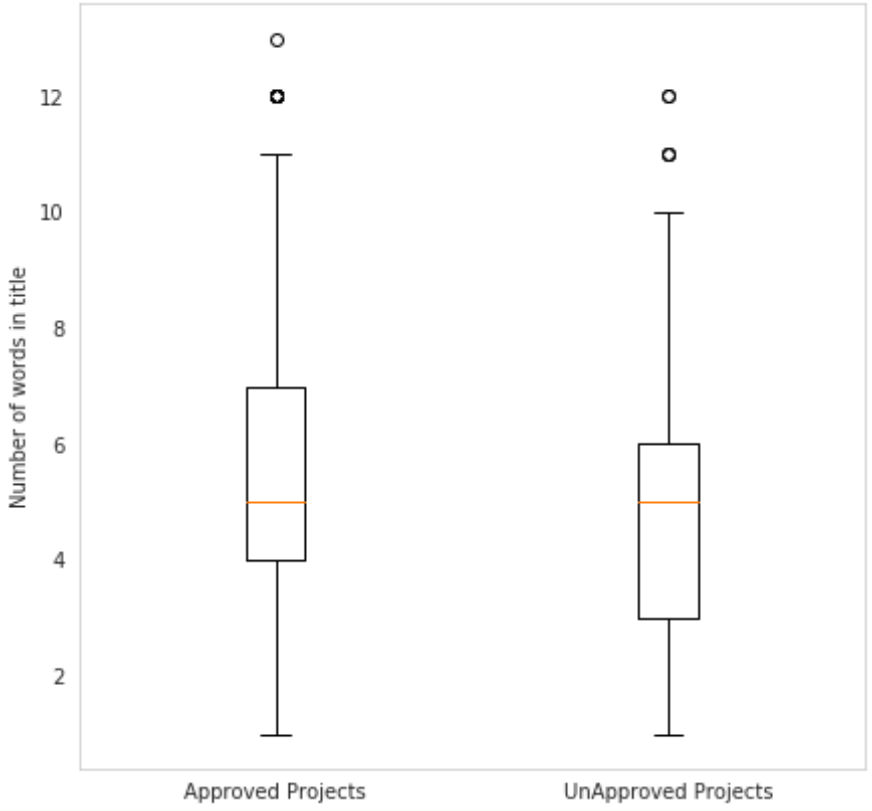
| | number_of_words | number_of_projects |
|---|---|---|
| **0** | 13 | 1 |
| **1** | 12 | 11 |
| **2** | 11 | 30 |
| **3** | 1 | 31 |
| **4** | 10 | 3968 |
| **5** | 9 | 5383 |
| **6** | 8 | 7289 |
| **7** | 2 | 8733 |
| **8** | 7 | 10631 |
| **9** | 6 | 14824 |
| **10** | 3 | 18691 |

| | number_of_words | number_of_projects |
|---|---|---|
| **11** | 5 | 19677 |
| **12** | 4 | 19979 |

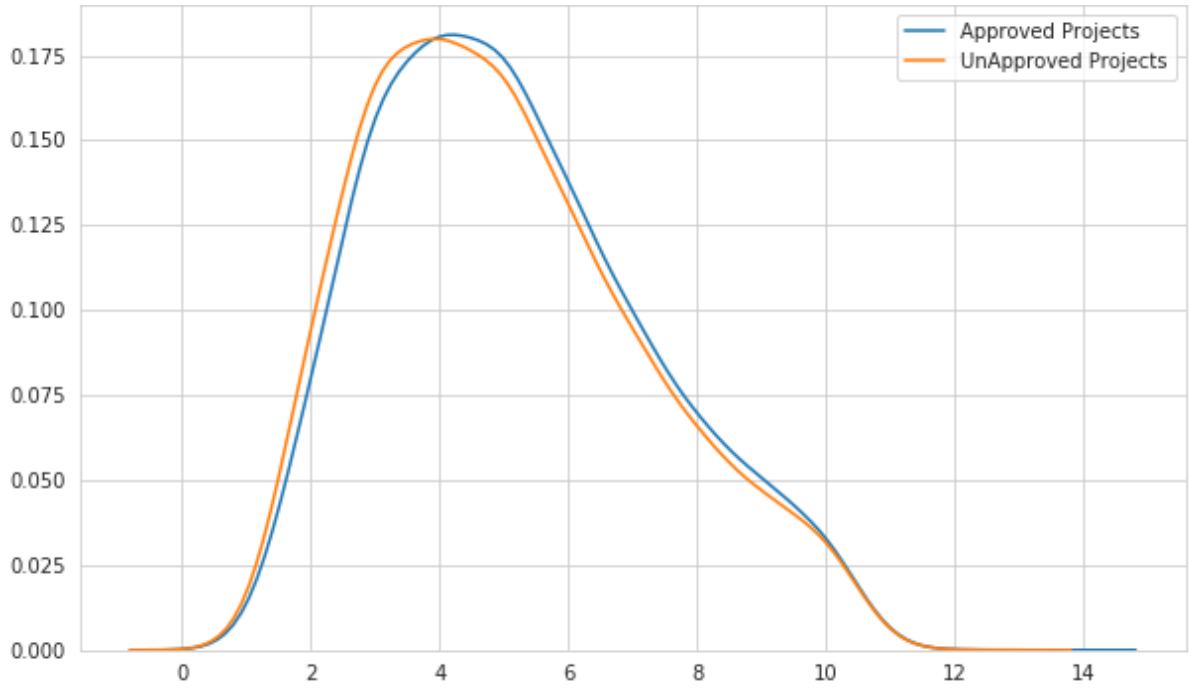



In [113]:
```
approvedNumberOfProjects = projectsData[projectsData.project_is_approve
d == 1]['project_title'].str.split().apply(len);
approvedNumberOfProjects = approvedNumberOfProjects.values
unApprovedNumberOfProjects = projectsData[projectsData.project_is_appro
ved == 0]['project_title'].str.split().apply(len);
unApprovedNumberOfProjects = unApprovedNumberOfProjects.values
```

```
plt.boxplot([approvedNumberOfProjects, unApprovedNumberOfProjects]);
plt.grid();
plt.xticks([1, 2], ['Approved Projects', 'UnApproved Projects']);
plt.ylabel('Number of words in title');
plt.show();
```




In [114]:
```
plt.figure(figsize = (10, 6));
sbrn.kdeplot(approvedNumberOfProjects, label = "Approved Projects", bw
= 0.6);
sbrn.kdeplot(unApprovedNumberOfProjects, label = "UnApproved Projects",
 bw = 0.6);
plt.legend();
plt.show();
```




### Observations:

1. Most of the approved projects have between 4 to 8 number of words in their project_title.
2. Most of the rejected projects have between 3 to 6 number of words in their project_title.

### Univariate Analysis: project_essay_1,2,3,4

In [115]:
```
projectsData['project_essay'] = projectsData['project_essay_1'].map(str
) + projectsData['project_essay_2'].map(str) + \
                                projectsData['project_essay_3'].map(str
) + projectsData['project_essay_4'].map(str);
projectsData.head(5)
```

Out[115]:

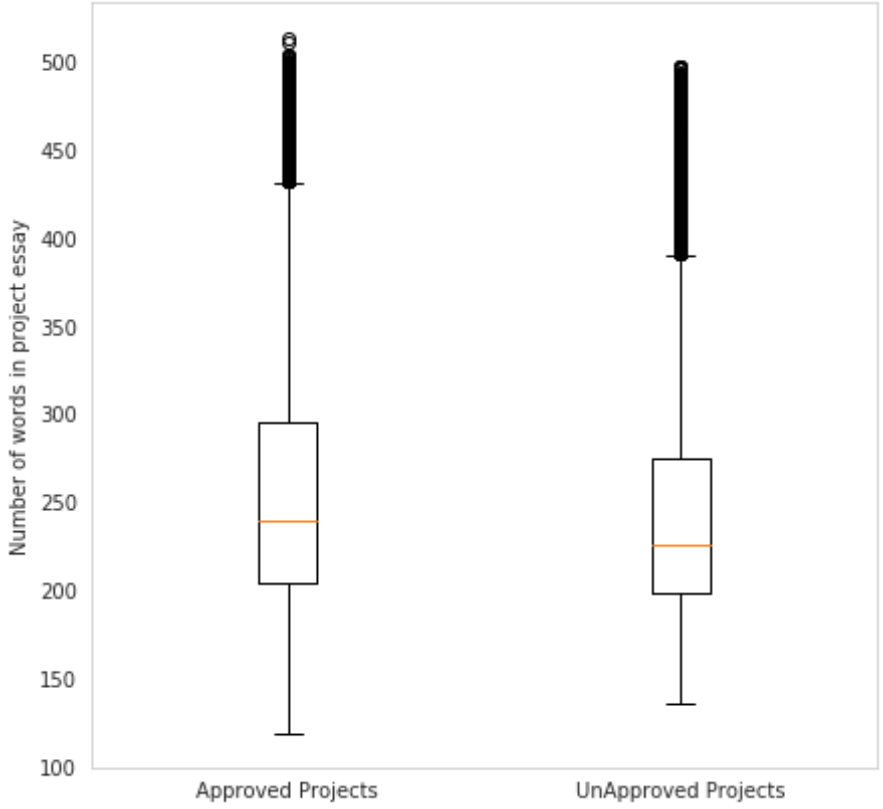
Out[115]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY |
| **4** | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX |

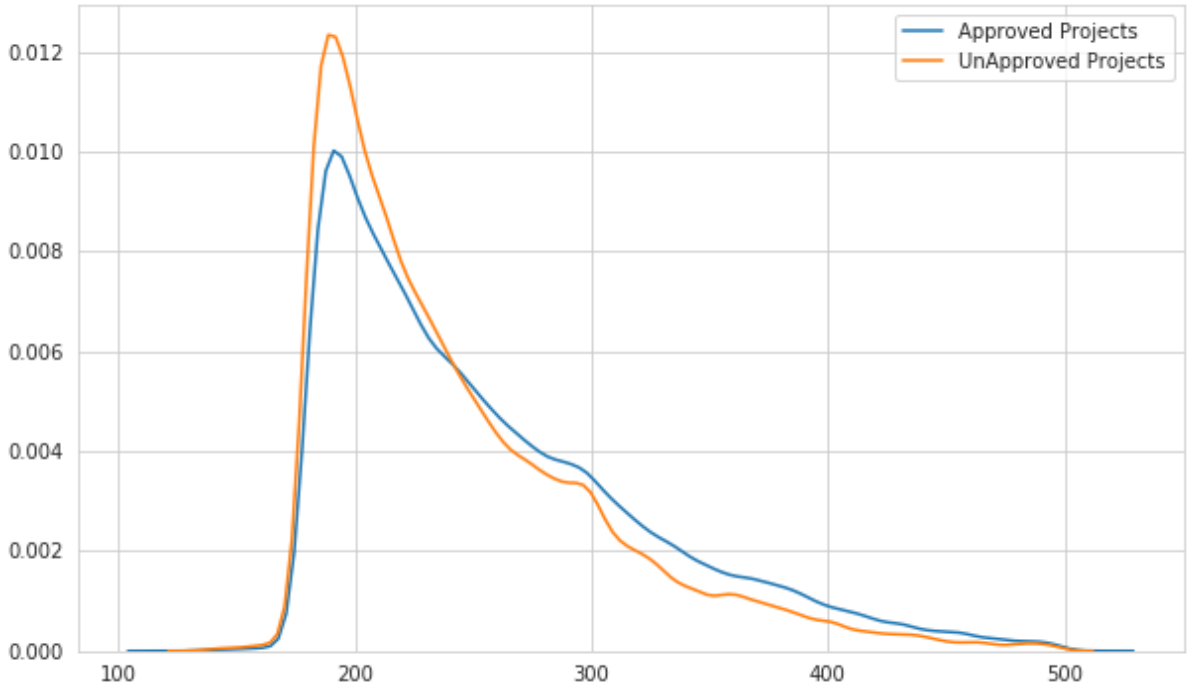
In [116]:

```
approvedNumberOfProjects = projectsData[projectsData.project_is_approve
d == 1]['project_essay'].str.split().apply(len);
approvedNumberOfProjects = approvedNumberOfProjects.values
unApprovedNumberOfProjects = projectsData[projectsData.project_is_appro
ved == 0]['project_essay'].str.split().apply(len);
```

```
unApprovedNumberOfProjects = unApprovedNumberOfProjects.values
plt.boxplot([approvedNumberOfProjects, unApprovedNumberOfProjects]);
plt.grid();
plt.xticks([1, 2], ['Approved Projects', 'UnApproved Projects']);
plt.ylabel('Number of words in project essay');
plt.show();
```




In [117]:
```
plt.figure(figsize = (10, 6));
sbrn.kdeplot(approvedNumberOfProjects, label = "Approved Projects", bw
= 5);
sbrn.kdeplot(unApprovedNumberOfProjects, label = "UnApproved Projects",
 bw = 5);
plt.legend();
plt.show();
```




### Observation:

1. The approved and rejected projects overlap largely when plotted based on number of words in project_essay. So we cannot predict any observation which will be useful for classification.

### Univariate Analysis: price

In [118]:
```
projectsData.head(5)
```

Out[118]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY |
| **4** | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX |

In [119]:
```
resourcesData.head(5)
```

Out[119]:

| | id | description | quantity | price |
|---|---|---|---|---|

| | id | description | quantity | price |
|---|---|---|---|---|
| **0** | p233245 | LC652 - Lakeshore Double-Space Mobile Drying Rack | 1 | 149.00 |
| **1** | p069063 | Bouncy Bands for Desks (Blue support pipes) | 3 | 14.95 |
| **2** | p069063 | Cory Stories: A Kid's Book About Living With Adhd | 1 | 8.45 |
| **3** | p069063 | Dixon Ticonderoga Wood-Cased #2 HB Pencils, Bo... | 2 | 13.59 |
| **4** | p069063 | EDUCATIONAL INSIGHTS FLUORESCENT LIGHT FILTERS... | 3 | 24.95 |

In [120]:
```
# https://stackoverflow.com/questions/22407798/how-to-reset-a-dataframes-indexes-for-all-groups-in-one-step
priceAndQuantityData = resourcesData.groupby('id').agg({'price': 'sum',
 'quantity': 'sum'}).reset_index();
priceAndQuantityData.head(5)
```

Out[120]:

| | id | price | quantity |
|---|---|---|---|
| **0** | p000001 | 459.56 | 7 |
| **1** | p000002 | 515.89 | 21 |
| **2** | p000003 | 298.97 | 4 |
| **3** | p000004 | 1113.69 | 98 |
| **4** | p000005 | 485.99 | 8 |

In [121]:
```
projectsData.shape
```

Out[121]: (109248, 20)

In [122]:
```
projectsData = pd.merge(projectsData, priceAndQuantityData, on = 'id',
how = 'left');
print(projectsData.shape);
projectsData.head(3)
```

```
(109248, 22)
```

Out[122]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |

3 rows × 22 columns

In [123]:
```
projectsData[projectsData['id'] == 'p253737']
```

Out[123]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | p |
|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2 |

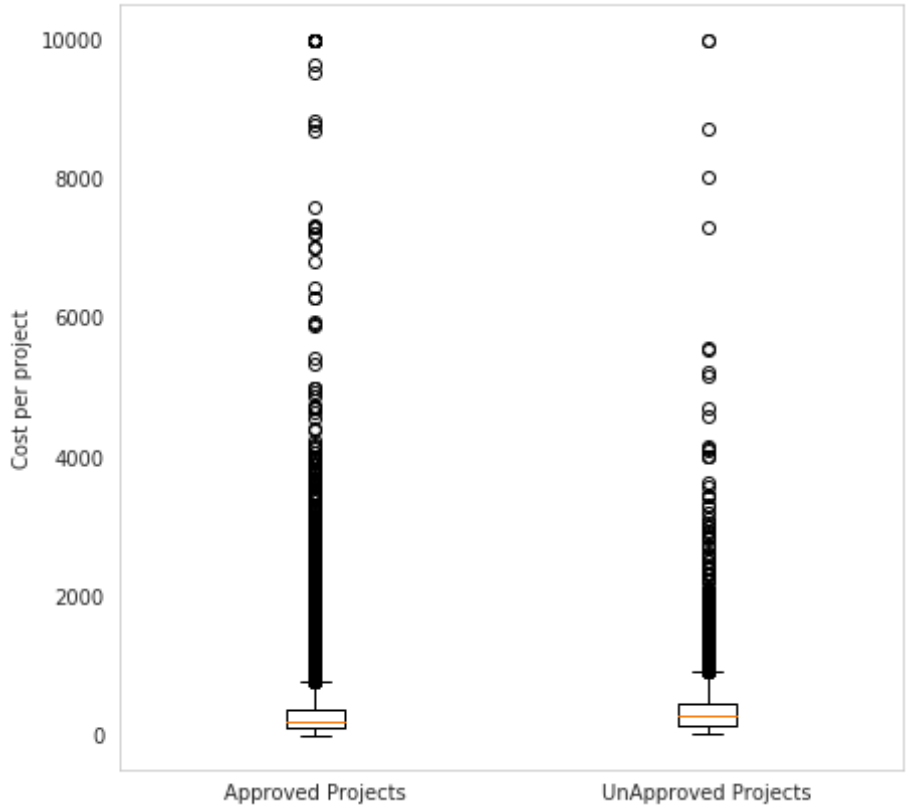
1 rows × 22 columns

In [124]:
```
priceAndQuantityData[priceAndQuantityData['id'] == 'p253737']
```

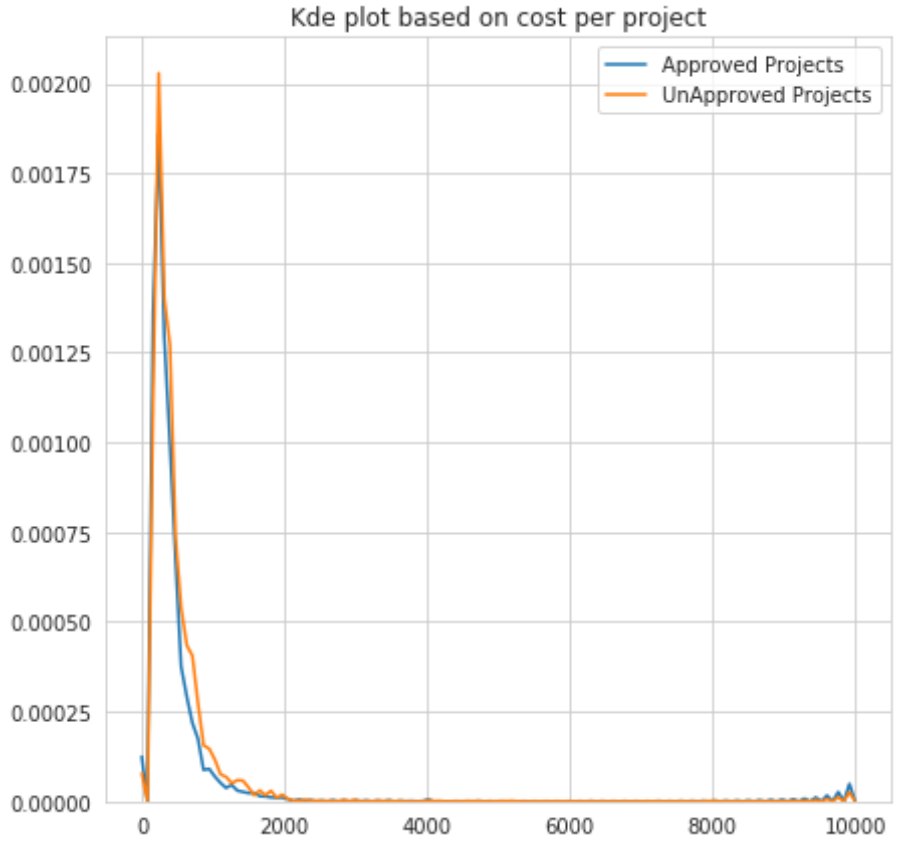
Out[124]:

| | id | price | quantity |
|---|---|---|---|
| **253736** | p253737 | 154.6 | 23 |

In [125]:
```
approvedProjectsPrice = projectsData[projectsData['project_is_approved'
] == 1].price;
unApprovedProjectsPrice = projectsData[projectsData['project_is_approve
d'] == 0].price;
plt.boxplot([approvedProjectsPrice, unApprovedProjectsPrice]);
plt.grid();
plt.xticks([1, 2], ['Approved Projects', 'UnApproved Projects']);
plt.ylabel('Cost per project');
plt.show();
```




In [126]:
```
plt.title("Kde plot based on cost per project");
sbrn.kdeplot(approvedProjectsPrice, label = "Approved Projects", bw =
0.6);
sbrn.kdeplot(unApprovedProjectsPrice, label = "UnApproved Projects", bw
 = 0.6);
plt.legend();
plt.show();
```




In [127]:
```
pricePercentilesApproved = [round(np.percentile(approvedProjectsPrice, percentile), 3) for percentile in np.arange(0, 100, 5)];
pricePercentilesUnApproved = [round(np.percentile(unApprovedProjectsPrice, percentile), 3) for percentile in np.arange(0, 100, 5)];
percentileValuePricesData = pd.DataFrame({'Percentile': np.arange(0, 100, 5), 'Approved projects': pricePercentilesApproved, 'UnApproved Projects': pricePercentilesUnApproved});
percentileValuePricesData
```

Out[127]:

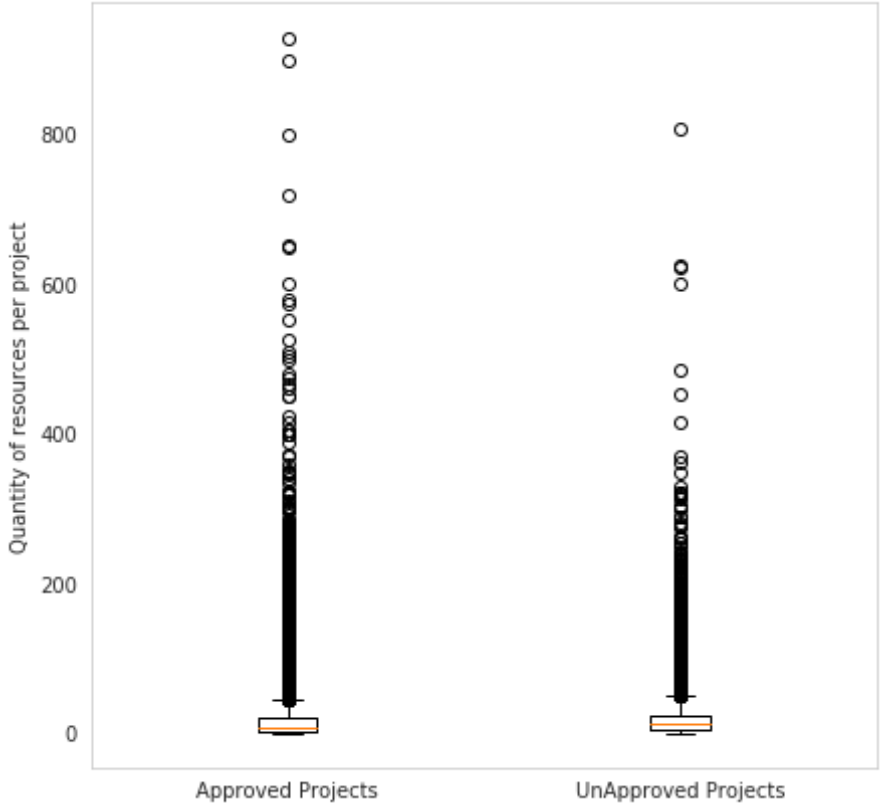
| | Percentile | Approved projects | UnApproved Projects |
|---|---|---|---|
| **0** | 0 | 0.660 | 1.970 |

| | Percentile | Approved projects | UnApproved Projects |
|---|---|---|---|
| **1** | 5 | 13.590 | 41.900 |
| **2** | 10 | 33.880 | 73.670 |
| **3** | 15 | 58.000 | 99.109 |
| **4** | 20 | 77.380 | 118.560 |
| **5** | 25 | 99.950 | 140.892 |
| **6** | 30 | 116.680 | 162.230 |
| **7** | 35 | 137.232 | 184.014 |
| **8** | 40 | 157.000 | 208.632 |
| **9** | 45 | 178.265 | 235.106 |
| **10** | 50 | 198.990 | 263.145 |
| **11** | 55 | 223.990 | 292.610 |
| **12** | 60 | 255.630 | 325.144 |
| **13** | 65 | 285.412 | 362.390 |
| **14** | 70 | 321.225 | 399.990 |
| **15** | 75 | 366.075 | 449.945 |
| **16** | 80 | 411.670 | 519.282 |
| **17** | 85 | 479.000 | 618.276 |
| **18** | 90 | 593.110 | 739.356 |
| **19** | 95 | 801.598 | 992.486 |

## Observation:

1. Most of the projects proposed are of less cost.

In [128]:
```
approvedProjectsQuantity = projectsData[projectsData['project_is_approved'] == 1].quantity;
unApprovedProjectsQuantity = projectsData[projectsData['project_is_approved'] == 0].quantity;
plt.boxplot([approvedProjectsQuantity, unApprovedProjectsQuantity]);
plt.grid();
plt.xticks([1, 2], ['Approved Projects', 'UnApproved Projects']);
plt.ylabel('Quantity of resources per project');
plt.show();
```




In [129]:
```
plt.title("Kde plot based on quantity of resources per project");
sbrn.kdeplot(approvedProjectsQuantity, label = "Approved Projects", bw = 0.6);
```

```
sbrn.kdeplot(unApprovedProjectsQuantity, label = "UnApproved Projects",
 bw = 0.6);
plt.legend();
plt.show();
```


In [130]:
```
quantityPercentilesApproved = [round(np.percentile(approvedProjectsQuan
tity, percentile), 3) for percentile in np.arange(0, 100, 5)];
quantityPercentilesUnApproved = [round(np.percentile(unApprovedProjects
Quantity, percentile), 3) for percentile in np.arange(0, 100, 5)];
percentileValueQuantitiesData = pd.DataFrame({'Percentile': np.arange(0
, 100, 5), 'Approved projects': quantityPercentilesApproved, 'UnApprove
d Projects': quantityPercentilesUnApproved});
percentileValueQuantitiesData
```

Out[130]:

| | Percentile | Approved projects | UnApproved Projects |
|---|---|---|---|
| **0** | 0 | 1.0 | 1.0 |
| **1** | 5 | 1.0 | 2.0 |
| **2** | 10 | 1.0 | 3.0 |
| **3** | 15 | 2.0 | 4.0 |
| **4** | 20 | 3.0 | 5.0 |
| **5** | 25 | 3.0 | 6.0 |
| **6** | 30 | 4.0 | 7.0 |
| **7** | 35 | 5.0 | 8.0 |
| **8** | 40 | 6.0 | 9.0 |
| **9** | 45 | 7.0 | 10.0 |
| **10** | 50 | 8.0 | 12.0 |
| **11** | 55 | 10.0 | 13.0 |
| **12** | 60 | 11.0 | 15.0 |
| **13** | 65 | 14.0 | 18.0 |
| **14** | 70 | 16.0 | 20.0 |
| **15** | 75 | 20.0 | 24.0 |
| **16** | 80 | 25.0 | 29.0 |
| **17** | 85 | 30.0 | 35.0 |
| **18** | 90 | 38.0 | 45.0 |
| **19** | 95 | 56.0 | 63.0 |

In [131]:
```
sbrn.set_style('whitegrid');
```

```
sbrn.FacetGrid(projectsData, hue = 'project_is_approved', size = 6) \
    .map(plt.scatter, 'price', 'quantity') \
    .add_legend();
plt.title("Scatter plot between price and quantity based project approv
al and rejection");
plt.show();
```


## Observation:

1. When plotted scatter plot between approved and rejected projects based on price and quantity there is huge overlap. So the projects approval is not actually depending on

price and quantity resources of the project.

## Univariate Analysis: teacher_number_of_previously_posted_projects

In [132]:
```
projectsData.head(5)
```

Out[132]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY |

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **4** | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX |

5 rows × 22 columns

In [133]:
```
previouslyPostedApprovedNumberData = projectsData.groupby('teacher_numb
er_of_previously_posted_projects')['project_is_approved'].agg(lambda x:
 x.eq(1).sum()).reset_index();
previouslyPostedRejectedNumberData = projectsData.groupby('teacher_numb
er_of_previously_posted_projects')['project_is_approved'].agg(lambda x:
 x.eq(0).sum()).reset_index();
print("Total number of projects approved: ", len(projectsData[projectsD
ata['project_is_approved'] == 1]));
print("Total number of projects rejected: ", len(projectsData[projectsD
ata['project_is_approved'] == 0]));
print("Number of projects approved categorized by previously_posted: ",
 previouslyPostedApprovedNumberData['project_is_approved'].sum());
print("Number of projects rejected categorized by previously_posted: ",
 previouslyPostedRejectedNumberData['project_is_approved'].sum());
previouslyPostedNumberData = pd.merge(previouslyPostedApprovedNumberDat
a, previouslyPostedRejectedNumberData, on = 'teacher_number_of_previous
ly_posted_projects', how = 'inner');
previouslyPostedNumberData.head(5)
```

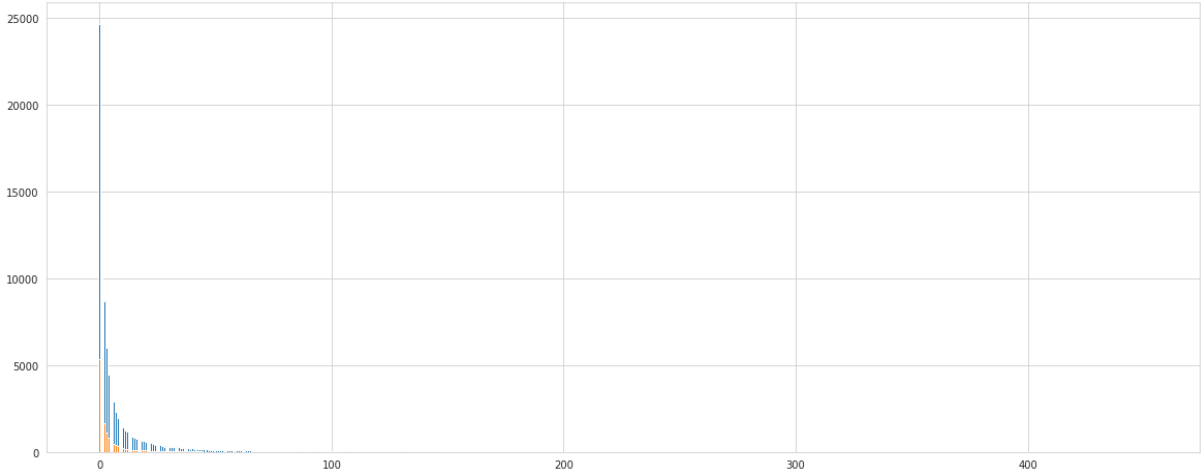
```
Total number of projects approved:  92706
Total number of projects rejected:  16542
Number of projects approved categorized by previously_posted:  92706
Number of projects rejected categorized by previously_posted:  16542
```

Out[133]:

| | teacher_number_of_previously_posted_projects | project_is_approved_x | project_is_ap |
|---|---|---|---|

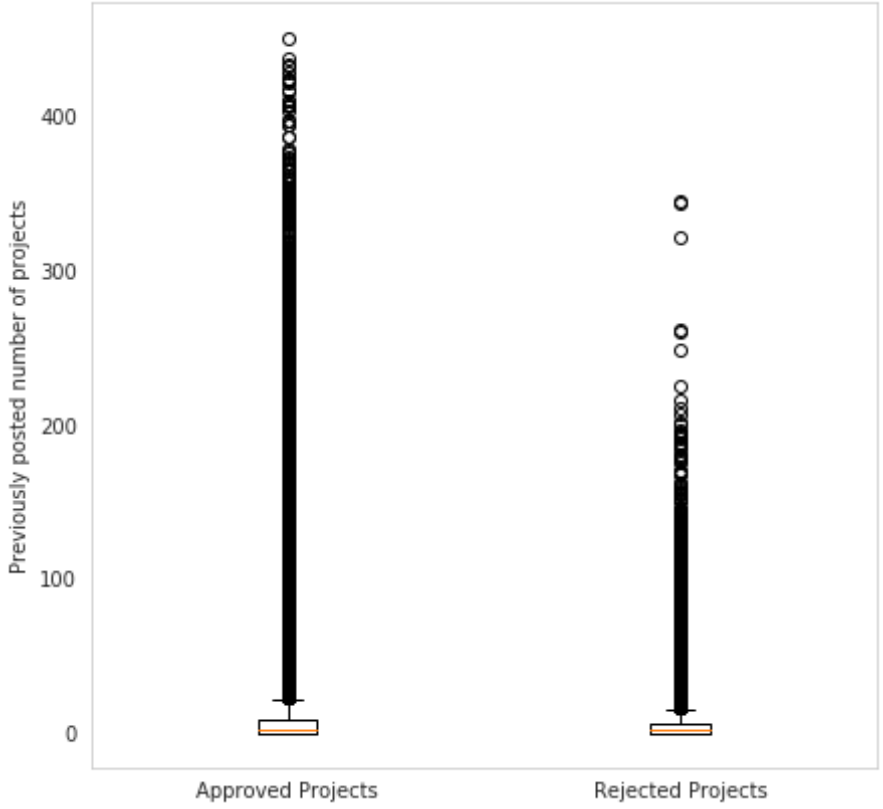
| | teacher_number_of_previously_posted_projects | project_is_approved_x | project_is_ap |
|---|---|---|---|
| **0** | 0 | 24652 | 5362 |
| **1** | 1 | 13329 | 2729 |
| **2** | 2 | 8705 | 1645 |
| **3** | 3 | 5997 | 1113 |
| **4** | 4 | 4452 | 814 |

In [134]:
```
plt.figure(figsize = (20, 8));
plt.bar(previouslyPostedNumberData.teacher_number_of_previously_posted_
projects, previouslyPostedNumberData.project_is_approved_x);
plt.bar(previouslyPostedNumberData.teacher_number_of_previously_posted_
projects, previouslyPostedNumberData.project_is_approved_y);
plt.show();
```

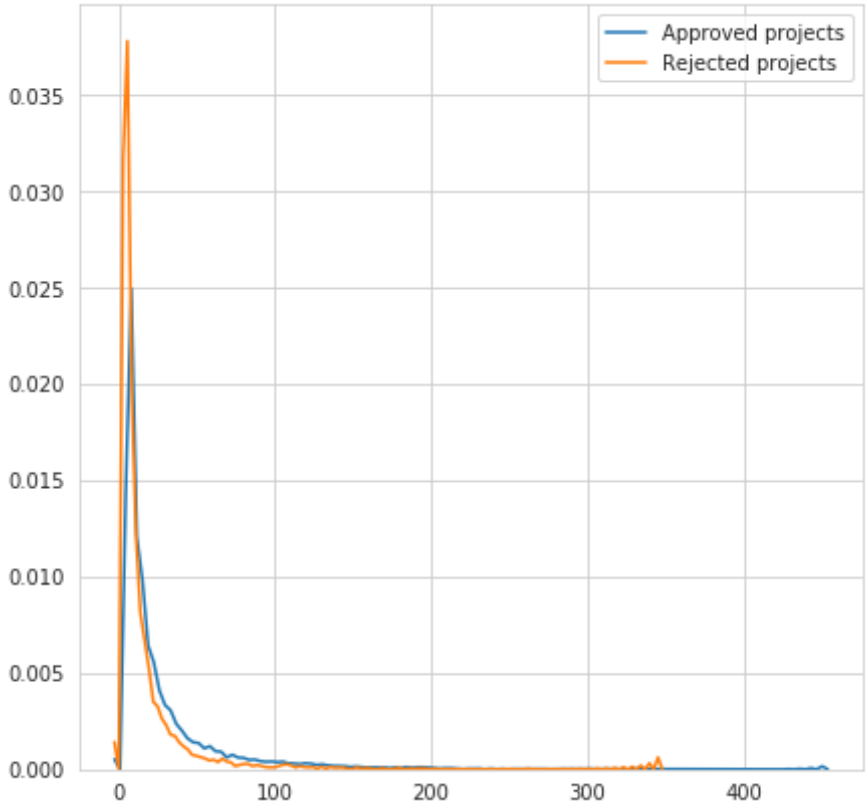



In [135]:
```
previouslyPostedApprovedData = projectsData[projectsData['project_is_ap
proved'] == 1].teacher_number_of_previously_posted_projects;
previouslyPostedRejectedData = projectsData[projectsData['project_is_ap
proved'] == 0].teacher_number_of_previously_posted_projects;
```

```
plt.boxplot([previouslyPostedApprovedData, previouslyPostedRejectedData
]);
plt.grid();
plt.xticks([1, 2], ['Approved Projects', 'Rejected Projects']);
plt.ylabel('Previously posted number of projects');
plt.show();
```




In [136]:
```
sbrn.kdeplot(previouslyPostedApprovedData, label = "Approved projects",
 bw = 1);
sbrn.kdeplot(previouslyPostedRejectedData, label = "Rejected projects",
 bw = 1);
plt.show();
```




## Observation:

1. Most of the projects approved and rejected are with less number of teacher_number_of_previously_posted_projects. So the approval is not much depending on how many number of projects proposed by teacher previously.

In [0]:
```
def stringContainsNumbers(string):
    return any([character.isdigit() for character in string])
```

In [138]:
```
numericResourceApprovedData = projectsData[(projectsData['project_resou
rce_summary'].apply(stringContainsNumbers) == True) & (projectsData['pr
```

```
oject_is_approved'] == 1)]
textResourceApprovedData = projectsData[(projectsData['project_resource
_summary'].apply(stringContainsNumbers) == False) & (projectsData['proj
ect_is_approved'] == 1)]
numericResourceRejectedData = projectsData[(projectsData['project_resou
rce_summary'].apply(stringContainsNumbers) == True) & (projectsData['pr
oject_is_approved'] == 0)]
textResourceRejectedData = projectsData[(projectsData['project_resource
_summary'].apply(stringContainsNumbers) == False) & (projectsData['proj
ect_is_approved'] == 0)]
print("Checking whether numbers in resource summary will be useful for
 project approval?");
equalsBorder(70);
print("Number of approved projects with numbers in resource summary: ",
 numericResourceApprovedData.shape[0]);
print("Number of rejected projects with numbers in resource summary: ",
 numericResourceRejectedData.shape[0]);
print("Number of approved projects without numbers in resource summary:
 ", textResourceApprovedData.shape[0]);
print("Number of rejected projects without numbers in resource summary:
 ", textResourceRejectedData.shape[0]);
```

```
Checking whether numbers in resource summary will be useful for project
approval?
======================================================================
Number of approved projects with numbers in resource summary:  14090
Number of rejected projects with numbers in resource summary:  1666
Number of approved projects without numbers in resource summary:  78616
Number of rejected projects without numbers in resource summary:  14876
```

## Observation:

1. The rejection rate of project is less when projects resource summary has numbers in it.
2. Even the number of projects approved without numbers in resource summary is high which means that the classification does not actually depends on whether resource summary contains numerical digits or not.

## Conclusion of univariate analysis:

1. There is huge overlap of approved and rejected projects when taken for all single featurs. So, this project cannot be classified using single features.
2. project_title is some what better in text type of feature because of less overlap than others.
3. The project approval is not depending on resources cost, but the probability of project rejection is more when resources cost is more.

## Preprocessing data

In [0]:
```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'no
t'
# All stopwords that are needed to be removed in the text
stopWords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'o
urs', 'ourselves', 'you', "you're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselve
s', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'it
s', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'th
is', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'h
ave', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between',
'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',
'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'h
ow', 'all', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 's
o', 'than', 'too', 'very', \
```

```
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should',
"should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
 "shouldn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"]);
def preProcessingWithAndWithoutStopWords(texts):
    """
    This function takes list of texts and returns preprocessed list of
 texts one with
    stop words and one without stopwords.
    """
    # Variable for storing preprocessed text with stop words
    preProcessedTextsWithStopWords = [];
    # Variable for storing preprocessed text without stop words
    preProcessedTextsWithoutStopWords = [];

    # Looping over list of texts for performing pre processing
    for text in tqdm(texts, total = len(texts)):
        # Removing all links in the text
        text = re.sub(r"http\S+", "", text);

        # Removing all html tags in the text
        text = re.sub(r"<\w+/>", "", text);
        text = re.sub(r"<\w+>", "", text);

        # https://stackoverflow.com/a/47091490/4084039
        # Replacing all below words with adverbs
        text = re.sub(r"won't", "will not", text)
        text = re.sub(r"can\'t", "can not", text)
        text = re.sub(r"n\'t", " not", text)
        text = re.sub(r"\'re", " are", text)
        text = re.sub(r"\'s", " is", text)
        text = re.sub(r"\'d", " would", text)
        text = re.sub(r"\'ll", " will", text)
        text = re.sub(r"\'t", " not", text)
```

```
        text = re.sub(r"\'ve", " have", text)
        text = re.sub(r"\'m", " am", text)

        # Removing backslash symbols in text
        text = text.replace('\\r', ' ');
        text = text.replace('\\n', ' ');
        text = text.replace('\\"', ' ');

        # Removing all special characters of text
        text = re.sub(r"[^a-zA-Z0-9]+", " ", text);

        # Converting whole review text into lower case
        text = text.lower();

        # adding this preprocessed text without stopwords to list
        preProcessedTextsWithStopWords.append(text);

        # removing stop words from text
        textWithoutStopWords = ' '.join([word for word in text.split()
if word not in stopWords]);
        # adding this preprocessed text without stopwords to list
        preProcessedTextsWithoutStopWords.append(textWithoutStopWords);

    return [preProcessedTextsWithStopWords, preProcessedTextsWithoutSto
pWords];
```

In [140]:

```
texts = [projectsData['project_essay'].values[0]]
preProcessedTextsWithStopWords, preProcessedTextsWithoutStopWords = pre
ProcessingWithAndWithoutStopWords(texts);
print("Example project essay without pre-processing: ");
equalsBorder(70);
print(texts);
equalsBorder(70);
print("Example project essay with stop words and pre-processing: ");
equalsBorder(70);
print(preProcessedTextsWithStopWords);
equalsBorder(70);
print("Example project essay without stop words and pre-processing: ");
```

```
equalsBorder(70);
print(preProcessedTextsWithoutStopWords);
```

```
Example project essay without pre-processing:
======================================================================
['My students are English learners that are working on English as their
second or third languages. We are a melting pot of refugees, immigrant
s, and native-born Americans bringing the gift of language to our schoo
l. \\r\\n\\r\\n We have over 24 languages represented in our English Le
arner program with students at every level of mastery.  We also have ov
er 40 countries represented with the families within our school.  Each
student brings a wealth of knowledge and experiences to us that open ou
r eyes to new cultures, beliefs, and respect.\\"The limits of your lang
uage are the limits of your world.\\"-Ludwig Wittgenstein  Our English
learner\'s have a strong support system at home that begs for more reso
urces.  Many times our parents are learning to read and speak English a
long side of their children.  Sometimes this creates barriers for paren
ts to be able to help their child learn phonetics, letter recognition,
and other reading skills.\\r\\n\\r\\nBy providing these dvd\'s and play
ers, students are able to continue their mastery of the English languag
e even if no one at home is able to assist.  All families with students
within the Level 1 proficiency status, will be a offered to be a part o
f this program.  These educational videos will be specially chosen by t
he English Learner Teacher and will be sent home regularly to watch.  T
he videos are to help the child develop early reading skills.\\r\\n\\r
\\nParents that do not have access to a dvd player will have the opport
unity to check out a dvd player to use for the year.  The plan is to us
e these videos and educational dvd\'s for the years to come for other E
L students.\\r\\nnannan']
======================================================================
Example project essay with stop words and pre-processing:
======================================================================
['my students are english learners that are working on english as their
second or third languages we are a melting pot of refugees immigrants a
nd native born americans bringing the gift of language to our school we
have over 24 languages represented in our english learner program with
students at every level of mastery we also have over 40 countries repre
sented with the families within our school each student brings a wealth
of knowledge and experiences to us that open our eyes to new cultures b
eliefs and respect the limits of your language are the limits of your w
```

```
orld ludwig wittgenstein our english learner is have a strong support s
ystem at home that begs for more resources many times our parents are l
earning to read and speak english along side of their children sometime
s this creates barriers for parents to be able to help their child lear
n phonetics letter recognition and other reading skills by providing th
ese dvd is and players students are able to continue their mastery of t
he english language even if no one at home is able to assist all famili
es with students within the level 1 proficiency status will be a offere
d to be a part of this program these educational videos will be special
ly chosen by the english learner teacher and will be sent home regularl
y to watch the videos are to help the child develop early reading skill
s parents that do not have access to a dvd player will have the opportu
nity to check out a dvd player to use for the year the plan is to use t
hese videos and educational dvd is for the years to come for other el s
tudents nannan']
==========================================================================
Example project essay without stop words and pre-processing:
==========================================================================
['students english learners working english second third languages melt
ing pot refugees immigrants native born americans bringing gift languag
e school 24 languages represented english learner program students ever
y level mastery also 40 countries represented families within school st
udent brings wealth knowledge experiences us open eyes new cultures bel
iefs respect limits language limits world ludwig wittgenstein english l
earner strong support system home begs resources many times parents lea
rning read speak english along side children sometimes creates barriers
parents able help child learn phonetics letter recognition reading skil
ls providing dvd players students able continue mastery english languag
e even no one home able assist families students within level 1 profici
ency status offered part program educational videos specially chosen en
glish learner teacher sent home regularly watch videos help child devel
op early reading skills parents not access dvd player opportunity check
dvd player use year plan use videos educational dvd years come el stude
nts nannan']
```

In [141]:
```
projectEssays = projectsData['project_essay'];
preProcessedEssaysWithStopWords, preProcessedEssaysWithoutStopWords = p
reProcessingWithAndWithoutStopWords(projectEssays);
```

In [142]:
```
preProcessedEssaysWithoutStopWords[0:3]
```

Out[142]:
```
['students english learners working english second third languages melt
ing pot refugees immigrants native born americans bringing gift languag
e school 24 languages represented english learner program students ever
y level mastery also 40 countries represented families within school st
udent brings wealth knowledge experiences us open eyes new cultures bel
iefs respect limits language limits world ludwig wittgenstein english l
earner strong support system home begs resources many times parents lea
rning read speak english along side children sometimes creates barriers
parents able help child learn phonetics letter recognition reading skil
ls providing dvd players students able continue mastery english languag
e even no one home able assist families students within level 1 profici
ency status offered part program educational videos specially chosen en
glish learner teacher sent home regularly watch videos help child devel
op early reading skills parents not access dvd player opportunity check
dvd player use year plan use videos educational dvd years come el stude
nts nannan',
 'students arrive school eager learn polite generous strive best know e
ducation succeed life help improve lives school focuses families low in
comes tries give student education deserve not much students use materi
als given best projector need school crucial academic improvement stude
nts technology continues grow many resources internet teachers use grow
th students however school limited resources particularly technology wi
thout disadvantage one things could really help classrooms projector pr
ojector not crucial instruction also growth students projector show pre
sentations documentaries photos historical land sites math problems muc
h projector make teaching learning easier also targeting different type
s learners classrooms auditory visual kinesthetic etc nannan',
 'true champions not always ones win guts mia hamm quote best describes
students cholla middle school approach playing sports especially girls
boys soccer teams teams made 7th 8th grade students not opportunity pla
y organized sport due family financial difficulties teach title one mid
dle school urban neighborhood 74 students qualify free reduced lunch ma
ny come activity sport opportunity poor homes students love participate
sports learn new skills apart team atmosphere school lacks funding meet
students needs concerned lack exposure not prepare participating sports
```

```
teams high school end school year goal provide students opportunity lea
rn variety soccer skills positive qualities person actively participate
s team students campus come school knowing face uphill battle comes par
ticipating organized sports players would thrive field confidence appro
priate soccer equipment play soccer best abilities students experience
helpful person part team teaches positive supportive encouraging others
students using soccer equipment practice games daily basis learn practi
ce necessary skills develop strong soccer team experience create opport
unity students learn part team positive contribution teammates students
get opportunity learn practice variety soccer skills use skills game ac
cess type experience nearly impossible without soccer equipment student
s players utilize practice games nannan']
```

In [143]:
```
projectTitles = projectsData['project_title'];
preProcessedProjectTitlesWithStopWords, preProcessedProjectTitlesWithou
tStopWords = preProcessingWithAndWithoutStopWords(projectTitles);
preProcessedProjectTitlesWithoutStopWords[0:5]
```

Out[143]:
```
['educational support english learners home',
 'wanted projector hungry learners',
 'soccer equipment awesome middle school students',
 'techie kindergarteners',
 'interactive math tools']
```

In [144]:
```
projectsData['preprocessed_titles'] = preProcessedProjectTitlesWithoutS
topWords;
projectsData['preprocessed_essays'] = preProcessedEssaysWithoutStopWord
s;
projectsData.shape
```

Out[144]:
```
(109248, 24)
```

## Preparing data for classification and modelling

In [0]:
```
pd.DataFrame(projectsData.columns, columns = ['All features in projects
 data'])
```

Out[0]:

| | All features in projects data |
|---|---|
| **0** | Unnamed: 0 |
| **1** | id |
| **2** | teacher_id |
| **3** | teacher_prefix |
| **4** | school_state |
| **5** | project_submitted_datetime |
| **6** | project_grade_category |
| **7** | project_subject_categories |
| **8** | project_subject_subcategories |
| **9** | project_title |
| **10** | project_essay_1 |
| **11** | project_essay_2 |
| **12** | project_essay_3 |
| **13** | project_essay_4 |
| **14** | project_resource_summary |
| **15** | teacher_number_of_previously_posted_projects |
| **16** | project_is_approved |
| **17** | cleaned_categories |
| **18** | cleaned_sub_categories |
| **19** | project_essay |
| **20** | price |

| | All features in projects data |
|---|---|
| **21** | quantity |

## Useful features:

**Here we will consider only below features for classification and we can ignore the other features**

***Categorical data:***

1. **school_state** - categorical data
2. **project_grade_category** - categorical data
3. **cleaned_categories** - categorical data
4. **cleaned_sub_categories** - categorical data
5. **teacher_prefix** - categorical data

***Text data:***

1. **project_resource_summary** - text data
2. **project_title - text data**
3. **project_resource_summary** - text data

***Numerical data:***

1. **teacher_number_of_previously_posted_projects** - numerical data
2. **price** - numerical data
3. **quantity** - numerical data

# Vectorizing categorical data

### 1. Vectorizing cleaned_categories(project_subject_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_categories
subjectsCategoriesVectorizer = CountVectorizer(vocabulary = list(sorted
CategoriesDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with cleaned_categories values
subjectsCategoriesVectorizer.fit(projectsData['cleaned_categories'].val
ues);
# Vectorizing categories using one-hot-encoding
categoriesVectors = subjectsCategoriesVectorizer.transform(projectsData
['cleaned_categories'].values);
```

In [0]:
```
print("Features used in vectorizing categories: ");
equalsBorder(70);
print(subjectsCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", categoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(categoriesVectors[0:4])
```

```
Features used in vectorizing categories:
======================================================================
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearn
ing', 'SpecialNeeds', 'Health_Sports', 'Math_Science', 'Literacy_Langua
ge']
======================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (109248, 9)
======================================================================
Sample vectors of categories:
======================================================================
  (0, 8)        1
  (1, 2)        1
  (1, 6)        1
```

```
  (2, 6)        1
  (3, 7)        1
  (3, 8)        1
```

### 2. Vectorizing cleaned_sub_categories(project_subject_sub_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_sub_categories
subjectsSubCategoriesVectorizer = CountVectorizer(vocabulary = list(sor
tedDictionarySubCategories.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with cleaned_sub_categories values
subjectsSubCategoriesVectorizer.fit(projectsData['cleaned_sub_categorie
s'].values);
# Vectorizing sub categories using one-hot-encoding
subCategoriesVectors = subjectsSubCategoriesVectorizer.transform(projec
tsData['cleaned_sub_categories'].values);
```

In [0]:
```
print("Features used in vectorizing subject sub categories: ");
equalsBorder(70);
print(subjectsSubCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", subCategoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(subCategoriesVectors[0:4])
```

```
Features used in vectorizing subject sub categories:
======================================================================
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolveme
nt', 'Extracurricular', 'Civics_Government', 'ForeignLanguages', 'Nutri
tionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingA
rts', 'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPre
p', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopme
```

```
nt', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Healt
h_Wellness', 'AppliedSciences', 'SpecialNeeds', 'Literature_Writing',
'Mathematics', 'Literacy']
==================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (109248, 30)
==================================================================
Sample vectors of categories:
==================================================================
  (0, 20)       1
  (0, 29)       1
  (1, 5)        1
  (1, 13)       1
  (2, 13)       1
  (2, 24)       1
  (3, 28)       1
  (3, 29)       1
```

### 3. Vectorizing teacher_prefix - One Hot Encoding

In [0]:
```
def giveCounter(data):
    counter = Counter();
    for dataValue in data:
        counter.update(str(dataValue).split());
    return counter
```

In [0]:
```
giveCounter(projectsData['teacher_prefix'].values)
```

Out[0]:
```
Counter({'Mrs.': 57269,
         'Mr.': 10648,
         'Ms.': 38955,
         'Teacher': 2360,
         'nan': 3,
         'Dr.': 13})
```

In [0]:
```
projectsData = projectsData.dropna(subset = ['teacher_prefix']);
projectsData.shape
```

Out[0]: (109245, 22)

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(projectsData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(projectsData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(projectsData[
'teacher_prefix'].values);
```

In [0]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
ing): ", teacherPrefixVectors.shape);
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:100]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs.', 'Mr.', 'Ms.', 'Teacher', 'Dr.']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(109245, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (27, 3)       1
  (75, 3)       1
  (82, 3)       1
  (88, 3)       1
```

In [0]:
```
teacherPrefixes = [prefix.replace('.', '') for prefix in projectsData[
'teacher_prefix'].values];
teacherPrefixes[0:5]
```

Out[0]:
```
['Mrs', 'Mr', 'Ms', 'Mrs', 'Mrs']
```

In [0]:
```
projectsData['teacher_prefix'] = teacherPrefixes;
projectsData.head(3)
```

Out[0]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr | FL |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms | AZ |

3 rows × 22 columns

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(projectsData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
```

```
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(projectsData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(projectsData[
'teacher_prefix'].values);
```

In [0]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
ing): ", teacherPrefixVectors.shape);
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:4]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs', 'Mr', 'Ms', 'Teacher', 'Dr']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(109245, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 2)        1
  (3, 0)        1
```

### 4. Vectorizing school_state - One Hot Encoding

In [0]:
```
schoolStateDictionary = dict(giveCounter(projectsData['school_state'].v
```

```
alues));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique school states
schoolStateVectorizer = CountVectorizer(vocabulary = list(schoolStateDi
ctionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with school_state values
schoolStateVectorizer.fit(projectsData['school_state'].values);
# Vectorizing school_state using one-hot-encoding
schoolStateVectors = schoolStateVectorizer.transform(projectsData['scho
ol_state'].values);
```

In [0]:

```
print("Features used in vectorizing school_state: ");
equalsBorder(70);
print(schoolStateVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of school_state matrix after vectorization(one-hot-encodin
g): ", schoolStateVectors.shape);
equalsBorder(70);
print("Sample vectors of school_state: ");
equalsBorder(70);
print(schoolStateVectors[0:4]);
```

```
Features used in vectorizing school_state:
======================================================================
['IN', 'FL', 'AZ', 'KY', 'TX', 'CT', 'GA', 'SC', 'NC', 'CA', 'NY', 'O
K', 'MA', 'NV', 'OH', 'PA', 'AL', 'LA', 'VA', 'AR', 'WA', 'WV', 'ID',
'TN', 'MS', 'CO', 'UT', 'IL', 'MI', 'HI', 'IA', 'RI', 'NJ', 'MO', 'DE',
'MN', 'ME', 'WY', 'ND', 'OR', 'AK', 'MD', 'WI', 'SD', 'NE', 'NM', 'DC',
'KS', 'MT', 'NH', 'VT']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (1
09245, 51)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 2)        1
  (3, 3)        1
```

## 5. Vectorizing project_grade_category - One Hot Encoding

In [0]:
```
giveCounter(projectsData['project_grade_category'])
```

Out[0]:
```
Counter({'Grades': 109245,
         'PreK-2': 44225,
         '6-8': 16923,
         '3-5': 37135,
         '9-12': 10962})
```

In [0]:
```
cleanedGrades = []
for grade in projectsData['project_grade_category'].values:
    grade = grade.replace(' ', '');
    grade = grade.replace('-', 'to');
    cleanedGrades.append(grade);
cleanedGrades[0:4]
```

Out[0]:
```
['GradesPreKto2', 'Grades6to8', 'Grades6to8', 'GradesPreKto2']
```

In [0]:
```
projectsData['project_grade_category'] = cleanedGrades
projectsData.head(4)
```

Out[0]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs | IN |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr | FL |

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state |
|---|---|---|---|---|---|
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms | AZ |
| **3** | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs | KY |

4 rows × 22 columns

In [0]:
```
projectGradeDictionary = dict(giveCounter(projectsData['project_grade_category'].values));
# Using CountVectorizer for performing one-hot-encoding by setting vocabulary as list of all unique project grade categories
projectGradeVectorizer = CountVectorizer(vocabulary = list(projectGradeDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with project_grade_category values
projectGradeVectorizer.fit(projectsData['project_grade_category'].values);
# Vectorizing project_grade_category using one-hot-encoding
projectGradeVectors = projectGradeVectorizer.transform(projectsData['project_grade_category'].values);
```

In [0]:
```
print("Features used in vectorizing project_grade_category: ");
equalsBorder(70);
print(projectGradeVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of school_state matrix after vectorization(one-hot-encoding): ", projectGradeVectors.shape);
equalsBorder(70);
```

```
print("Sample vectors of school_state: ");
equalsBorder(70);
print(projectGradeVectors[0:4]);
```

```
Features used in vectorizing project_grade_category:
======================================================================
['GradesPreKto2', 'Grades6to8', 'Grades3to5', 'Grades9to12']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (1
09245, 4)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 1)        1
  (3, 0)        1
```

In [0]:
```
projectsDataSub = projectsData[0:40000];
preProcessedEssaysWithoutStopWordsSub = preProcessedEssaysWithoutStopWo
rds[0:40000];
preProcessedProjectTitlesWithoutStopWordsSub = preProcessedProjectTitle
sWithoutStopWords[0:40000];
```

# Vectorizing Text Data

## Bag of Words

#### 1. Vectorizing project_essay

In [0]:
```
# Initializing countvectorizer for bag of words vectorization of prepro
cessed project essays
bowEssayVectorizer = CountVectorizer(min_df = 10);
# Transforming the preprocessed essays to bag of words vectors
```

```
bowEssayModel = bowEssayVectorizer.fit_transform(preProcessedEssaysWith
outStopWordsSub);
```

In [0]:
```
print("Some of the Features used in vectorizing preprocessed essays: "
);
equalsBorder(70);
print(bowEssayVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed essay matrix after vectorization: ", bowEs
sayModel.shape);
equalsBorder(70);
print("Sample bag-of-words vector of preprocessed essay: ");
equalsBorder(70);
print(bowEssayModel[0])
```

```
Some of the Features used in vectorizing preprocessed essays:
======================================================================
['yeats', 'yell', 'yelling', 'yellow', 'yemen', 'yes', 'yesterday', 'ye
t', 'yield', 'yields', 'yoga', 'york', 'younannan', 'young', 'younger',
'youngest', 'youngsters', 'youth', 'youthful', 'youths', 'youtube', 'yu
mmy', 'zeal', 'zearn', 'zen', 'zenergy', 'zero', 'zest', 'zip', 'ziplo
c', 'zippers', 'zipping', 'zone', 'zoned', 'zones', 'zoo', 'zoom', 'zoo
ming', 'zoos', 'zumba']
======================================================================
Shape of preprocessed essay matrix after vectorization:  (40000, 11077)
======================================================================
Sample bag-of-words vector of preprocessed essay:
======================================================================
  (0, 6533)     1
  (0, 3306)     1
  (0, 1981)     1
  (0, 11036)    1
  (0, 7347)     1
  (0, 11029)    1
  (0, 10530)    2
  (0, 1734)     1
  (0, 6855)     1
  (0, 7374)     2
  (0, 232)      1
```

```
(0, 6687)     1
(0, 3211)     1
(0, 2805)     1
(0, 10766)    1
(0, 8133)     1
(0, 8803)     1
(0, 9831)     1
(0, 1794)     1
(0, 9237)     1
(0, 10639)    3
(0, 3274)     2
(0, 7068)     1
(0, 6798)     1
(0, 9399)     1
:     :
(0, 6123)     2
(0, 5785)     2
(0, 3613)     1
(0, 7703)     2
(0, 5732)     3
(0, 8269)     2
(0, 67)       1
(0, 8670)     2
(0, 5664)     3
(0, 4383)     1
(0, 1339)     1
(0, 553)      1
(0, 1248)     1
(0, 6549)     1
(0, 5003)     1
(0, 8116)     1
(0, 7501)     1
(0, 6207)     1
(0, 5665)     2
(0, 9968)     1
(0, 8736)     1
(0, 10964)    1
(0, 5733)     1
```

```
  (0, 3449)     7
  (0, 9553)     5
```

**2. Vectorizing project_title**

In [0]:
```
# Initializing countvectorizer for bag of words vectorization of prepro
cessed project titles
bowTitleVectorizer = CountVectorizer(min_df = 10);
# Transforming the preprocessed project titles to bag of words vectors
bowTitleModel = bowTitleVectorizer.fit_transform(preProcessedProjectTit
lesWithoutStopWordsSub);
```

In [0]:
```
print("Some of the Features used in vectorizing preprocessed titles: "
);
equalsBorder(70);
print(bowTitleVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after vectorization: ", bowTi
tleModel.shape);
equalsBorder(70);
print("Sample bag-of-words vector of preprocessed title: ");
equalsBorder(70);
print(bowTitleModel[0])
```

```
Some of the Features used in vectorizing preprocessed titles:
======================================================================
['wireless', 'wise', 'wish', 'within', 'without', 'wizards', 'wo', 'wob
ble', 'wobbles', 'wobbling', 'wobbly', 'wonder', 'wonderful', 'wonder
s', 'word', 'words', 'work', 'workers', 'working', 'works', 'workshop',
'world', 'worlds', 'worms', 'worth', 'would', 'wow', 'write', 'writer',
'writers', 'writing', 'ye', 'year', 'yearbook', 'yes', 'yoga', 'young',
'youth', 'zone', 'zoom']
======================================================================
Shape of preprocessed title matrix after vectorization:  (40000, 1774)
======================================================================
Sample bag-of-words vector of preprocessed title:
======================================================================
  (0, 766)      1
```

```
  (0, 906)      1
  (0, 514)      1
  (0, 1553)     1
  (0, 483)      1
```

## Tf-Idf Vectorization

### 1. Vectorizing project_essay

In [0]:
```
# Intializing tfidf vectorizer for tf-idf vectorization of preprocessed
 project essays
tfIdfEssayVectorizer = TfidfVectorizer(min_df = 10);
# Transforming the preprocessed project essays to tf-idf vectors
tfIdfEssayModel = tfIdfEssayVectorizer.fit_transform(preProcessedEssays
WithoutStopWordsSub);
```

In [0]:
```
print("Some of the Features used in tf-idf vectorizing preprocessed ess
ays: ");
equalsBorder(70);
print(tfIdfEssayVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after tf-idf vectorization: "
, tfIdfEssayModel.shape);
equalsBorder(70);
print("Sample Tf-Idf vector of preprocessed essay: ");
equalsBorder(70);
print(tfIdfEssayModel[0])
```

```
Some of the Features used in tf-idf vectorizing preprocessed essays:
======================================================================
['yeats', 'yell', 'yelling', 'yellow', 'yemen', 'yes', 'yesterday', 'ye
t', 'yield', 'yields', 'yoga', 'york', 'younannan', 'young', 'younger',
'youngest', 'youngsters', 'youth', 'youthful', 'youths', 'youtube', 'yu
mmy', 'zeal', 'zearn', 'zen', 'zenergy', 'zero', 'zest', 'zip', 'ziplo
c', 'zippers', 'zipping', 'zone', 'zoned', 'zones', 'zoo', 'zoom', 'zoo
ming', 'zoos', 'zumba']
```

```
==================================================================================
Shape of preprocessed title matrix after tf-idf vectorization:  (40000,
11077)
==================================================================================
Sample Tf-Idf vector of preprocessed essay:
==================================================================================
  (0, 9553)     0.07732161197654648
  (0, 3449)     0.2978137199079083
  (0, 5733)     0.03611311825070974
  (0, 10964)    0.03819325396356506
  (0, 8736)     0.04966730436190034
  (0, 9968)     0.05933894161734909
  (0, 5665)     0.13189136979245247
  (0, 6207)     0.09909858268088724
  (0, 7501)     0.09797369103397546
  (0, 8116)     0.09716121418147701
  (0, 5003)     0.09174889764250635
  (0, 6549)     0.07739523816315956
  (0, 1248)     0.09041771504928811
  (0, 553)      0.09502243963232913
  (0, 1339)     0.07922532406820633
  (0, 4383)     0.08387324724715874
  (0, 5664)     0.12052414724469786
  (0, 8670)     0.03565737676523101
  (0, 67)       0.0797508795755641
  (0, 8269)     0.18440093271700464
  (0, 5732)     0.23244852084297085
  (0, 7703)     0.0932371184396508
  (0, 3613)     0.033250154942777416
  (0, 5785)     0.08336998078832462
  (0, 6123)     0.18451571587493337
  :     :
  (0, 9399)     0.0680639151319745
  (0, 6798)     0.08632328546640713
  (0, 7068)     0.046135007257522224
  (0, 3274)     0.10489683635458984
  (0, 10639)    0.2063461965343629
  (0, 9237)     0.1100116652395096
  (0, 1794)     0.07900547931629058
```

```
  (0, 9831)     0.03792376194008962
  (0, 8803)     0.09740047454864696
  (0, 8133)     0.09001501053091984
  (0, 10766)    0.07024528926492071
  (0, 2805)     0.05089165427462248
  (0, 3211)     0.06222851802675729
  (0, 6687)     0.022226920710368445
  (0, 232)      0.040248356980164615
  (0, 7374)     0.1846309297399045
  (0, 6855)     0.03799907965204156
  (0, 1734)     0.07743897673831124
  (0, 10530)    0.05491069896079749
  (0, 11029)    0.030886589234837624
  (0, 7347)     0.06268239285732621
  (0, 11036)    0.04610937510882687
  (0, 1981)     0.02654012905964554
  (0, 3306)     0.1031894334469226
  (0, 6533)     0.016043824658976313
```

**2. Vectorizing project_title**

In [0]:
```
# Intializing tfidf vectorizer for tf-idf vectorization of preprocessed
 project titles
tfIdfTitleVectorizer = TfidfVectorizer(min_df = 10);
# Transforming the preprocessed project titles to tf-idf vectors
tfIdfTitleModel = tfIdfTitleVectorizer.fit_transform(preProcessedProjec
tTitlesWithoutStopWordsSub);
```

In [0]:
```
print("Some of the Features used in tf-idf vectorizing preprocessed tit
les: ");
equalsBorder(70);
print(tfIdfTitleVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after tf-idf vectorization: "
, tfIdfTitleModel.shape);
equalsBorder(70);
print("Sample Tf-Idf vector of preprocessed title: ");
```

```
equalsBorder(70);
print(tfIdfTitleModel[0])
```

```
Some of the Features used in tf-idf vectorizing preprocessed titles:
==================================================================
['wireless', 'wise', 'wish', 'within', 'without', 'wizards', 'wo', 'wob
ble', 'wobbles', 'wobbling', 'wobbly', 'wonder', 'wonderful', 'wonder
s', 'word', 'words', 'work', 'workers', 'working', 'works', 'workshop',
'world', 'worlds', 'worms', 'worth', 'would', 'wow', 'write', 'writer',
'writers', 'writing', 'ye', 'year', 'yearbook', 'yes', 'yoga', 'young',
'youth', 'zone', 'zoom']
==================================================================
Shape of preprocessed title matrix after tf-idf vectorization:  (40000,
1774)
==================================================================
Sample Tf-Idf vector of preprocessed title:
==================================================================
  (0, 483)      0.5356140846908081
  (0, 1553)     0.4441059196924978
  (0, 514)      0.4615835742389133
  (0, 906)      0.3400969810242112
  (0, 766)      0.4326223894644794
```

### Average Word2Vector Vectorization

In [0]:
```
# stronging variables into pickle files python: http://www.jessicayung.
com/how-to-use-pickle-to-save-and-load-variables-in-python/
# We should have glove_vectors file for creating below model
with open('glove_vectors', 'rb') as f:
    gloveModel = pickle.load(f)
    gloveWords =  set(gloveModel.keys())
```

In [0]:
```
print("Glove vector of sample word: ");
equalsBorder(70);
print(gloveModel['technology']);
equalsBorder(70);
print("Shape of glove vector: ", gloveModel['technology'].shape);
```

```
Glove vector of sample word:
========================================================================
[-0.26078   -0.36898   -0.022831   0.21666    0.16672   -0.20268
 -3.1219     0.33057    0.71512    0.28874    0.074368  -0.033203
  0.23783    0.21052    0.076562   0.13007   -0.31706   -0.45888
 -0.45463   -0.13191    0.49761    0.072704   0.16811    0.18846
 -0.16688   -0.21973    0.08575   -0.19577   -0.2101    -0.32436
 -0.56336    0.077996  -0.22758   -0.66569    0.14824    0.038945
  0.50881   -0.1352     0.49966   -0.4401    -0.022335  -0.22744
  0.22086    0.21865    0.36647    0.30495   -0.16565    0.038759
  0.28108   -0.2167     0.12453    0.65401    0.34584   -0.2557
 -0.046363  -0.31111   -0.020936  -0.17122   -0.77114    0.29289
 -0.14625    0.39541   -0.078938   0.051127   0.15076    0.085126
  0.183     -0.06755    0.26312    0.0087276  0.0066415  0.37033
  0.03496   -0.12627   -0.052626  -0.34897    0.14672    0.14799
 -0.21821   -0.042785   0.2661    -1.1105     0.31789    0.27278
  0.054468  -0.27458    0.42732   -0.44101   -0.19302   -0.32948
  0.61501   -0.22301   -0.36354   -0.34983   -0.16125   -0.17195
 -3.363      0.45146   -0.13753    0.31107    0.2061     0.33063
  0.45879    0.24256    0.042342   0.074837  -0.12869    0.12066
  0.42843   -0.4704    -0.18937    0.32685    0.26079    0.20518
 -0.18432   -0.47658    0.69193    0.18731   -0.12516    0.35447
 -0.1969    -0.58981   -0.88914    0.5176     0.13177   -0.078557
  0.032963  -0.19411    0.15109    0.10547   -0.1113    -0.61533
  0.0948    -0.3393    -0.20071   -0.30197    0.29531    0.28017
  0.16049    0.25294   -0.44266   -0.39412    0.13486    0.25178
 -0.044114   1.1519     0.32234   -0.34323   -0.10713   -0.15616
  0.031206   0.46636   -0.52761   -0.39296   -0.068424  -0.04072
  0.41508   -0.34564    0.71001   -0.364      0.2996     0.032281
  0.34035    0.23452    0.78342    0.48045   -0.1609     0.40102
 -0.071795  -0.16531    0.082153   0.52065    0.24194    0.17113
  0.33552   -0.15725   -0.38984    0.59337   -0.19388   -0.39864
 -0.47901    1.0835     0.24473    0.41309    0.64952    0.46846
  0.024386  -0.72087   -0.095061   0.10095   -0.025229   0.29435
 -0.57696    0.53166   -0.0058338 -0.3304     0.19661   -0.085206
  0.34225    0.56262    0.19924   -0.027111  -0.44567    0.17266
  0.20887   -0.40702    0.63954    0.50708   -0.31862   -0.39602
 -0.1714    -0.040006  -0.45077   -0.32482   -0.0316     0.54908
 -0.1121     0.12951   -0.33577   -0.52768   -0.44592   -0.45388
```

```
 0.66145    0.33023   -1.9089     0.5318     0.21626   -0.13152
 0.48258    0.68028   -0.84115   -0.51165    0.40017    0.17233
-0.033749   0.045275   0.37398   -0.18252    0.19877    0.1511
 0.029803   0.16657   -0.12987   -0.50489    0.55311   -0.22504
 0.13085   -0.78459    0.36481   -0.27472    0.031805   0.53052
-0.20078    0.46392   -0.63554    0.040289  -0.19142   -0.0097011
 0.068084  -0.10602    0.25567    0.096125  -0.10046    0.15016
-0.26733   -0.26494    0.057888   0.062678  -0.11596    0.28115
 0.25375   -0.17954    0.20615    0.24189    0.062696   0.27719
-0.42601   -0.28619   -0.44697   -0.082253  -0.73415   -0.20675
-0.60289   -0.06728    0.15666   -0.042614   0.41368   -0.17367
-0.54012    0.23883    0.23075    0.13608   -0.058634  -0.089705
 0.18469    0.023634   0.16178    0.23384    0.24267    0.091846 ]
==========================================================================
Shape of glove vector:  (300,)
```

In [0]:
```
def getWord2VecVectors(texts):
    word2VecTextsVectors = [];
    for preProcessedText in tqdm(texts):
        word2VecTextVector = np.zeros(300);
        numberOfWordsInText = 0;
        for word in preProcessedText.split():
            if word in gloveWords:
                word2VecTextVector += gloveModel[word];
                numberOfWordsInText += 1;
        if numberOfWordsInText != 0:
            word2VecTextVector = word2VecTextVector / numberOfWordsInTe
xt;
        word2VecTextsVectors.append(word2VecTextVector);
    return word2VecTextsVectors;
```

**1. Vectorizing project_essay**

In [0]:
```
word2VecEssaysVectors = getWord2VecVectors(preProcessedEssaysWithoutSto
pWords);
```

In [0]:
```
print("Shape of Word2Vec vectorization matrix of essays: {},{}".format(
```

```
len(word2VecEssaysVectors), len(word2VecEssaysVectors[0])));
equalsBorder(70);
print("Sample essay: ");
equalsBorder(70);
print(preProcessedEssaysWithoutStopWords[0]);
equalsBorder(70);
print("Word2Vec vector of sample essay: ");
equalsBorder(70);
print(word2VecEssaysVectors[0]);
```

```
Shape of Word2Vec vectorization matrix of essays: 109248,300
======================================================================
Sample essay:
======================================================================
students english learners working english second third languages meltin
g pot refugees immigrants native born americans bringing gift language
school 24 languages represented english learner program students every
level mastery also 40 countries represented families within school stud
ent brings wealth knowledge experiences us open eyes new cultures belie
fs respect limits language limits world ludwig wittgenstein english lea
rner strong support system home begs resources many times parents learn
ing read speak english along side children sometimes creates barriers p
arents able help child learn phonetics letter recognition reading skill
s providing dvd players students able continue mastery english language
even no one home able assist families students within level 1 proficien
cy status offered part program educational videos specially chosen engl
ish learner teacher sent home regularly watch videos help child develop
early reading skills parents not access dvd player opportunity check dv
d player use year plan use videos educational dvd years come el student
s nannan
======================================================================
Word2Vec vector of sample essay:
======================================================================
[-1.40030644e-02  8.78995685e-02  3.50108161e-02 -5.90358980e-03
 -5.93166809e-02 -6.21039893e-02 -2.96711248e+00  9.45840302e-02
 -8.18737785e-03  4.46964161e-02 -7.64722101e-02  6.97099444e-02
  8.44441262e-02 -1.22974138e-01 -3.55310208e-02 -8.90947154e-02
  1.20959579e-01 -1.21977699e-01  4.61334597e-02 -3.33640832e-02
  1.24900557e-01  7.18837631e-02 -6.14885114e-02 -2.67269047e-02
  6.82086621e-02 -3.60263034e-02  1.17172255e-01 -1.17868631e-01
```

```
-1.13467710e-01 -9.25920168e-02 -2.42461725e-01 -7.92963658e-02
 3.52513154e-03  1.79752468e-01 -4.69217812e-02 -3.56593007e-02
-7.95331477e-03 -6.71107383e-04 -1.80828067e-02 -1.16224805e-02
-3.69645852e-02  1.61287176e-01 -1.75201329e-01 -6.02256376e-02
 1.48811886e-02 -9.00106181e-02  7.72160490e-02  7.42989819e-02
-1.02682389e-02 -1.33311658e-01 -2.82030537e-02 -7.71051879e-03
 7.33988450e-02  3.54095087e-02 -5.80719597e-03 -8.70242758e-02
-3.57117638e-02  2.78475651e-02 -1.54957291e-01 -3.24157495e-02
-5.93266570e-02 -8.80254174e-02  2.18914318e-01 -1.22730395e-02
-1.05831485e-01  1.53985730e-01  7.15618933e-02 -3.97147470e-02
 1.47169116e-01 -4.50476644e-03 -1.49678829e-01  5.52201396e-02
 3.04915879e-02 -6.24086617e-02 -7.68483134e-02 -7.50149195e-02
-1.07105068e-01 -2.69954530e-02  1.28067340e-01 -3.42946330e-02
 4.24139667e-02 -4.49685043e-01  1.52793905e-01 -9.06178181e-02
-6.67951510e-02 -2.72063766e-02  7.37261792e-02 -8.64977130e-02
 1.64616877e-01  4.86745523e-02 -4.44542828e-02 -3.04823530e-02
 2.63897436e-02 -6.59345034e-02 -5.21813664e-02 -7.45015886e-02
-2.21975948e+00  8.57858456e-02  7.73778584e-02  1.14644799e-01
-1.50536483e-01 -5.17326940e-02  3.23826117e-02 -1.15700542e-01
 7.15651973e-02  9.15412617e-02  5.41334631e-02 -1.25451318e-01
 2.80941483e-02 -3.95890262e-02 -1.67010497e-02  1.74708879e-02
 4.58374505e-02  2.56664910e-01  3.74891134e-02  3.00990497e-02
-2.18904765e-01  9.37672966e-02  9.99403436e-02  5.26255996e-02
-6.67958718e-02  5.97650946e-02  4.14311192e-02 -6.85917603e-02
 1.72453235e-02  1.02485026e-01  3.02940430e-02  9.59998859e-03
 1.96364913e-02  1.22438477e-01  7.98410557e-02  1.92611322e-02
 6.44085906e-03  4.94252148e-03 -5.36137718e-03 -1.17976934e-01
 1.77991634e-01 -2.51954819e-02  8.02478188e-02  2.29125079e-01
 3.79080403e-02  1.22892819e-02  7.19621470e-02 -9.25031570e-02
-8.86571674e-02 -4.74898563e-02  1.68688409e-02 -1.15134901e-01
 1.76528904e-01 -6.30485141e-02 -4.99678329e-02 -1.00350507e-01
 1.25089302e-02 -4.08706114e-02  4.50565289e-02  2.49286074e-02
-1.29713758e-03 -3.21404376e-02 -2.52972249e-02 -9.63531510e-02
 8.42448993e-04 -7.29482953e-03 -3.77497893e-02 -9.35034987e-02
-3.45719793e-02  7.15921796e-02 -1.29330935e-01  1.28508101e-02
 4.24846988e-02 -8.43078228e-02  4.79772134e-02 -3.05753799e-02
-3.03772013e-02 -2.10572558e-01 -1.05464289e-03  5.18230436e-02
-4.39921874e-02  5.29591584e-02 -1.08551689e-01  2.88053128e-02
-4.88957058e-02  2.31962381e-01 -2.90986193e-02 -2.83725755e-02
```

```
-6.80350899e-02 -6.99966387e-02 -6.80414679e-02 -7.63552362e-02
-1.59287859e-02 -2.59947651e-03 -7.81848121e-03 -1.14299579e-01
-2.02054698e-02  1.21184430e-03  2.59984919e-02 -7.64172013e-02
 9.47882617e-03 -5.71751181e-02  1.25667972e-01 -4.60388139e-02
 5.51296403e-02 -6.73280980e-02 -2.06862389e-02  1.12049165e-01
-7.63451436e-02  4.71124027e-02  6.32404235e-02 -2.13828034e-02
 1.24239236e-01  5.08985235e-02  2.05136711e-03  1.45916498e-02
 4.25123886e-02 -9.41766832e-02 -3.08569389e-02 -2.57995470e-02
-3.53808765e-02 -7.16000389e-02  1.35426121e-02  4.57596799e-02
-1.85721693e-01 -6.62042523e-02 -1.45448285e-01  5.50366758e-02
-2.09367026e+00  1.23479489e-01 -1.46630889e-01 -8.86940765e-02
-7.32806463e-02 -1.48629733e-01  3.23867248e-03  7.08553181e-02
 1.10315906e-02 -2.35431879e-02 -7.69633283e-02 -1.13640894e-01
 9.96301846e-02 -5.70585054e-02 -5.45997987e-04  9.42995174e-02
-1.40422433e-01 -5.03571812e-04 -2.50305216e-01  3.79384141e-02
-6.44086637e-02 -1.53146188e-02 -2.55858274e-02 -1.10195376e-01
 1.62183899e-02 -1.61929591e-02  2.03421993e-02  1.21424534e-01
 5.02740463e-02  2.37900799e-02  9.07398322e-02  1.57962685e-02
 3.73036075e-02 -8.14876248e-02  1.37349395e-01 -8.17880913e-02
 9.27907812e-02  6.76093826e-03 -5.22928389e-02  6.02994188e-02
 8.28096711e-03 -1.05344042e-01 -1.02705751e-01  2.45275938e-02
-1.18970611e-02  9.86759282e-02 -1.92870134e-02  9.71936577e-03
-1.40249490e-01  1.61314103e-01 -4.55344879e-02  2.21929812e-02
 9.54108215e-02 -1.25028370e-02  2.89625007e-02  1.65818081e-02
-2.34467852e-02 -7.88610081e-02  3.34242148e-03  4.43269879e-02
-4.08419376e-02  6.06990416e-02  2.33916564e-02 -1.02773899e-02
 9.21596550e-02  9.90483805e-02  7.50525638e-03 -4.07725570e-03
-6.93980047e-02 -3.50341946e-02 -8.79849597e-02 -4.10474223e-02
 4.55004698e-03  2.27073689e-01  1.37340472e-01  4.43856114e-02]
```

**2. Vectorizing project_title**

In [0]:
```
word2VecTitlesVectors = getWord2VecVectors(preProcessedProjectTitlesWit
houtStopWords);
```

In [0]:
```
print("Shape of Word2Vec vectorization matrix of project titles: {}, {}
```

```
".format(len(word2VecTitlesVectors), len(word2VecTitlesVectors[0])));
equalsBorder(70);
print("Sample title: ");
equalsBorder(70);
print(preProcessedProjectTitlesWithoutStopWords[0]);
equalsBorder(70);
print("Word2Vec vector of sample title: ");
equalsBorder(70);
print(word2VecTitlesVectors[0]);
```

```
Shape of Word2Vec vectorization matrix of project titles: 109248, 300
======================================================================
Sample title:
======================================================================
educational support english learners home
======================================================================
Word2Vec vector of sample title:
======================================================================
[-4.1285000e-02  4.4970000e-02  1.4283080e-01  1.9901860e-02
 -8.4519200e-02 -4.3207400e-01 -2.8496800e+00 -2.2953320e-01
  2.1736960e-01  3.4239600e-01 -7.5568200e-02  1.8077600e-01
  1.3998316e-01 -1.6401800e-01 -2.9812820e-01 -2.5030200e-01
  2.0420960e-01 -1.6882720e-01  6.5439800e-02 -1.6061000e-01
  2.2179020e-01  2.9944900e-01  2.7358000e-02 -8.8528800e-02
  1.5856400e-01  6.2905000e-02  2.0427440e-01 -1.9312560e-01
 -9.2904600e-02 -2.2050020e-01 -5.7761060e-01 -1.2101294e-01
  1.6846980e-01  2.8212460e-01 -1.8210120e-01  1.7754000e-02
  1.4805200e-01  4.1059000e-02  3.1145000e-02 -9.5658000e-02
 -9.6840000e-03  2.4896520e-01 -2.5047440e-01  7.7859000e-02
 -3.7512000e-03 -2.7071920e-01  2.5586200e-02  2.3205600e-01
  1.0154800e-01 -5.2259200e-01 -1.3211440e-01  1.1908300e-01
  2.7147196e-01  5.6135400e-02 -5.3140200e-02 -1.4937160e-01
 -1.0488160e-01  1.2059600e-01 -1.2639620e-01 -1.4316640e-01
 -2.2147600e-01 -1.9137800e-01  1.6595340e-01 -5.6078000e-02
  3.9884400e-02  1.0854760e-01  1.5552920e-01  7.8204600e-02
  9.5928000e-02 -6.2156000e-03 -1.1407312e-01  3.6862800e-02
 -8.7530020e-02 -4.7668000e-02 -2.3264200e-01 -6.1687200e-02
 -3.1690916e-01 -1.1851380e-01  1.4931240e-01 -7.7857200e-02
  1.8634840e-01 -4.6202100e-01  2.7096800e-01 -3.0512800e-02
 -2.1226400e-01 -1.5356200e-02  1.0844260e-01 -8.2669200e-02
```

```
 2.8918600e-01  1.3372960e-01 -8.3522800e-02  4.6474200e-02
 2.0703580e-01 -2.1937640e-01 -1.0252400e-01 -2.5177000e-01
-2.8408000e+00  1.6622880e-01  1.1216234e-01  2.0837920e-01
-1.5711600e-01 -1.9159400e-01 -1.4992160e-01 -2.7392820e-01
 3.4989140e-01  1.3991600e-01  1.6275200e-01  1.3887200e-01
 1.8212760e-01 -3.2218600e-02  4.3172000e-02  1.8323640e-01
 1.2295780e-01  4.4706600e-01  2.1688400e-02 -3.8988200e-02
-3.2467400e-01  3.8389160e-01 -1.4416560e-01  1.1117380e-01
-1.6218300e-01  1.3871928e-01  1.4305240e-01 -7.6173200e-02
 8.9476800e-02  2.6043820e-01  5.1114000e-02  1.0619800e-01
 1.5968840e-01  1.0530680e-01  8.6300000e-02  1.4667260e-01
 1.2320460e-02 -6.6124620e-02 -1.1017760e-01 -1.5091940e-01
 2.1297280e-01 -3.2808520e-01  1.4493194e-01  2.1848680e-01
-4.1809800e-03  8.5340000e-02 -1.2410789e-01 -2.2308140e-01
 8.8026000e-02  1.9555000e-01 -3.7981400e-02 -1.7720080e-01
 3.4328600e-01 -3.7459600e-01 -1.7268200e-01 -2.1554400e-01
-1.1533400e-01  9.9680000e-02 -1.9032980e-01  8.6249800e-02
 7.6682200e-02 -9.1090380e-02 -9.3714000e-02 -1.7333260e-01
 8.6429960e-02 -6.7933600e-02 -8.6470600e-02 -2.2431600e-01
-2.8319800e-01  1.0138200e-01 -2.8114320e-01 -1.1168240e-01
 2.1770560e-02 -1.3971160e-01  2.1795080e-01 -1.1995600e-01
-1.3166600e-02 -3.4848260e-01 -3.0102000e-02  2.3396200e-02
 2.8840000e-02  2.8763000e-01 -2.3679600e-02  1.1806440e-01
-3.2261460e-01  2.2622920e-01  1.9506400e-02  1.4363200e-01
-1.3668380e-01 -1.0521880e-01 -3.9385400e-03 -4.6388000e-02
-7.7493780e-02 -2.4700800e-02 -5.2006200e-02 -2.6299360e-01
-2.5607520e-01  2.1704520e-01  5.6336000e-02 -6.3474400e-02
-1.0400400e-01 -1.7901000e-01  2.0326180e-01 -2.8708740e-01
 1.0132000e-01 -1.6278080e-01  1.2441440e-01  3.2699820e-01
-4.8321600e-02 -3.6052800e-02  2.2539620e-01 -8.2764000e-03
 3.1087258e-01  2.4090500e-01 -9.9590000e-02  1.2362460e-01
 1.7440000e-03 -1.6117280e-01  7.4570000e-02  3.1281120e-02
-1.1758000e-02 -1.8464800e-02 -2.0872020e-01 -3.9510000e-03
-5.7714400e-01 -1.8090080e-01 -2.8288200e-01 -2.4662120e-01
-1.8806540e+00  4.4765400e-01 -2.9412700e-01 -1.7280000e-02
-3.1931600e-01 -1.9190500e-01 -1.1642000e-02  1.7475600e-01
 1.3068840e-01  1.1943000e-01 -1.7219524e-01  1.9224000e-02
 2.2620000e-01 -1.0821980e-01  1.3789060e-01  2.6989320e-01
-2.4364960e-01 -1.3650800e-01 -3.0984180e-01 -3.9546200e-02
```

```
-1.1410800e-01 -6.6744640e-02  1.6330620e-01 -4.0601000e-01
 9.3793000e-02 -8.3026800e-02  9.0567600e-02  3.1595600e-01
 1.6786620e-01  1.0099860e-01  3.5043600e-02  6.6221200e-02
-3.5907800e-02 -2.4589760e-01  2.6006800e-01 -8.0637000e-02
 1.5359624e-01 -1.1078680e-01 -5.6956400e-02  2.2253080e-01
 3.5808000e-02 -1.8873860e-01 -2.5032660e-01  3.6167400e-02
-2.2424700e-01  2.7863640e-01  2.2622600e-02  1.3753300e-01
-2.3369620e-01  2.8058040e-01  5.0818000e-02 -3.4805800e-02
 1.7916600e-01 -7.5374000e-02  7.1228900e-02  1.7556000e-01
-5.8004120e-01 -2.0522500e-01 -1.3367960e-01  1.3656000e-02
-2.9052200e-02  1.3698600e-02  1.1746340e-01 -2.3288400e-02
 2.7706200e-01  1.6106000e-01 -2.0183340e-01  5.7781800e-02
-2.0954400e-01 -1.4111260e-02 -3.1186860e-01 -2.9536360e-02
-1.7226500e-01  3.5709400e-01  2.9448200e-01  8.5600000e-05]
```

## Tf-Idf Weighted Word2Vec Vectorization

### 1. Vectorizing project_essay

In [0]:
```
# Initializing tfidf vectorizer
tfIdfEssayTempVectorizer = TfidfVectorizer();
# Vectorizing preprocessed essays using tfidf vectorizer initialized ab
ove
tfIdfEssayTempVectorizer.fit(preProcessedEssaysWithoutStopWords);
# Saving dictionary in which each word is key and it's idf is value
tfIdfEssayDictionary = dict(zip(tfIdfEssayTempVectorizer.get_feature_na
mes(), list(tfIdfEssayTempVectorizer.idf_)));
# Creating set of all unique words used by tfidf vectorizer
tfIdfEssayWords = set(tfIdfEssayTempVectorizer.get_feature_names());
```

In [0]:
```
# Creating list to save tf-idf weighted vectors of essays
tfIdfWeightedWord2VecEssaysVectors = [];
# Iterating over each essay
for essay in tqdm(preProcessedEssaysWithoutStopWords):
    # Sum of tf-idf values of all words in a particular essay
    cumulativeSumTfIdfWeightOfEssay = 0;
```

```
    # Tf-Idf weighted word2vec vector of a particular essay
    tfIdfWeightedWord2VecEssayVector = np.zeros(300);
    # Splitting essay into list of words
    splittedEssay = essay.split();
    # Iterating over each word
    for word in splittedEssay:
        # Checking if word is in glove words and set of words used by t
fIdf essay vectorizer
        if (word in gloveWords) and (word in tfIdfEssayWords):
            # Tf-Idf value of particular word in essay
            tfIdfValueWord = tfIdfEssayDictionary[word] * (essay.count(
word) / len(splittedEssay));
            # Making tf-idf weighted word2vec
            tfIdfWeightedWord2VecEssayVector += tfIdfValueWord * gloveM
odel[word];
            # Summing tf-idf weight of word to cumulative sum
            cumulativeSumTfIdfWeightOfEssay += tfIdfValueWord;
    if cumulativeSumTfIdfWeightOfEssay != 0:
        # Taking average of sum of vectors with tf-idf cumulative sum
        tfIdfWeightedWord2VecEssayVector = tfIdfWeightedWord2VecEssayVe
ctor / cumulativeSumTfIdfWeightOfEssay;
    # Appending the above calculated tf-idf weighted vector of particul
ar essay to list of vectors of essays
    tfIdfWeightedWord2VecEssaysVectors.append(tfIdfWeightedWord2VecEssa
yVector);
```

In [0]:

```
print("Shape of Tf-Idf weighted Word2Vec vectorization matrix of projec
t essays: {}, {}".format(len(tfIdfWeightedWord2VecEssaysVectors), len(t
fIdfWeightedWord2VecEssaysVectors[0])));
equalsBorder(70);
print("Sample Essay: ");
equalsBorder(70);
print(preProcessedEssaysWithoutStopWords[0]);
equalsBorder(70);
print("Tf-Idf Weighted Word2Vec vector of sample essay: ");
equalsBorder(70);
print(tfIdfWeightedWord2VecEssaysVectors[0]);
```

```
Shape of Tf-Idf weighted Word2Vec vectorization matrix of project essay
s: 109248, 300
==========================================================================
Sample Essay:
==========================================================================
students english learners working english second third languages meltin
g pot refugees immigrants native born americans bringing gift language
school 24 languages represented english learner program students every
level mastery also 40 countries represented families within school stud
ent brings wealth knowledge experiences us open eyes new cultures belie
fs respect limits language limits world ludwig wittgenstein english lea
rner strong support system home begs resources many times parents learn
ing read speak english along side children sometimes creates barriers p
arents able help child learn phonetics letter recognition reading skill
s providing dvd players students able continue mastery english language
even no one home able assist families students within level 1 proficien
cy status offered part program educational videos specially chosen engl
ish learner teacher sent home regularly watch videos help child develop
early reading skills parents not access dvd player opportunity check dv
d player use year plan use videos educational dvd years come el student
s nannan
==========================================================================
Tf-Idf Weighted Word2Vec vector of sample essay:
==========================================================================
[-5.37582850e-02  7.68689598e-02  7.85741822e-02  4.38958976e-02
 -8.56874440e-02 -1.20832331e-01 -2.68120986e+00  7.17018732e-02
  1.03799206e-04 -5.17255299e-03 -2.67529751e-02  7.40185988e-02
  1.36881934e-01 -8.62706493e-02 -6.35020145e-02 -8.44084597e-02
  1.27523921e-01 -1.77105602e-01  3.68451284e-02 -5.74471880e-02
  1.86477259e-01  9.28786009e-02 -9.73137896e-02 -1.15230456e-02
  4.41962185e-02 -9.32894883e-02  1.11912943e-01 -1.17540961e-01
 -1.22150893e-01 -9.14028838e-02 -1.73918944e-01 -4.54143189e-02
 -7.82036060e-02  3.05617633e-01 -8.71850266e-02  6.31466708e-03
  1.15683161e-01  1.71477594e-02 -5.52983597e-02  9.08989585e-02
 -3.89808292e-04  1.97696142e-01 -4.08078376e-01 -5.39990199e-02
 -1.20129600e-02 -1.12456389e-01  2.92046345e-02  1.37924729e-01
  2.83465620e-02 -2.26817169e-01 -2.29639267e-02  6.94257143e-03
  5.80535394e-02  2.86454339e-02 -7.51508216e-02 -6.21569354e-02
 -1.41805544e-01  2.78707358e-02 -1.63165999e-01 -1.29716251e-01
```

```
-5.67625355e-02 -8.59507500e-02  3.54019902e-01 -4.96274469e-02
-6.88414062e-02  1.58623510e-01  1.24798600e-01  4.29711440e-02
 7.82814323e-02 -1.73260116e-02 -1.23679491e-01  1.47617250e-01
 4.27083617e-02 -1.16531047e-01 -1.27122530e-01 -5.93638332e-03
-1.99224414e-01 -8.66160391e-02  2.47701354e-01  1.61218205e-02
 3.56880345e-02 -3.71320273e-01  2.65501745e-01 -4.56454865e-02
-7.85433814e-02 -5.99177835e-02  4.42212779e-02 -8.20739267e-02
 2.14031939e-01  2.42131497e-02 -1.34069697e-01  7.15871686e-03
 4.00667270e-02 -6.75881497e-02 -7.07967357e-02 -2.15984749e-02
-2.09734597e+00  1.02300477e-01  6.61169899e-02  5.70146517e-02
-1.91302495e-01 -1.38114014e-01 -1.10709961e-01 -1.66994098e-01
 9.17800823e-02  1.35327093e-01  2.20333244e-02 -3.83844831e-02
 2.57206511e-02 -5.54503565e-02 -3.41973653e-03  1.99777588e-02
 4.85050396e-02  2.13190534e-01  4.64281665e-02  6.51171751e-02
-5.80015838e-02  1.19900386e-01  1.18803830e-01  7.05550873e-02
-1.87330886e-01  1.41219129e-01  1.33569574e-01  1.00530000e-01
 4.14498415e-02  1.39860952e-01 -7.95709830e-02  9.70242332e-02
 1.07442882e-01  9.00794808e-02  7.47745032e-02  4.18772282e-02
-7.10347826e-03 -7.62379756e-03 -7.31715828e-02 -1.16370646e-01
 2.82271708e-01 -5.30885621e-02  4.51472249e-02  2.61376253e-01
 1.29080066e-02  3.96843846e-02  1.04430681e-01 -1.30495811e-01
-1.17999239e-01 -1.02810089e-01 -6.52713784e-02 -1.81350799e-01
 1.55415740e-01 -4.43517889e-02 -8.34350788e-02 -1.31445407e-01
-8.87524029e-02 -1.15321245e-02  8.67587067e-03  3.55646447e-02
-4.32365925e-02  2.44285859e-03  2.73165854e-02 -1.91651165e-01
 6.70942750e-03  1.45533103e-02 -5.95191056e-02 -9.78336553e-02
-4.61200683e-02  1.04017495e-02 -1.68129330e-01 -5.53455289e-02
-1.95353920e-02 -3.24088827e-03  9.94121739e-02 -2.20584067e-02
 1.36190091e-02 -3.13014669e-01  4.46748268e-02  6.11251996e-02
-5.59088914e-02  8.07071841e-02 -7.80920682e-02  1.05535003e-02
-8.49705076e-02  1.87800458e-01 -5.53305425e-02 -4.05296946e-02
-1.68105655e-02 -9.64697267e-02 -1.00114054e-01 -1.25303984e-01
-6.77861115e-02  1.38106300e-02  4.97948787e-02 -1.04414463e-01
 3.12147536e-03 -2.46650333e-02  1.56250756e-02 -3.41987984e-02
 2.90197738e-02 -1.30795750e-01  1.71425098e-01 -1.33199913e-01
-4.35452619e-02 -1.52841321e-01  3.37717104e-02  2.11400042e-01
-1.08493100e-01  6.64905827e-02  4.45687503e-02 -3.38898797e-03
 1.47302984e-01  3.10931848e-02  6.94873935e-03 -3.79090162e-02
 3.97055902e-02 -3.12563998e-02  2.99815273e-02 -9.30892230e-03
```

```
 -3.37192802e-02 -7.79667288e-02  4.20509297e-02  4.33535394e-02
-2.38238094e-01 -4.11188300e-02 -1.93930088e-01  1.15012485e-01
-2.14605373e+00  1.36975648e-01 -1.79026305e-01 -1.42630498e-01
-1.37558424e-01 -1.55433436e-01 -6.96701214e-02  1.05328488e-01
 3.43486342e-02 -2.37676310e-03 -6.80980842e-02 -1.92470331e-01
 1.54727348e-01 -7.47455695e-02 -1.58054203e-02  3.33369549e-02
-1.70510752e-01 -5.74331307e-02 -2.38994456e-01  5.64188931e-02
-8.55051184e-02 -5.52984572e-02 -5.00408589e-02 -6.81572658e-02
 5.15848477e-03 -3.58487773e-02  7.00056842e-02  1.33127170e-01
 5.57938159e-02  1.03106840e-01  4.18598320e-02 -2.78162076e-03
 8.83131944e-02 -1.31482831e-01  1.34875022e-01 -8.31772344e-02
 1.62319378e-01  9.25839856e-02 -7.07548194e-02  1.74355644e-01
 1.53106818e-02 -1.74504449e-01 -5.39158255e-02 -1.16968555e-02
-1.37824311e-01  1.07713713e-01  4.48548015e-02  1.07272158e-01
-1.59084558e-01  1.94342786e-01 -4.73514319e-02 -4.87250503e-02
 2.82023483e-02 -4.18474756e-02  8.04397595e-02 -3.34005484e-02
-1.00808502e-01 -1.15380334e-01  7.05894205e-02  2.92052920e-02
-5.72604859e-02 -7.39274088e-03  1.44106517e-02 -2.64282237e-02
 2.31512689e-01  1.50161666e-01 -5.21462274e-02 -1.00796916e-02
-4.47392305e-02  4.83958092e-02 -2.21927272e-01 -9.69846899e-02
-5.91211767e-03  2.52508756e-01  1.08677704e-01  5.05047869e-02]
```

**2. Vectorizing project_title**

In [0]:
```
# Initializing tfidf vectorizer
tfIdfTitleTempVectorizer = TfidfVectorizer();
# Vectorizing preprocessed titles using tfidf vectorizer initialized ab
ove
tfIdfTitleTempVectorizer.fit(preProcessedProjectTitlesWithoutStopWords
);
# Saving dictionary in which each word is key and it's idf is value
tfIdfTitleDictionary = dict(zip(tfIdfTitleTempVectorizer.get_feature_na
mes(), list(tfIdfTitleTempVectorizer.idf_)));
# Creating set of all unique words used by tfidf vectorizer
tfIdfTitleWords = set(tfIdfTitleTempVectorizer.get_feature_names());
```

In [0]:
```
# Creating list to save tf-idf weighted vectors of project titles
```

```
tfIdfWeightedWord2VecTitlesVectors = [];
# Iterating over each title
for title in tqdm(preProcessedProjectTitlesWithoutStopWords):
    # Sum of tf-idf values of all words in a particular project title
    cumulativeSumTfIdfWeightOfTitle = 0;
    # Tf-Idf weighted word2vec vector of a particular project title
    tfIdfWeightedWord2VecTitleVector = np.zeros(300);
    # Splitting title into list of words
    splittedTitle = title.split();
    # Iterating over each word
    for word in splittedTitle:
        # Checking if word is in glove words and set of words used by t
fIdf title vectorizer
        if (word in gloveWords) and (word in tfIdfTitleWords):
            # Tf-Idf value of particular word in title
            tfIdfValueWord = tfIdfTitleDictionary[word] * (title.count(
word) / len(splittedTitle));
            # Making tf-idf weighted word2vec
            tfIdfWeightedWord2VecTitleVector += tfIdfValueWord * gloveM
odel[word];
            # Summing tf-idf weight of word to cumulative sum
            cumulativeSumTfIdfWeightOfTitle += tfIdfValueWord;
    if cumulativeSumTfIdfWeightOfTitle != 0:
        # Taking average of sum of vectors with tf-idf cumulative sum
        tfIdfWeightedWord2VecTitleVector = tfIdfWeightedWord2VecTitleVe
ctor / cumulativeSumTfIdfWeightOfTitle;
    # Appending the above calculated tf-idf weighted vector of particul
ar title to list of vectors of project titles
    tfIdfWeightedWord2VecTitlesVectors.append(tfIdfWeightedWord2VecTitl
eVector);
```

In [0]:

```
print("Shape of Tf-Idf weighted Word2Vec vectorization matrix of projec
t titles: {}, {}".format(len(tfIdfWeightedWord2VecTitlesVectors), len(t
fIdfWeightedWord2VecTitlesVectors[0])));
equalsBorder(70);
print("Sample Title: ");
equalsBorder(70);
print(preProcessedProjectTitlesWithoutStopWords[0]);
```

```
equalsBorder(70);
print("Tf-Idf Weighted Word2Vec vector of sample title: ");
equalsBorder(70);
print(tfIdfWeightedWord2VecTitlesVectors[0]);
```

```
Shape of Tf-Idf weighted Word2Vec vectorization matrix of project title
s: 109248, 300
======================================================================
Sample Title:
======================================================================
educational support english learners home
======================================================================
Tf-Idf Weighted Word2Vec vector of sample title:
======================================================================
[-3.23904891e-02  5.58064810e-02  1.32666911e-01  3.84227573e-02
 -6.71984492e-02 -4.30940397e-01 -2.84607947e+00 -2.45905055e-01
  1.96794858e-01  3.19604663e-01 -6.12568872e-02  1.59218099e-01
  1.25129027e-01 -1.67580327e-01 -2.82644062e-01 -2.47555536e-01
  2.18304104e-01 -1.57431101e-01  7.66481545e-02 -1.61436633e-01
  2.38451267e-01  2.86712258e-01  2.70730890e-02 -9.74962294e-02
  1.67511144e-01  7.18131102e-02  1.82846112e-01 -1.96778087e-01
 -8.19948978e-02 -2.25877630e-01 -5.54573752e-01 -1.28462870e-01
  1.61012606e-01  2.94412658e-01 -1.63196910e-01 -1.23217523e-02
  1.37466355e-01  4.45437696e-02  4.65691769e-02 -1.17867965e-01
 -2.41502151e-03  2.24350668e-01 -2.51274676e-01  8.29431360e-02
 -1.65996673e-02 -2.47747576e-01  1.45110611e-03  2.37117949e-01
  9.71345150e-02 -5.13516477e-01 -1.40296688e-01  1.42775548e-01
  2.89949805e-01  6.49771690e-02 -3.41581088e-02 -1.58076306e-01
 -1.07731741e-01  7.59015357e-02 -1.21511682e-01 -1.16519972e-01
 -2.27321940e-01 -1.63525257e-01  1.80860125e-01 -4.17314689e-02
  4.60171896e-02  1.00024674e-01  1.54588362e-01  8.25394911e-02
  7.45768118e-02 -1.80240543e-02 -1.22956246e-01 -4.97450371e-03
 -8.06577406e-02 -5.00614538e-02 -2.15836210e-01 -5.89271531e-02
 -3.26363335e-01 -1.32706775e-01  1.61236199e-01 -1.25038790e-01
  1.96493846e-01 -4.95095193e-01  2.34765396e-01 -4.44646606e-02
 -2.04266125e-01 -3.21415735e-02  8.48111983e-02 -7.27603472e-02
  2.79183660e-01  1.18968262e-01 -7.43300594e-02  6.34587771e-02
  1.99863053e-01 -2.13382053e-01 -1.01221319e-01 -2.49884070e-01
 -2.92249478e+00  1.60273141e-01  7.74579728e-02  1.85323805e-01
 -1.33255909e-01 -2.00013519e-01 -1.31974722e-01 -2.62288530e-01
```

```
 3.54852941e-01  1.18537924e-01  1.62207829e-01  1.24436802e-01
 1.98867481e-01 -4.87526944e-03  3.00886908e-02  2.09330567e-01
 1.17189984e-01  3.94887340e-01  2.52941492e-02 -5.13348554e-02
-2.91140828e-01  4.06939567e-01 -1.70319175e-01  1.17651155e-01
-1.66813086e-01  1.53049826e-01  1.41255472e-01 -8.10785736e-02
 9.57549943e-02  2.73610111e-01  5.85622995e-02  7.91410001e-02
 1.47619459e-01  9.75521835e-02  6.74487028e-02  1.53125504e-01
 2.02791106e-02 -5.59403852e-02 -1.02109913e-01 -1.22913427e-01
 1.99873969e-01 -3.21872719e-01  1.38343165e-01  2.17196179e-01
 4.95201760e-03  8.52128333e-02 -1.45880901e-01 -2.10862397e-01
 1.20343357e-01  2.15598061e-01 -1.14038072e-02 -1.72172799e-01
 3.24157324e-01 -3.82818101e-01 -1.87580283e-01 -2.00827204e-01
-1.41863370e-01  9.63016678e-02 -2.01659119e-01  6.74342164e-02
 7.12185747e-02 -1.04314039e-01 -9.08169483e-02 -1.63495605e-01
 9.68230169e-02 -5.01176209e-02 -8.34015616e-02 -1.88998660e-01
-2.84065057e-01  1.16975197e-01 -2.80836800e-01 -9.33191327e-02
 3.79583269e-02 -1.22755412e-01  2.30408258e-01 -1.31968890e-01
 9.72824714e-03 -3.44272546e-01 -2.09522211e-03  2.45944018e-02
 2.94077607e-02  2.67568157e-01 -2.69460269e-02  1.25412311e-01
-3.47031083e-01  2.09328241e-01  1.25385338e-02  1.55654760e-01
-1.41368915e-01 -1.01749781e-01 -4.77312036e-04 -4.82325465e-02
-7.15727478e-02 -3.63658602e-02 -4.33504397e-02 -2.71410315e-01
-2.40079853e-01  2.01171435e-01  6.39005674e-02 -4.86787485e-02
-1.48623863e-01 -1.72130906e-01  1.97761227e-01 -3.13043504e-01
 1.07772898e-01 -1.54518908e-01  1.31855435e-01  3.39703669e-01
-4.51652340e-02 -4.05998340e-02  2.03610454e-01  8.84982054e-03
 3.05974297e-01  2.54736700e-01 -1.06925907e-01  1.27066655e-01
-1.88835779e-02 -1.56632041e-01  8.45142200e-02  5.70681135e-02
 1.01119358e-02 -6.62387316e-03 -2.18552410e-01  1.20985419e-02
-5.54006219e-01 -1.72367117e-01 -2.90325016e-01 -2.34816399e-01
-1.94243114e+00  4.36715446e-01 -2.80713863e-01 -6.33991309e-03
-2.90035778e-01 -1.98732349e-01  2.96737137e-02  1.50873684e-01
 1.16943997e-01  1.39741722e-01 -1.82238609e-01  4.09714520e-02
 2.37176600e-01 -1.24515116e-01  1.41648743e-01  2.64206287e-01
-2.40551078e-01 -1.40415333e-01 -2.92432371e-01 -3.03761027e-02
-9.90320454e-02 -8.43648662e-02  1.81116706e-01 -4.05719699e-01
 1.22898740e-01 -8.80109292e-02  1.09543672e-01  2.96110858e-01
 1.85027885e-01  9.14976115e-02  9.63416424e-03  5.50340717e-02
-2.59328007e-02 -2.43942768e-01  2.54260096e-01 -1.03280950e-01
```

```
 1.56799018e-01 -9.58635926e-02 -4.31948365e-02  2.01228907e-01
 5.20033765e-02 -2.08030399e-01 -2.49149283e-01  3.11752465e-02
-2.39410711e-01  2.54421815e-01  3.50420005e-02  1.31625993e-01
-2.19027956e-01  2.75093693e-01  4.31276229e-02 -6.89266192e-02
 1.80694153e-01 -9.77254221e-02  6.52789959e-02  1.81468103e-01
-5.79288980e-01 -1.91501478e-01 -1.43298895e-01  1.56769073e-02
-2.28584041e-02 -7.96762354e-03  1.38764109e-01 -2.67804890e-02
 3.02808634e-01  1.63688874e-01 -1.98263925e-01  8.94007093e-02
-2.01132765e-01  8.29230669e-03 -3.17426319e-01 -4.07929287e-02
-1.63872993e-01  3.69860278e-01  2.90009047e-01  4.56005599e-02]
```

## Vectorizing numerical features

### 1. Vectorizing price

In [0]:
```
# Standardizing the price data using StandardScaler(Uses mean and std f
or standardization)
priceScaler = StandardScaler();
priceScaler.fit(projectsData['price'].values.reshape(-1, 1));
priceStandardized = priceScaler.transform(projectsData['price'].values.
reshape(-1, 1));
```

In [0]:
```
print("Shape of standardized matrix of prices: ", priceStandardized.sha
pe);
equalsBorder(70);
print("Sample original prices: ");
equalsBorder(70);
print(projectsData['price'].values[0:5]);
print("Sample standardized prices: ");
equalsBorder(70);
print(priceStandardized[0:5]);
```

```
Shape of standardized matrix of prices:  (109245, 1)
======================================================================
Sample original prices:
======================================================================
```

```
[154.6  299.   516.85 232.9   67.98]
Sample standardized prices:
======================================================================
[[-0.39052147]
 [ 0.00240752]
 [ 0.5952024 ]
 [-0.17745817]
 [-0.62622444]]
```

**2. Vectorizing quantity**

In [0]:
```
# Standardizing the quantity data using StandardScaler(Uses mean and st
d for standardization)
quantityScaler = StandardScaler();
quantityScaler.fit(projectsData['quantity'].values.reshape(-1, 1));
quantityStandardized = quantityScaler.transform(projectsData['quantity'
].values.reshape(-1, 1));
```

In [0]:
```
print("Shape of standardized matrix of quantities: ", quantityStandardi
zed.shape);
equalsBorder(70);
print("Sample original quantities: ");
equalsBorder(70);
print(projectsData['quantity'].values[0:5]);
print("Sample standardized quantities: ");
equalsBorder(70);
print(quantityStandardized[0:5]);
```

```
Shape of standardized matrix of quantities:  (109245, 1)
======================================================================
Sample original quantities:
======================================================================
[23  1 22  4  4]
Sample standardized quantities:
======================================================================
[[ 0.23045805]
 [-0.6097785 ]
 [ 0.19226548]
```

```
 [-0.49520079]
 [-0.49520079]]
```

### 3. Vectorizing teacher_number_of_previously_posted_projects

In [0]:
```
# Standardizing the teacher_number_of_previously_posted_projects data u
sing StandardScaler(Uses mean and std for standardization)
previouslyPostedScaler = StandardScaler();
previouslyPostedScaler.fit(projectsData['teacher_number_of_previously_p
osted_projects'].values.reshape(-1, 1));
previouslyPostedStandardized = previouslyPostedScaler.transform(project
sData['teacher_number_of_previously_posted_projects'].values.reshape(-1
, 1));
```

In [0]:
```
print("Shape of standardized matrix of teacher_number_of_previously_pos
ted_projects: ", previouslyPostedStandardized.shape);
equalsBorder(70);
print("Sample original quantities: ");
equalsBorder(70);
print(projectsData['teacher_number_of_previously_posted_projects'].valu
es[0:5]);
print("Sample standardized teacher_number_of_previously_posted_project
s: ");
equalsBorder(70);
print(previouslyPostedStandardized[0:5]);
```

```
Shape of standardized matrix of teacher_number_of_previously_posted_pro
jects:  (109245, 1)
======================================================================
Sample original quantities:
======================================================================
[0 7 1 4 1]
Sample standardized teacher_number_of_previously_posted_projects:
======================================================================
[[-0.40153083]
 [-0.14952695]
 [-0.36553028]
```

```
[-0.25752861]
[-0.36553028]]
```

## Taking 6k points(to avoid memory errors)

In [0]:
```
numberOfPoints = 6000;
# Categorical data
categoriesVectorsSub = categoriesVectors[0:numberOfPoints];
subCategoriesVectorsSub = subCategoriesVectors[0:numberOfPoints];
teacherPrefixVectorsSub = teacherPrefixVectors[0:numberOfPoints];
schoolStateVectorsSub = schoolStateVectors[0:numberOfPoints];
projectGradeVectorsSub = projectGradeVectors[0:numberOfPoints];

# Text data
bowEssayModelSub = bowEssayModel[0:numberOfPoints];
bowTitleModelSub = bowTitleModel[0:numberOfPoints];
tfIdfEssayModelSub = tfIdfEssayModel[0:numberOfPoints];
tfIdfTitleModelSub = tfIdfTitleModel[0:numberOfPoints];
word2VecEssaysVectorsSub = word2VecEssaysVectors[0:numberOfPoints];
word2VecTitlesVectorsSub = word2VecTitlesVectors[0:numberOfPoints];
tfIdfWeightedWord2VecEssaysVectorsSub = tfIdfWeightedWord2VecEssaysVect
ors[0:numberOfPoints];
tfIdfWeightedWord2VecTitlesVectorsSub = tfIdfWeightedWord2VecTitlesVect
ors[0:numberOfPoints];

# Numerical data
priceStandardizedSub = priceStandardized[0:numberOfPoints];
quantityStandardizedSub = quantityStandardized[0:numberOfPoints];
previouslyPostedStandardizedSub = previouslyPostedStandardized[0:number
OfPoints];
```

In [0]:
```
classesDataSub = projectsData['project_is_approved'][0:numberOfPoints].
values
```

In [0]:
```
classesDataSub.shape
```

Out[0]:
```
(6000,)
```

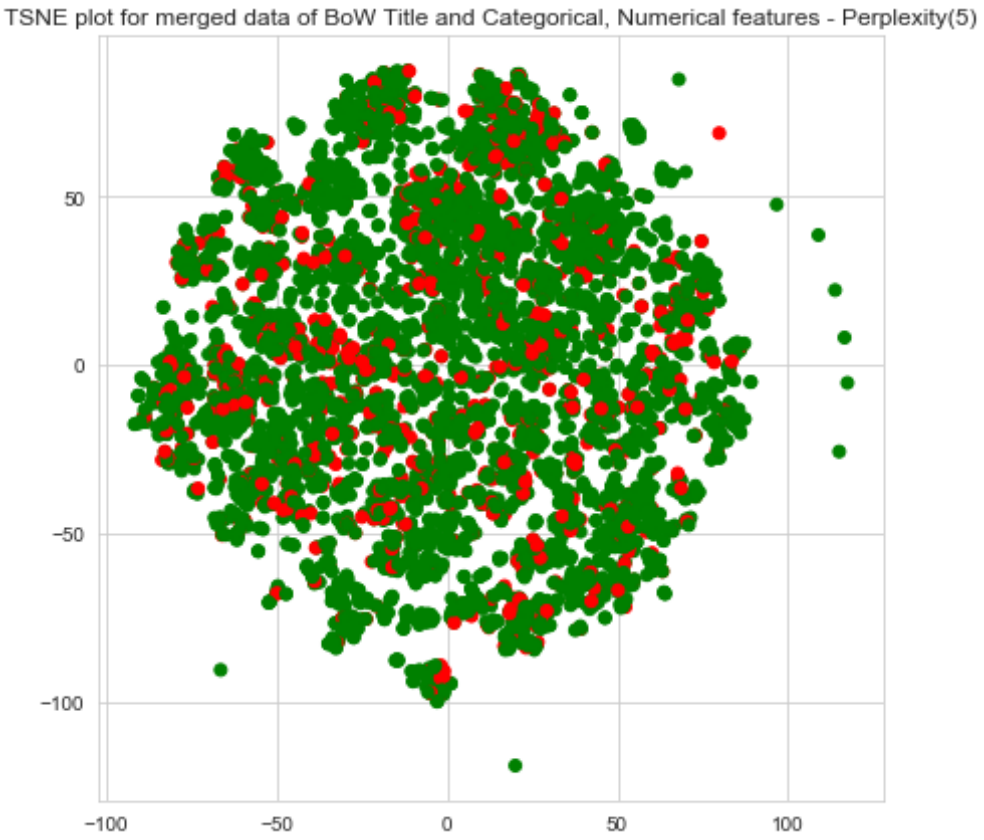
# Data Visualization using T-SNE

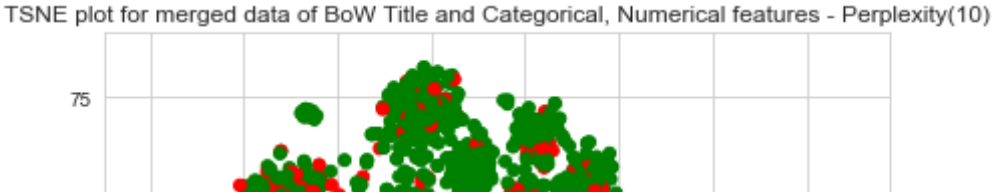
**Classification using data merged with bag of words vectorized title and all considered categorical, numerical features**

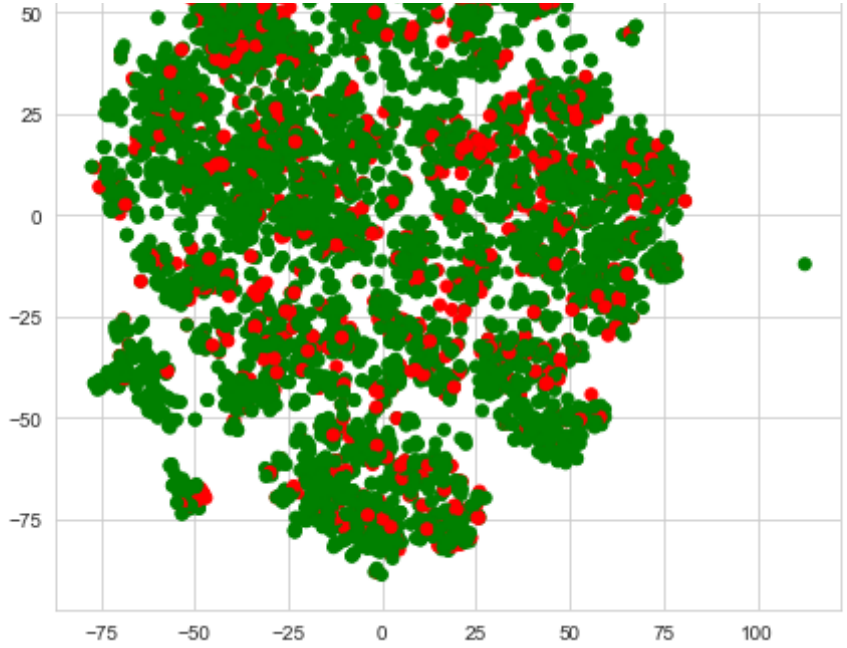
In [0]:
```
bowTitleAndOthers = hstack((bowTitleModelSub, categoriesVectorsSub, sub
CategoriesVectorsSub, teacherPrefixVectorsSub, schoolStateVectorsSub, p
rojectGradeVectorsSub, priceStandardizedSub, previouslyPostedStandardiz
edSub));
bowTitleAndOthers.shape
```

Out[0]:
```
(6000, 1875)
```

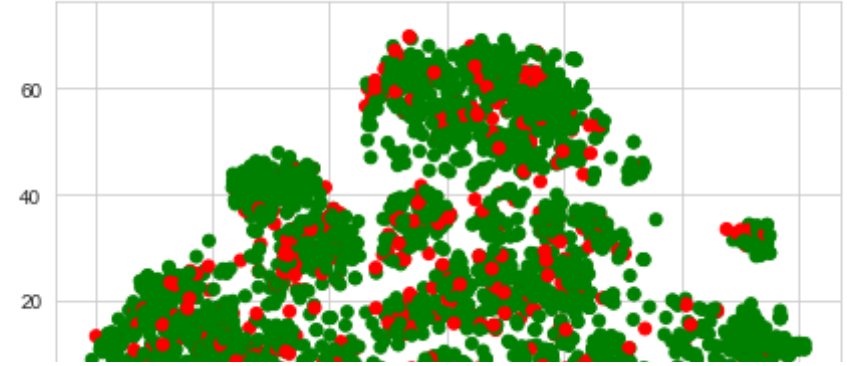
In [0]:
```
perplexityValues = [5, 10, 30, 50, 80, 100]
for perplexityValue in perplexityValues:
    tsne = TSNE(n_components = 2, perplexity = perplexityValue, learnin
g_rate = 200);
    bowTitleAndOthersEmbedded = tsne.fit_transform(bowTitleAndOthers.to
array());
    bowTitleAndOthersTsneData = np.hstack((bowTitleAndOthersEmbedded, c
lassesDataSub.reshape(-1, 1)));
    bowTitleAndOthersTsneDataFrame = pd.DataFrame(bowTitleAndOthersTsne
Data, columns = ['Dimension1', 'Dimension2', 'Class']);
    colors = {0.0:'red', 1.0:'green'}
    plt.title("TSNE plot for merged data of BoW Title and Categorical,
 Numerical features - Perplexity({})".format(perplexityValue));
    plt.scatter(bowTitleAndOthersTsneDataFrame['Dimension1'], bowTitleA
ndOthersTsneDataFrame['Dimension2'], c = bowTitleAndOthersTsneDataFrame
['Class'].apply(lambda x: colors[x]));
    plt.show();
```

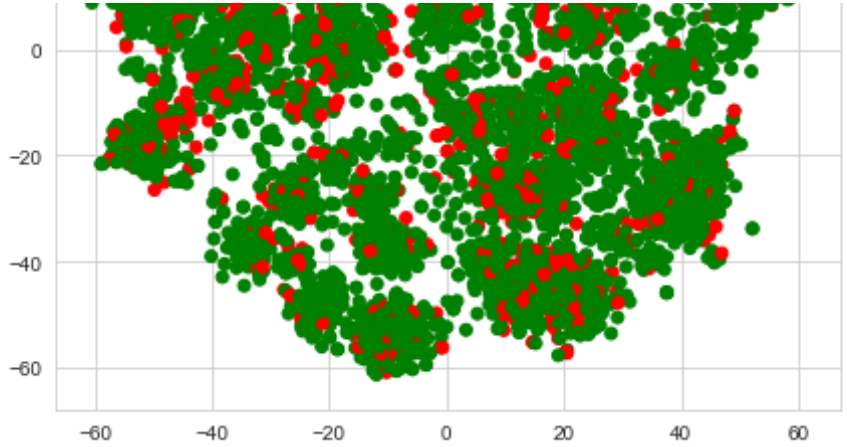


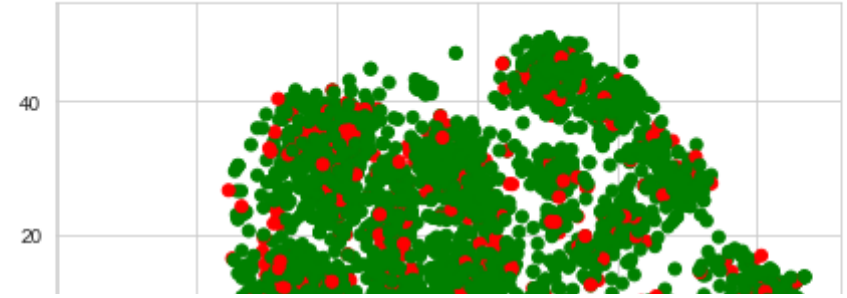


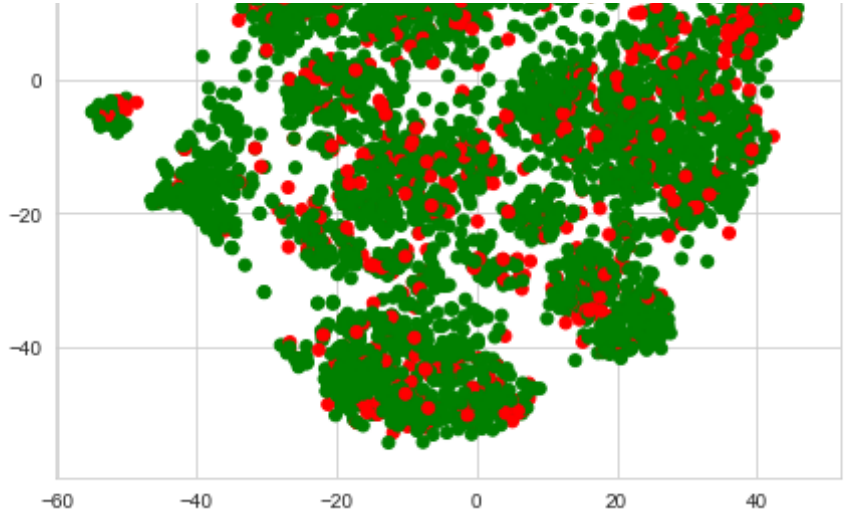

TSNE plot for merged data of BoW Title and Categorical, Numerical features - Perplexity(30)



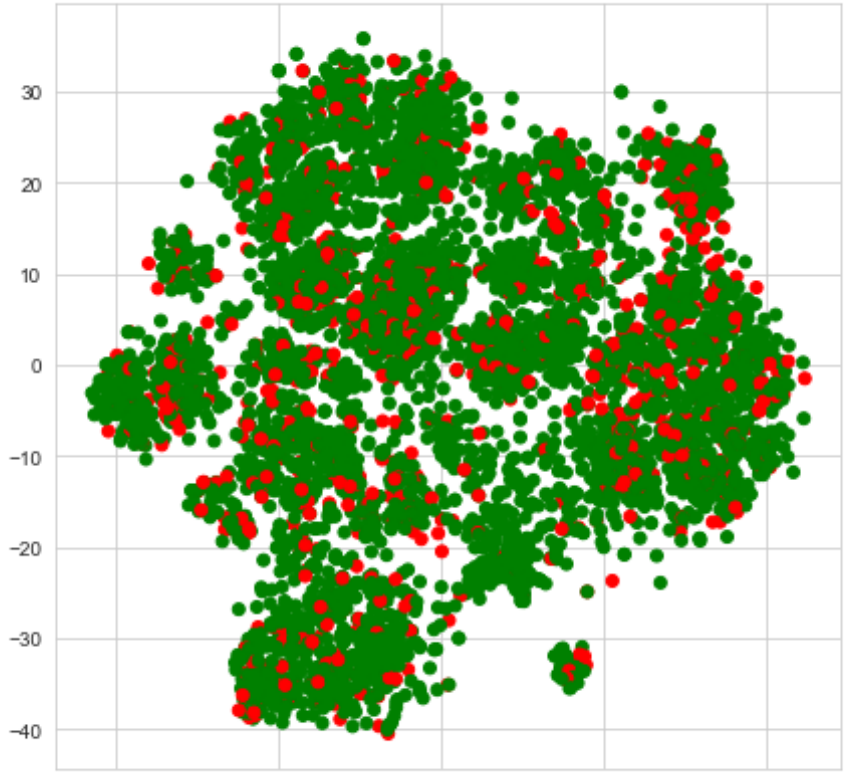


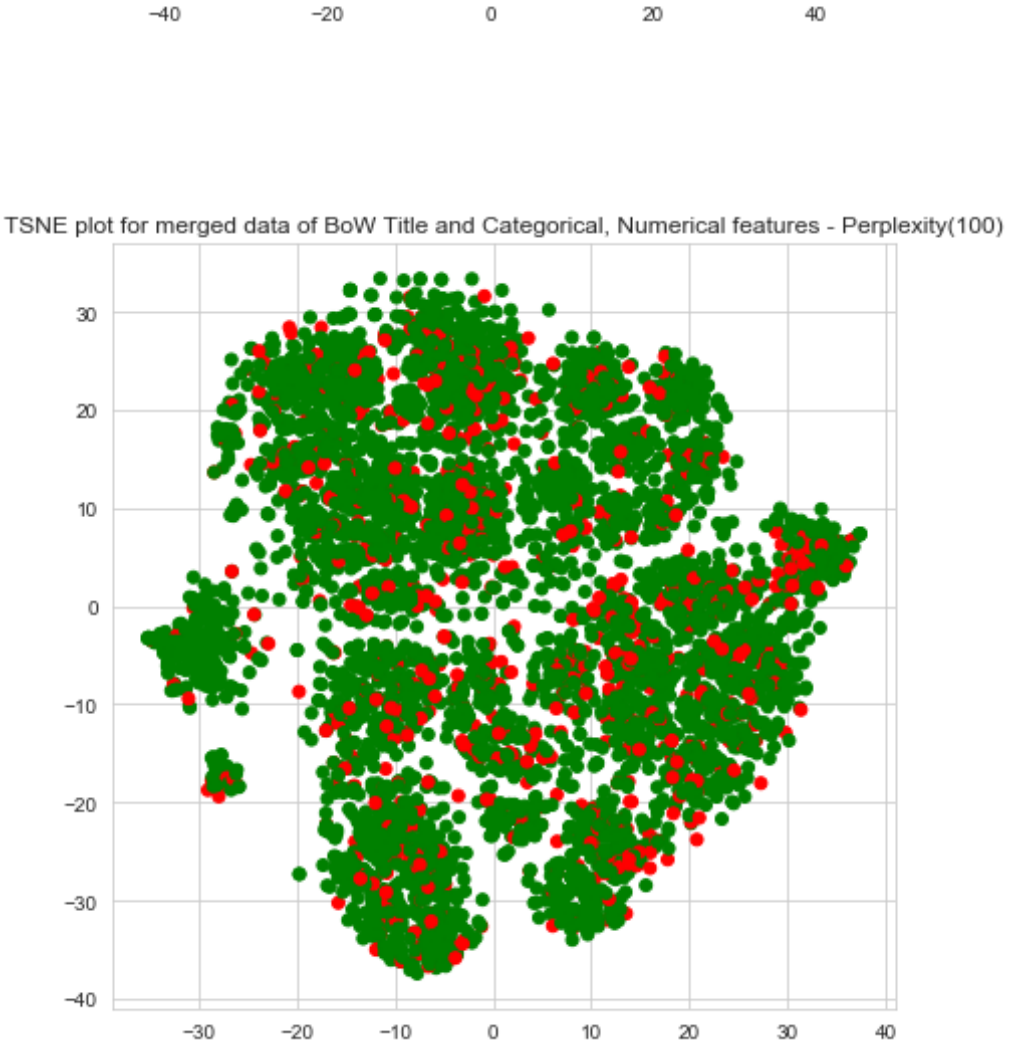

TSNE plot for merged data of BoW Title and Categorical, Numerical features - Perplexity(50)






TSNE plot for merged data of BoW Title and Categorical, Numerical features - Perplexity(80)






**Classification using data merged with Tf-Idf vectorized title and all considered categorical, numerical features**

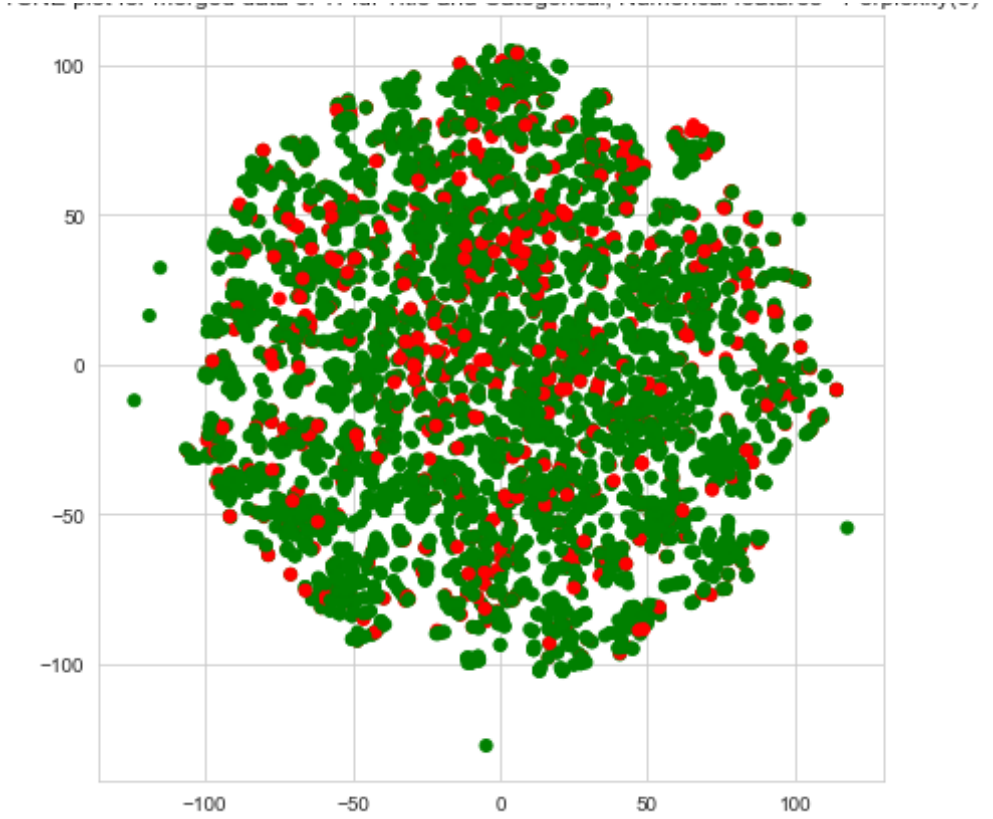
In [0]:
```
tfIdfTitleAndOthers = hstack((tfIdfTitleModelSub, categoriesVectorsSub,
 subCategoriesVectorsSub, teacherPrefixVectorsSub, schoolStateVectorsSu
b, projectGradeVectorsSub, priceStandardizedSub, previouslyPostedStanda
```

```
rdizedSub));
tfIdfTitleAndOthers.shape
```

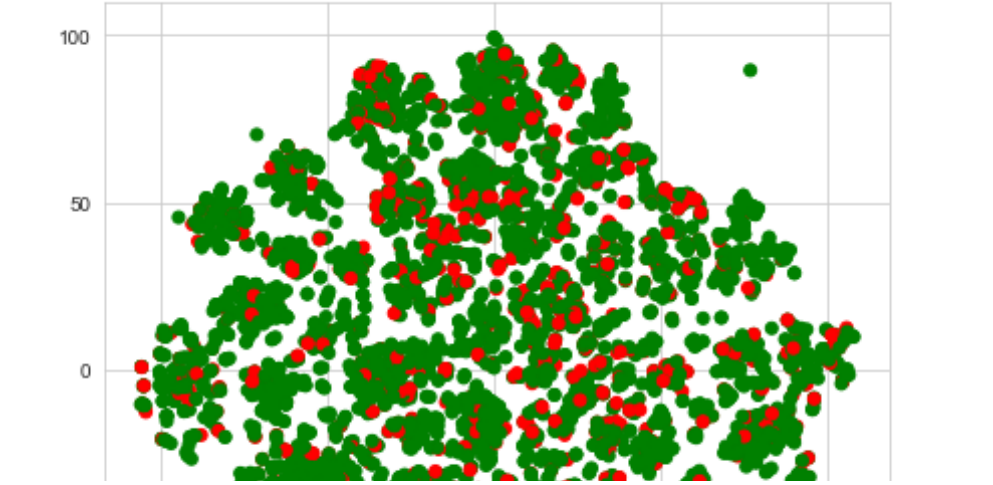
Out[0]: (6000, 1875)

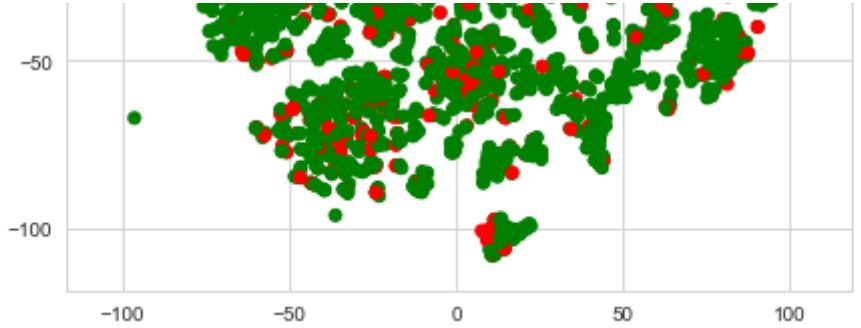
In [0]:
```
perplexityValues = [5, 10, 30, 50, 80, 100]
for perplexityValue in perplexityValues:
    tsne = TSNE(n_components = 2, perplexity = perplexityValue, learnin
g_rate = 200);
    tfIdfTitleAndOthersEmbedded = tsne.fit_transform(tfIdfTitleAndOther
s.toarray());
    tfIdfTitleAndOthersTsneData = np.hstack((tfIdfTitleAndOthersEmbedde
d, classesDataSub.reshape(-1, 1)));
    tfIdfTitleAndOthersTsneDataFrame = pd.DataFrame(tfIdfTitleAndOthers
TsneData, columns = ['Dimension1', 'Dimension2', 'Class']);
    colors = {0.0:'red', 1.0:'green'}
    plt.title("TSNE plot for merged data of Tf-Idf Title and Categorica
l, Numerical features - Perplexity({})".format(perplexityValue));
    plt.scatter(tfIdfTitleAndOthersTsneDataFrame['Dimension1'], tfIdfTi
tleAndOthersTsneDataFrame['Dimension2'], c = tfIdfTitleAndOthersTsneDat
aFrame['Class'].apply(lambda x: colors[x]));
    plt.show();
```

TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(5)

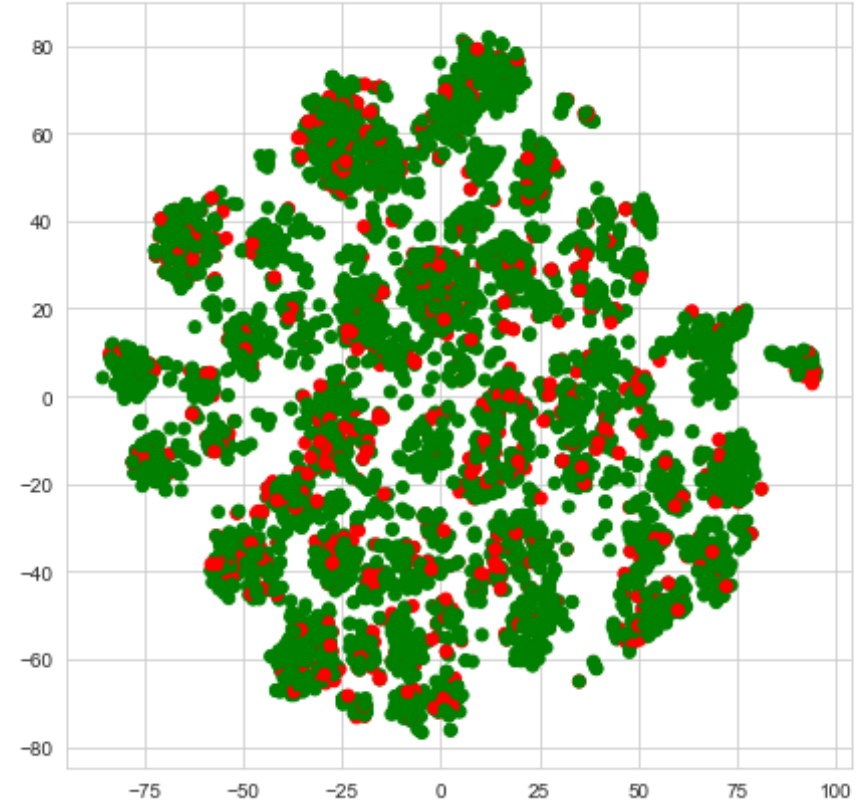



TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(10)

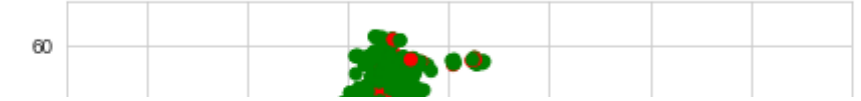


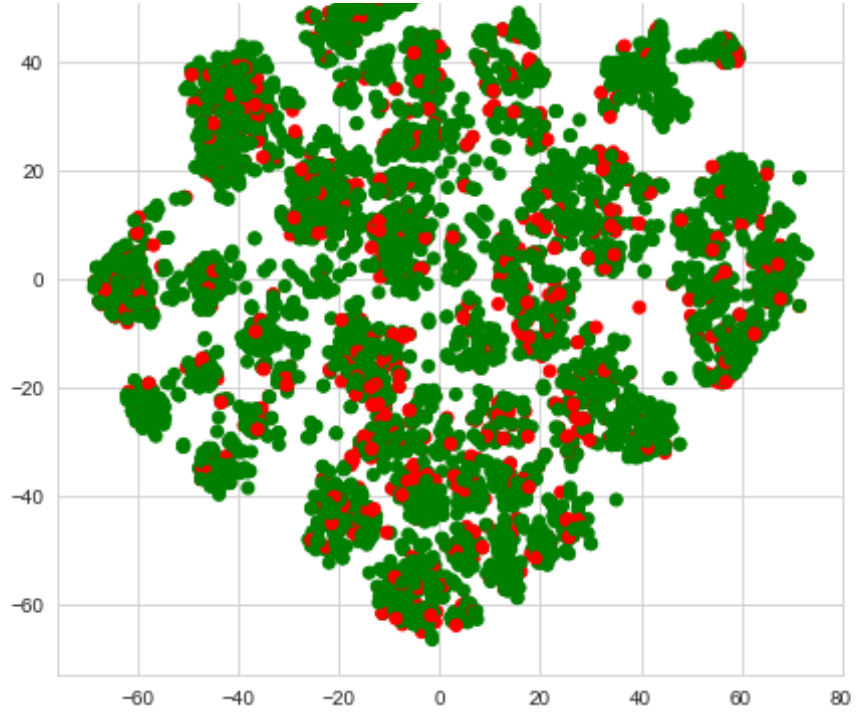



TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(30)

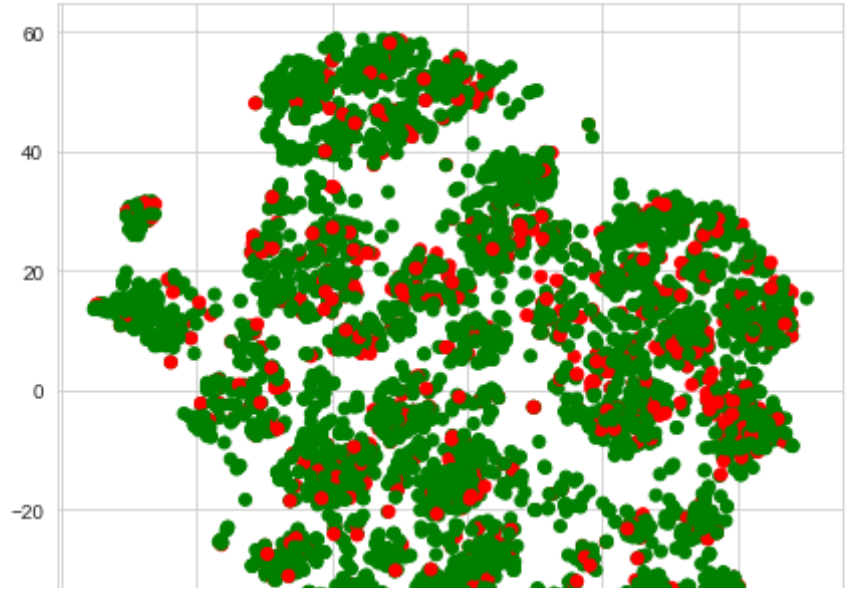


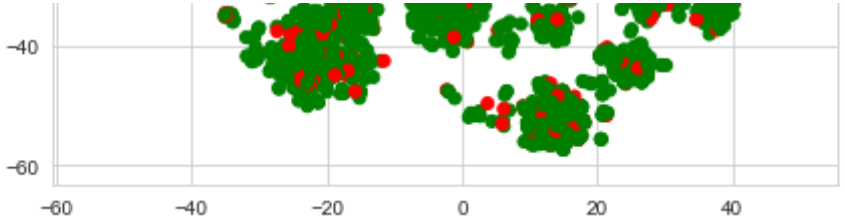

TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(50)



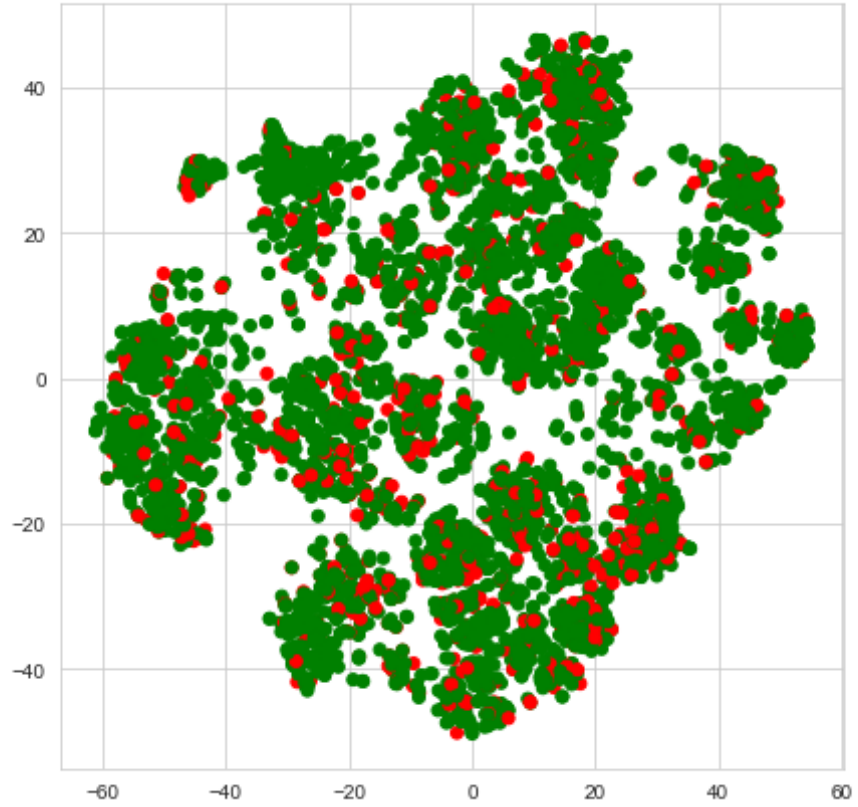



TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(80)






TSNE plot for merged data of Tf-Idf Title and Categorical, Numerical features - Perplexity(100)



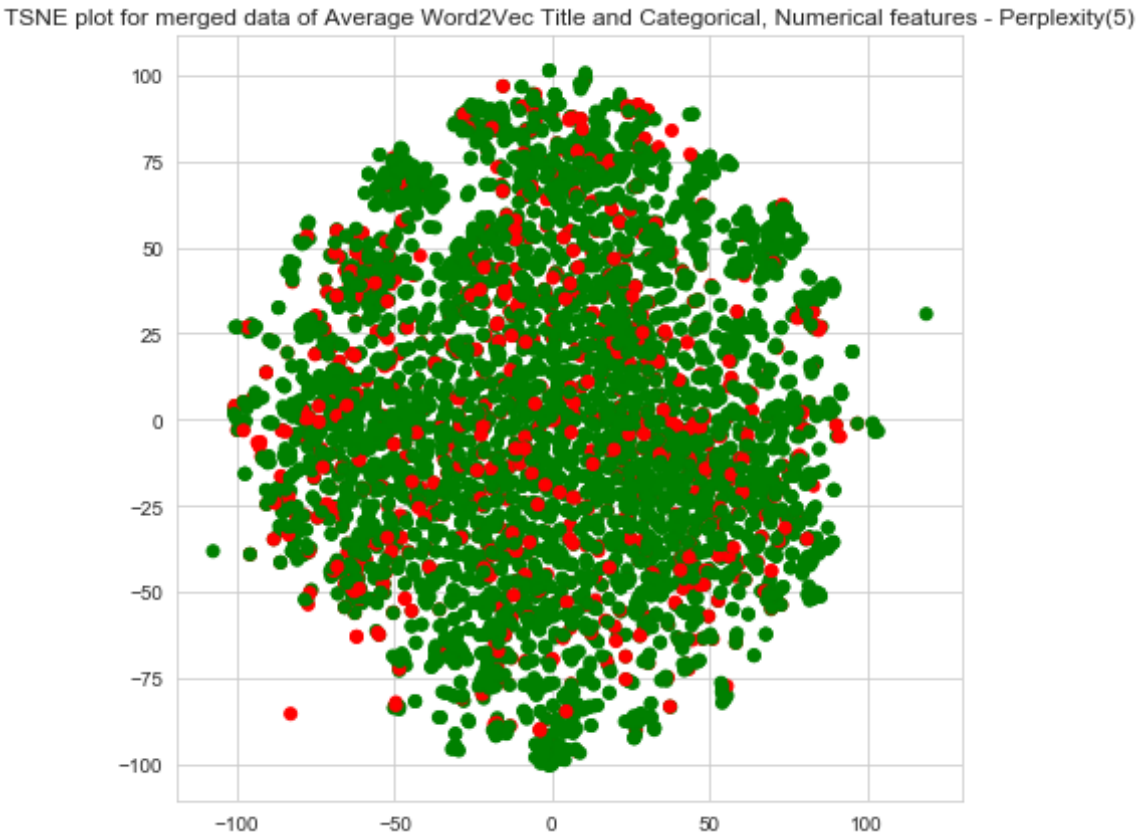

**Classification using data merged with Average Word2Vec vectorized title and all considered categorical, numerical features**

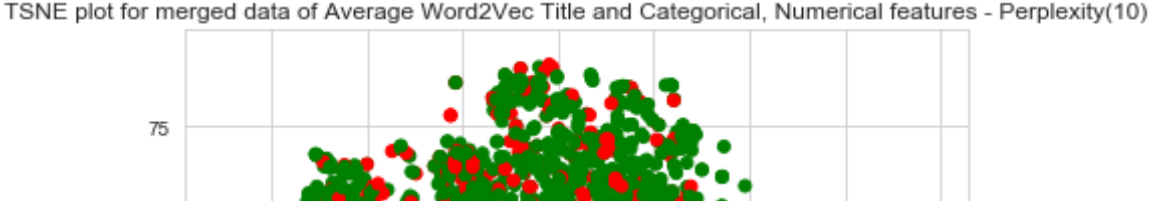
In [0]:
```
word2VecTitleAndOthers = hstack((word2VecTitlesVectorsSub, categoriesVe
ctorsSub, subCategoriesVectorsSub, teacherPrefixVectorsSub, schoolState
```

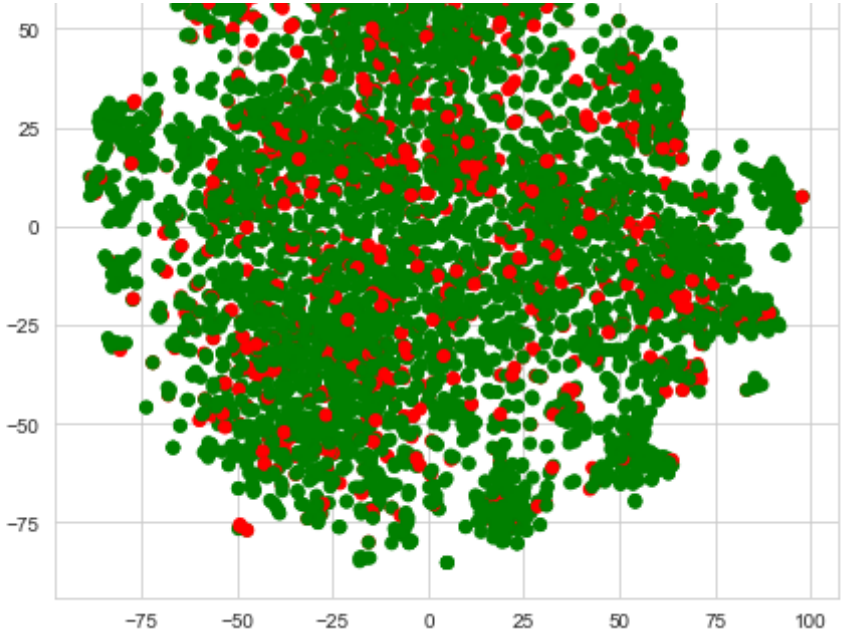
```
VectorsSub, projectGradeVectorsSub, priceStandardizedSub, previouslyPos
tedStandardizedSub));
word2VecTitleAndOthers.shape
```

Out[0]: (6000, 401)

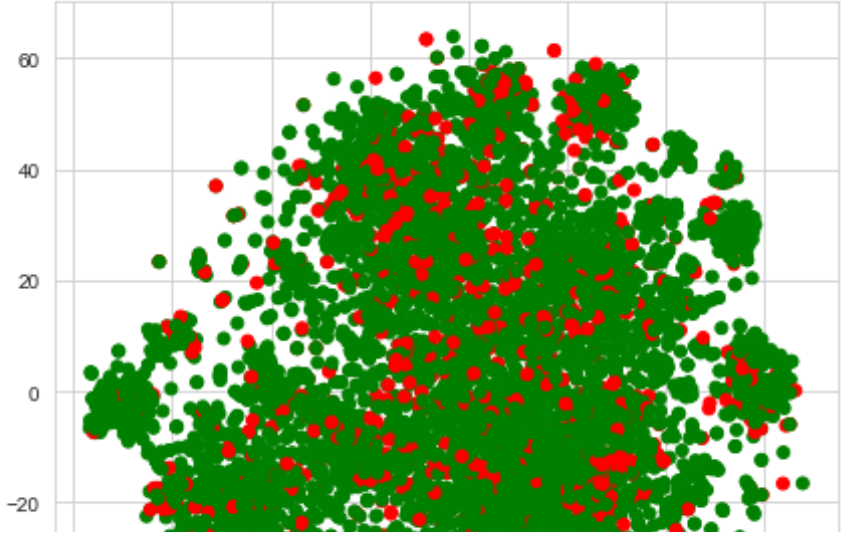
In [0]:
```
perplexityValues = [5, 10, 30, 50, 80, 100]
for perplexityValue in perplexityValues:
    tsne = TSNE(n_components = 2, perplexity = perplexityValue, learnin
g_rate = 200);
    word2VecTitleAndOthersEmbedded = tsne.fit_transform(word2VecTitleAn
dOthers.toarray());
    word2VecTitleAndOthersTsneData = np.hstack((word2VecTitleAndOthersE
mbedded, classesDataSub.reshape(-1, 1)));
    word2VecTitleAndOthersTsneDataFrame = pd.DataFrame(word2VecTitleAnd
OthersTsneData, columns = ['Dimension1', 'Dimension2', 'Class']);
    colors = {0.0:'red', 1.0:'green'}
    plt.title("TSNE plot for merged data of Average Word2Vec Title and
 Categorical, Numerical features - Perplexity({})".format(perplexityVal
ue));
    plt.scatter(word2VecTitleAndOthersTsneDataFrame['Dimension1'], word
2VecTitleAndOthersTsneDataFrame['Dimension2'], c = word2VecTitleAndOthe
rsTsneDataFrame['Class'].apply(lambda x: colors[x]));
    plt.show();
```

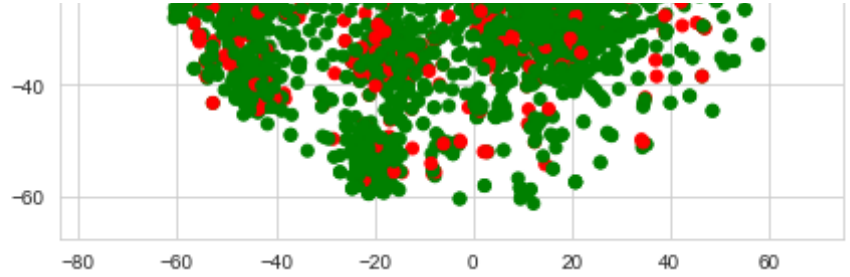


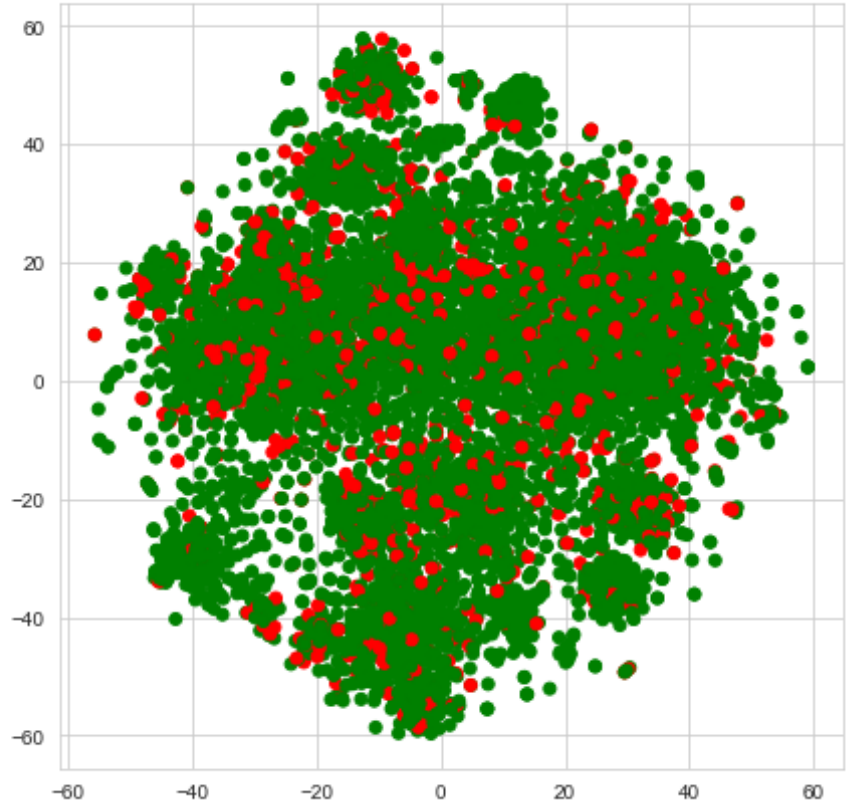


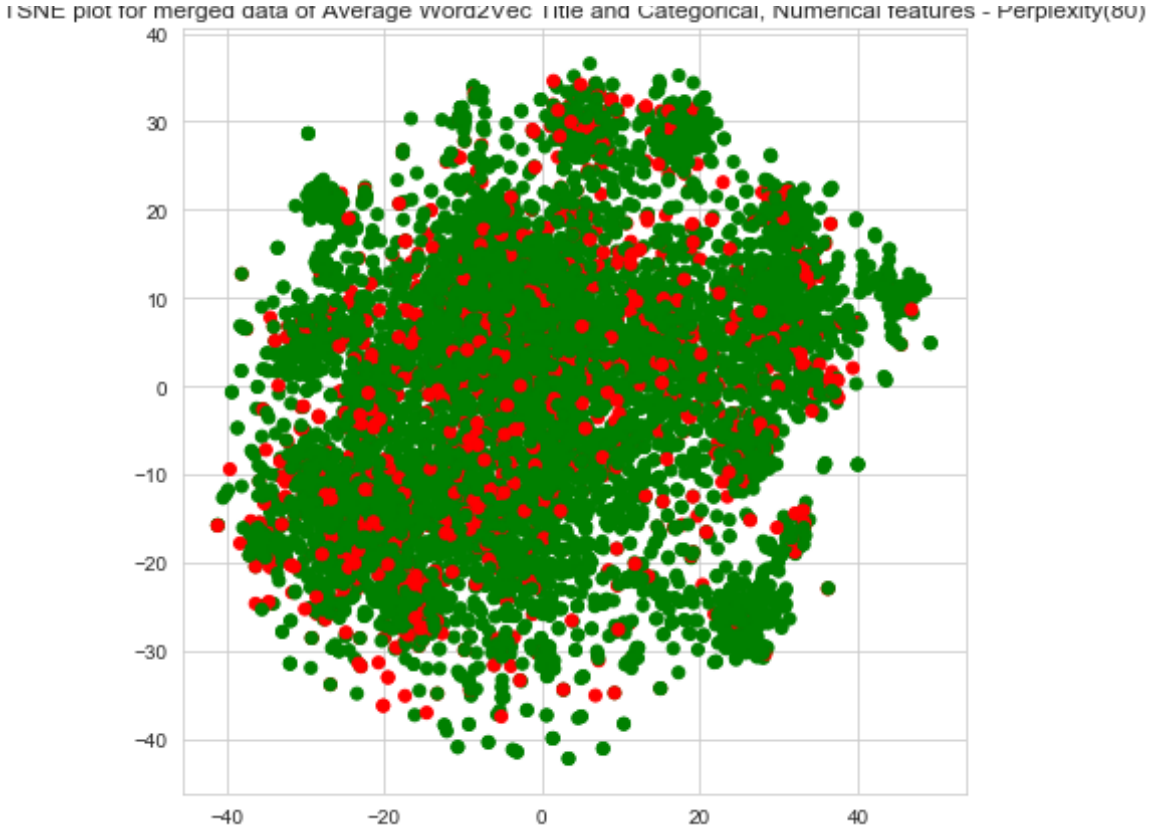

TSNE plot for merged data of Average Word2Vec Title and Categorical, Numerical features - Perplexity(30)



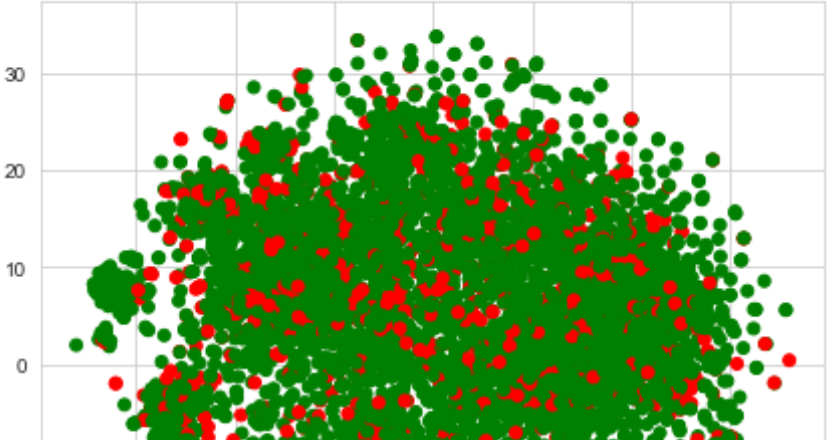


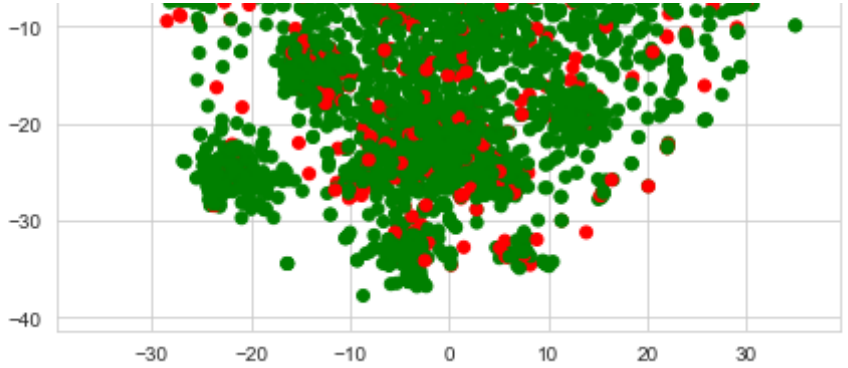

TSNE plot for merged data of Average Word2Vec Title and Categorical, Numerical features - Perplexity(50)






TSNE plot for merged data of Average Word2Vec Title and Categorical, Numerical features - Perplexity(100)





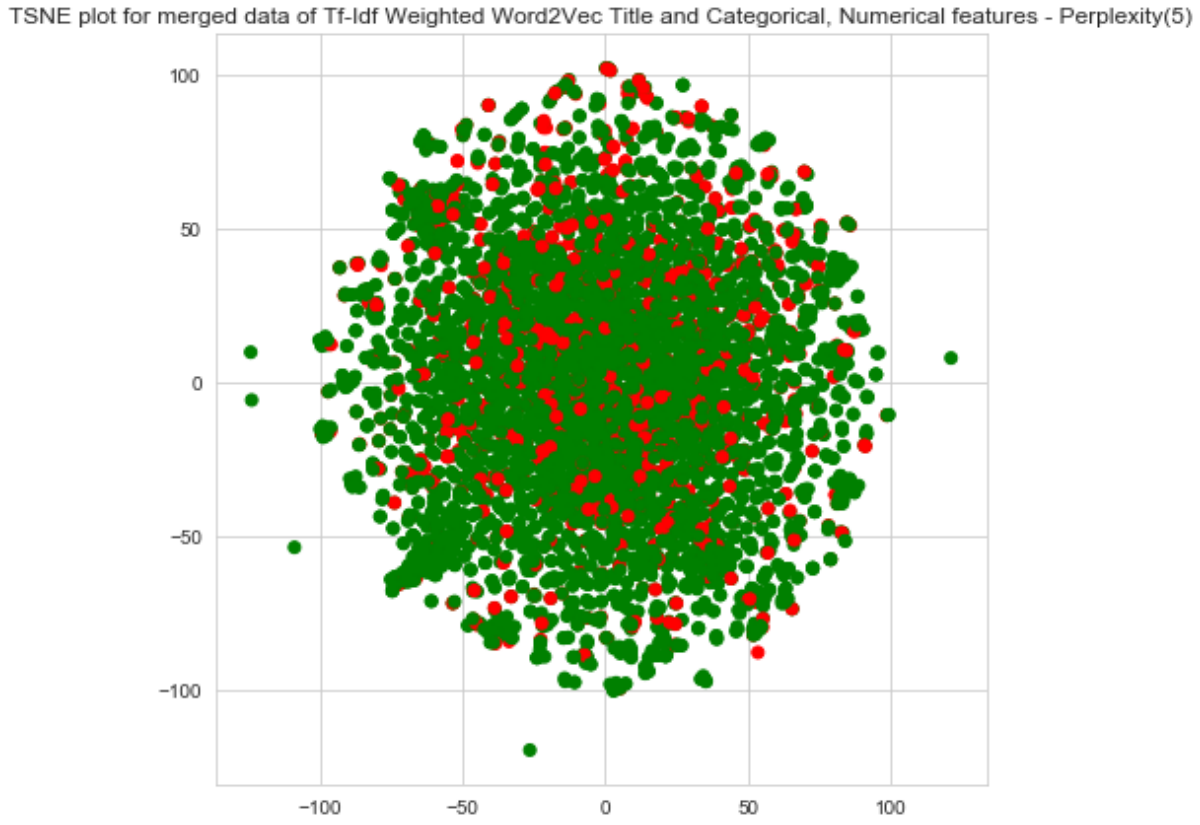

**Classification using data merged with Tf-idf Weighted Word2Vec vectorized title and all considered categorical, numerical features**

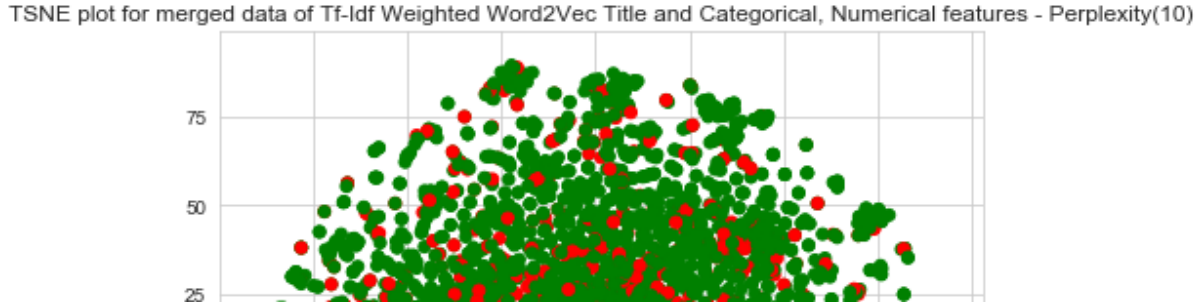
In [0]:
```
tfIdfWeightedWord2VecTitleAndOthers = hstack((tfIdfWeightedWord2VecTitlesVectorsSub, categoriesVectorsSub, subCategoriesVectorsSub, teacherPrefixVectorsSub, schoolStateVectorsSub, projectGradeVectorsSub, priceStandardizedSub, previouslyPostedStandardizedSub));
tfIdfWeightedWord2VecTitleAndOthers.shape
```

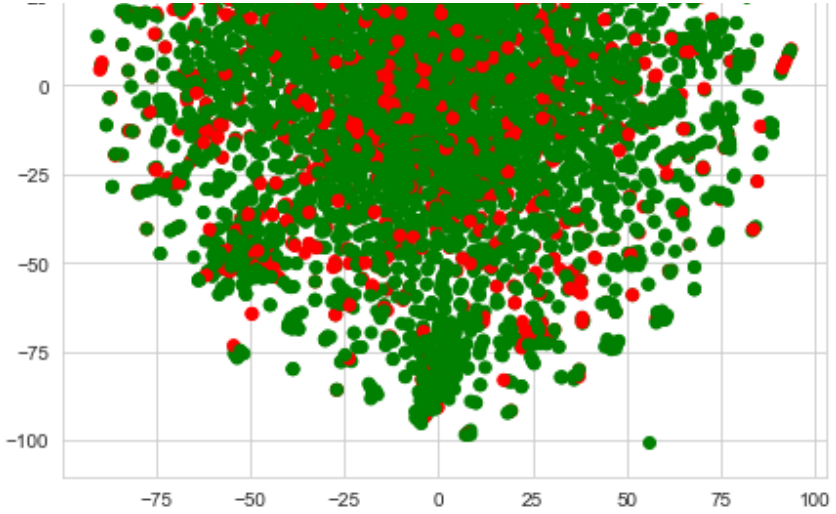
Out[0]: (6000, 401)

In [0]:
```
perplexityValues = [5, 10, 30, 50, 80, 100]
for perplexityValue in perplexityValues:
    tsne = TSNE(n_components = 2, perplexity = perplexityValue, learning_rate = 200);
    tfIdfWeightedWord2VecTitleAndOthersEmbedded = tsne.fit_transform(tfIdfWeightedWord2VecTitleAndOthers.toarray());
    tfIdfWeightedWord2VecTitleAndOthersTsneData = np.hstack((tfIdfWeightedWord2VecTitleAndOthersEmbedded, classesDataSub.reshape(-1, 1)));
    tfIdfWeightedWord2VecTitleAndOthersTsneDataFrame = pd.DataFrame(tfIdfWeightedWord2VecTitleAndOthersTsneData, columns = ['Dimension1', 'Dimension2', 'Class']);
    colors = {0.0:'red', 1.0:'green'}
    plt.title("TSNE plot for merged data of Tf-Idf Weighted Word2Vec Title and Categorical, Numerical features - Perplexity({})".format(perplexityValue));
```

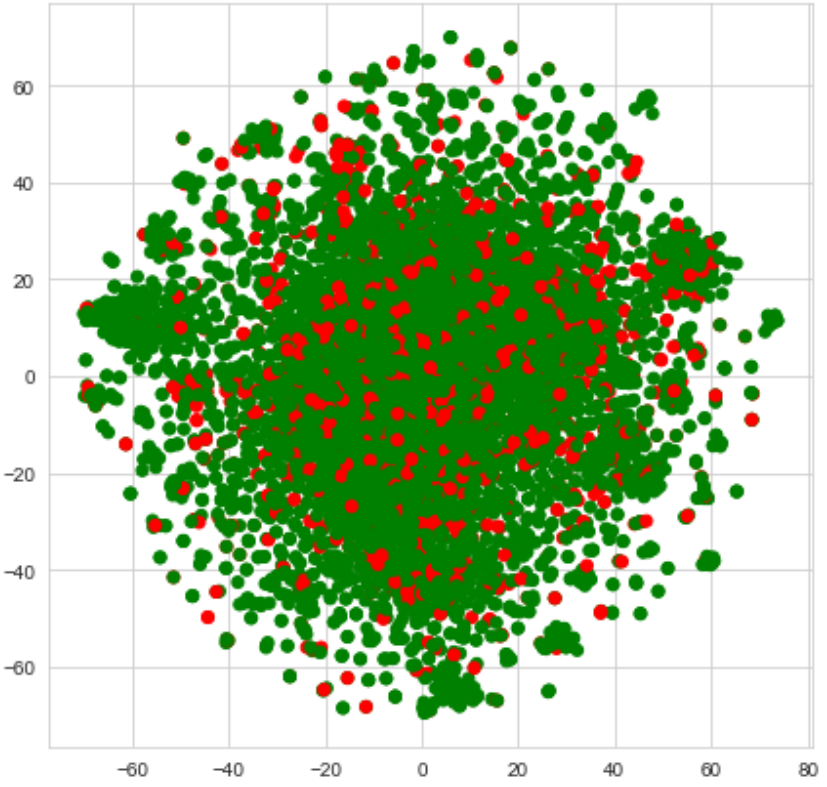
```
    plt.scatter(tfIdfWeightedWord2VecTitleAndOthersTsneDataFrame['Dimen
sion1'], tfIdfWeightedWord2VecTitleAndOthersTsneDataFrame['Dimension2'
], c = tfIdfWeightedWord2VecTitleAndOthersTsneDataFrame['Class'].apply(
lambda x: colors[x]));
    plt.show();
```

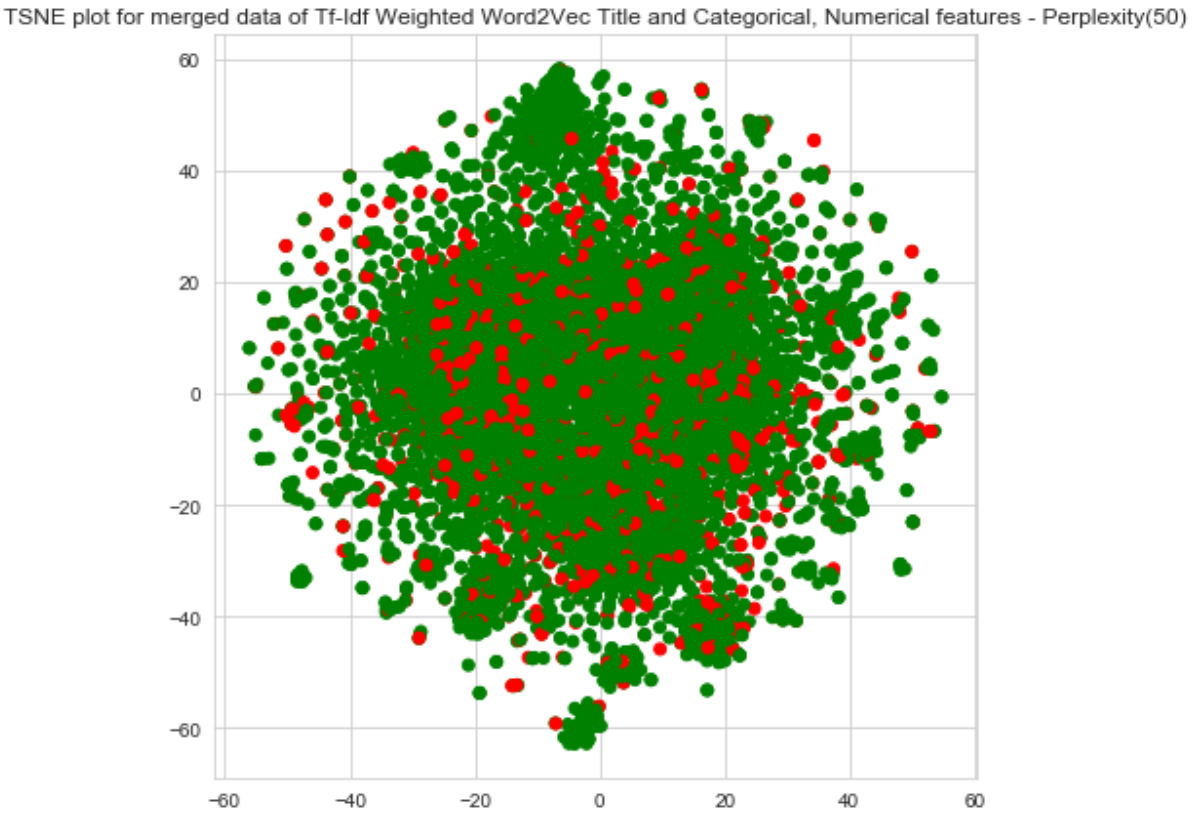


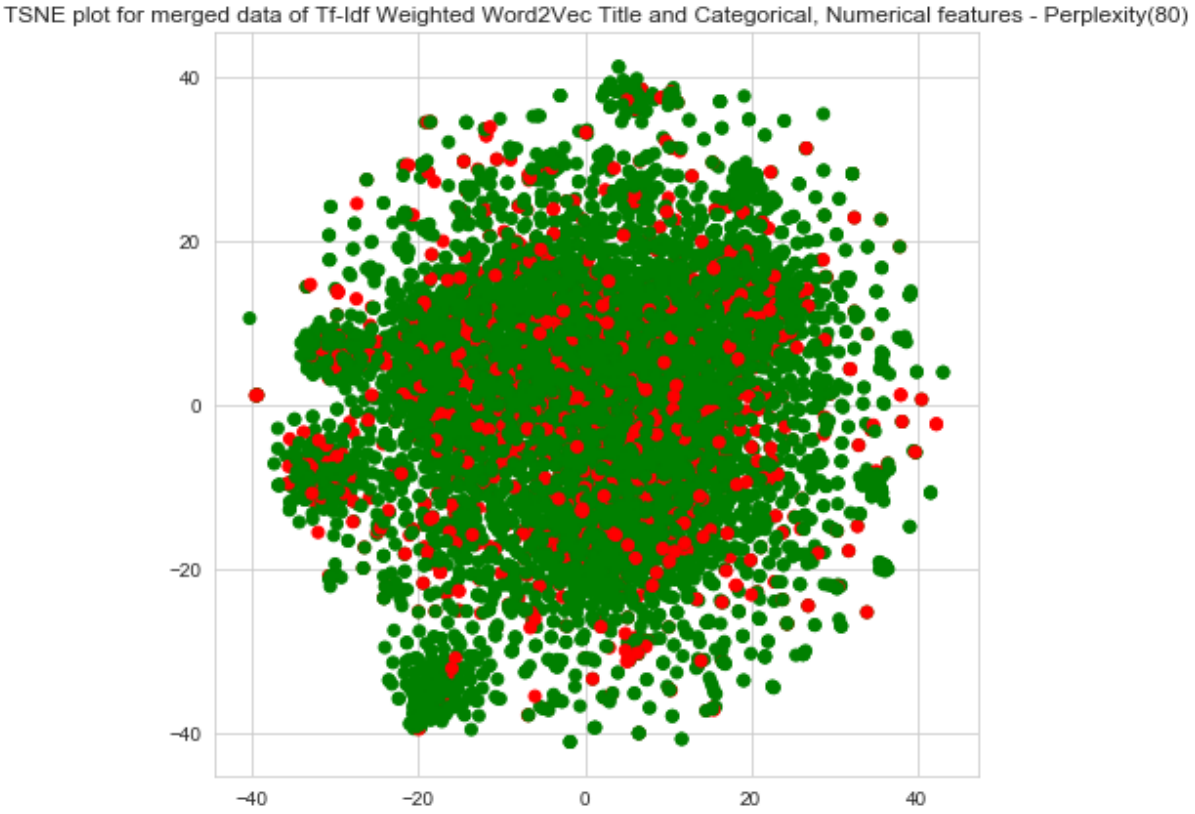



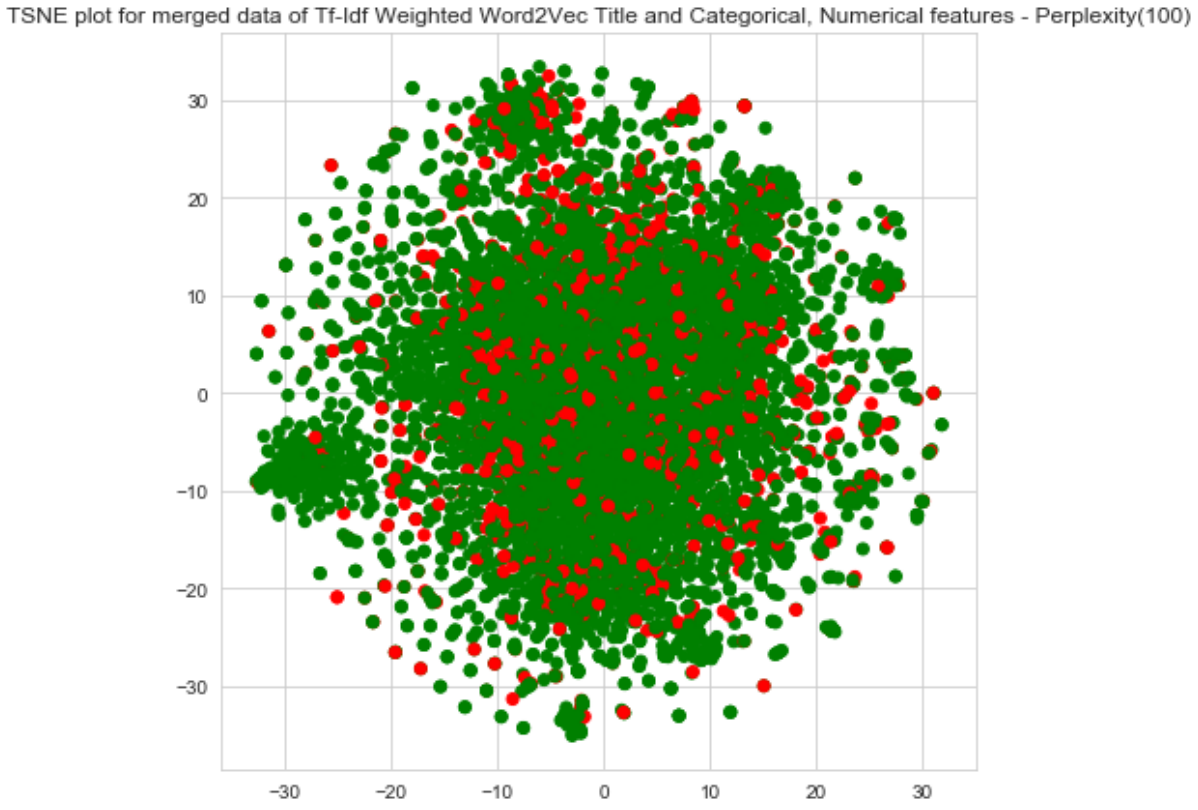



TSNE plot for merged data of Tf-Idf Weighted Word2Vec Title and Categorical, Numerical features - Perplexity(30)









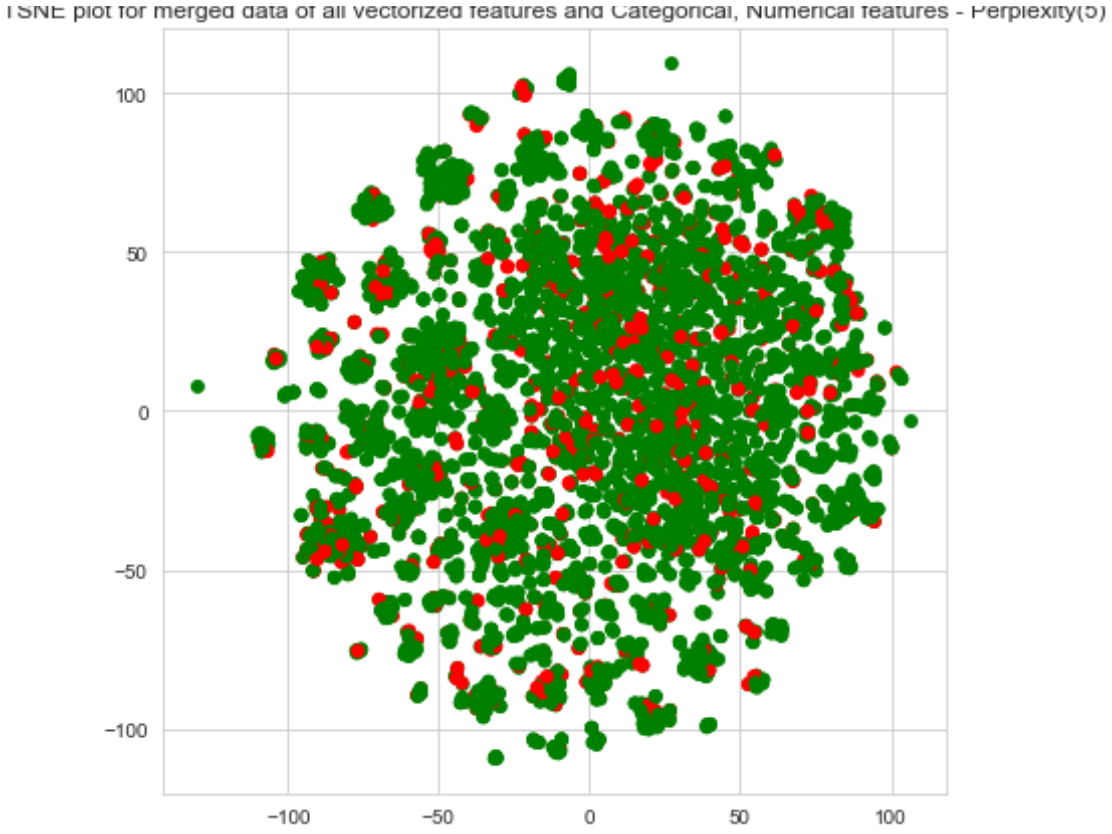

**Classification using data merged with all vectorizations of project_title and with all considered categorical, numerical features**

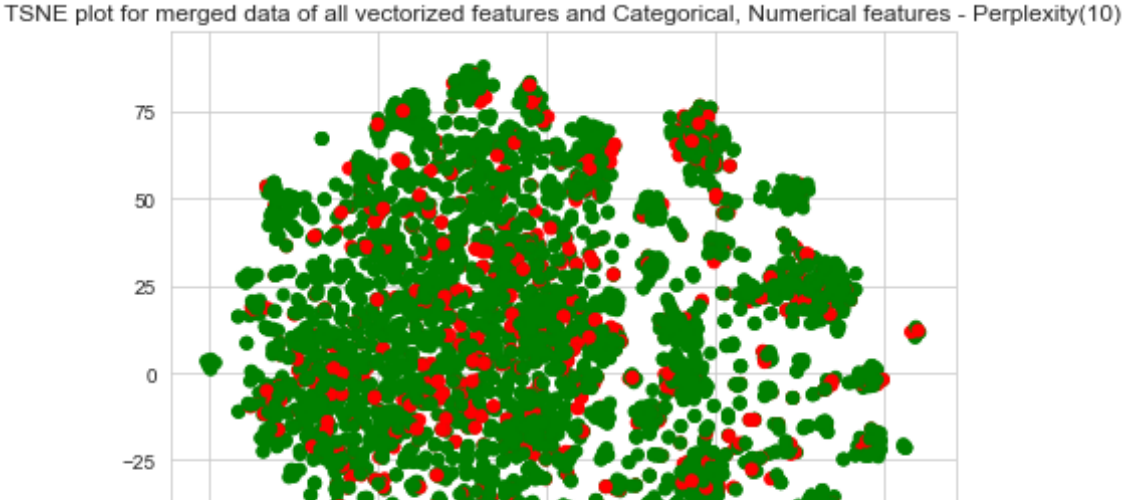
In [0]:
```
allFeatures = hstack((bowTitleModelSub, tfIdfTitleModelSub, word2VecTitlesVectorsSub, tfIdfWeightedWord2VecTitlesVectorsSub, categoriesVectorsSub, subCategoriesVectorsSub, teacherPrefixVectorsSub, schoolStateVectorsSub, projectGradeVectorsSub, priceStandardizedSub, previouslyPostedStandardizedSub))
print(allFeatures.shape)
```

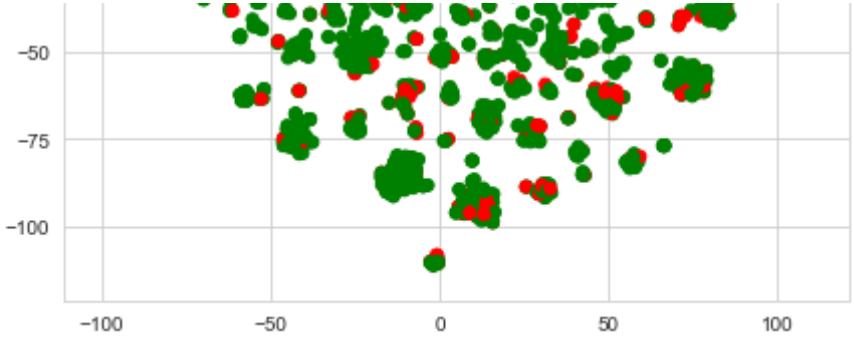
```
(6000, 4249)
```

In [0]:
```
perplexityValues = [5, 10, 30, 50, 80, 100]
```

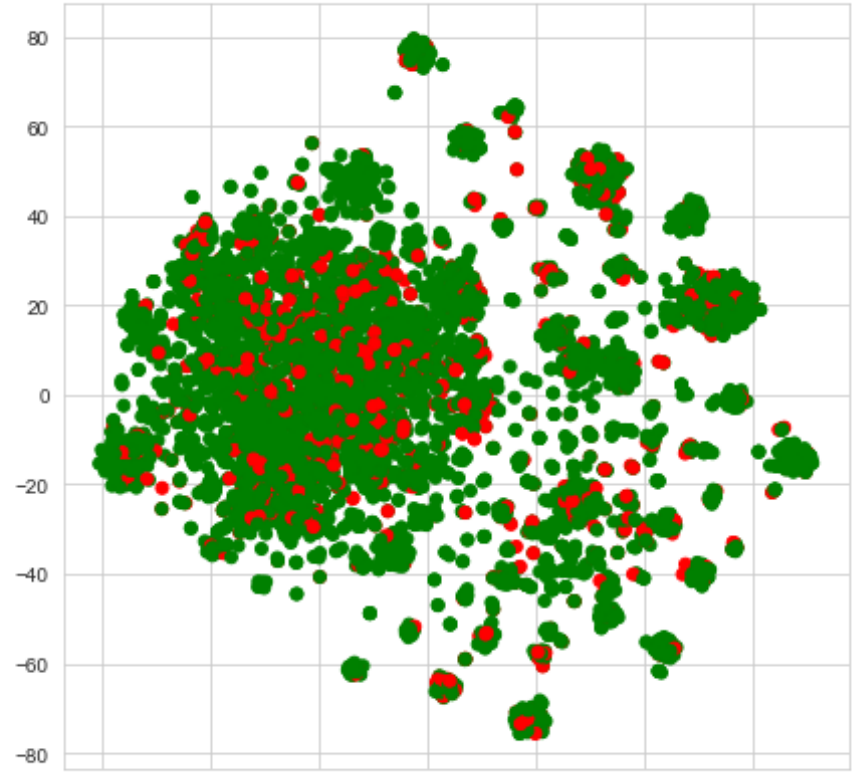
```
for perplexityValue in perplexityValues:
    tsne = TSNE(n_components = 2, perplexity = perplexityValue, learnin
g_rate = 200);
    allFeaturesEmbedded = tsne.fit_transform(allFeatures.toarray());
    allFeaturesTsneData = np.hstack((allFeaturesEmbedded, classesDataSu
b.reshape(-1, 1)));
    allFeaturesTsneDataFrame = pd.DataFrame(allFeaturesTsneData, column
s = ['Dimension1', 'Dimension2', 'Class']);
    colors = {0.0:'red', 1.0:'green'}
    plt.title("TSNE plot for merged data of all vectorized features and
 Categorical, Numerical features - Perplexity({})".format(perplexityVal
ue));
    plt.scatter(allFeaturesTsneDataFrame['Dimension1'], allFeaturesTsne
DataFrame['Dimension2'], c = allFeaturesTsneDataFrame['Class'].apply(la
mbda x: colors[x]));
    plt.show();
```

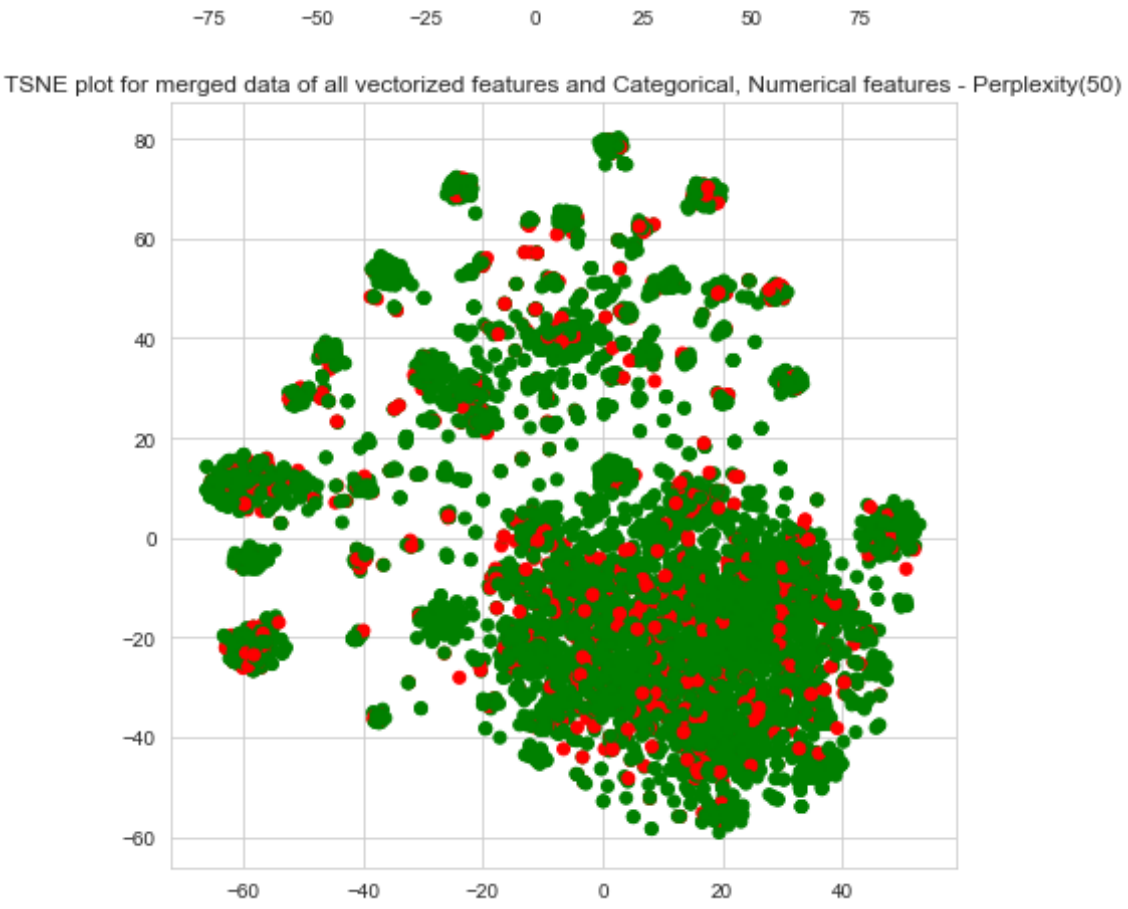


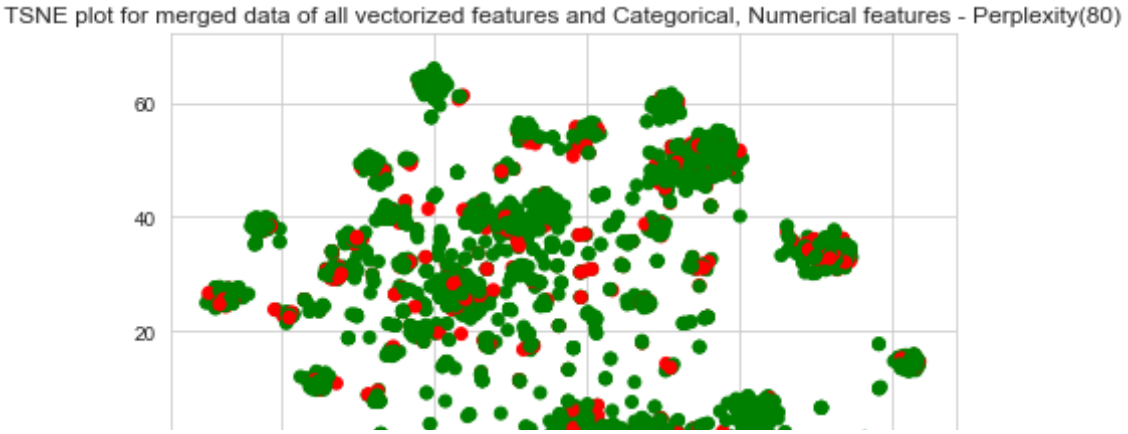



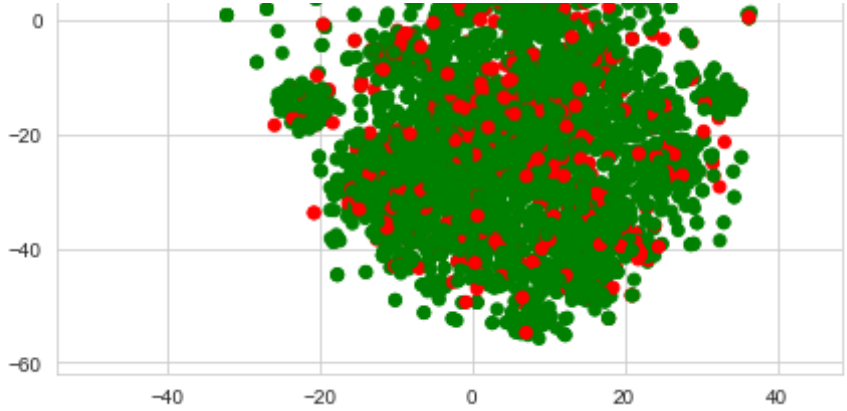



TSNE plot for merged data of all vectorized features and Categorical, Numerical features - Perplexity(30)

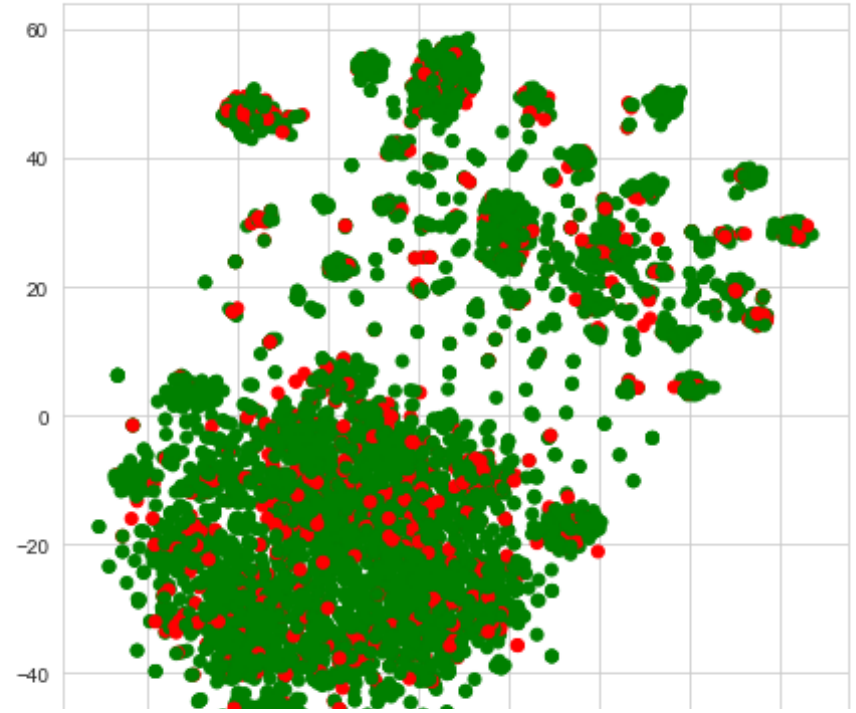


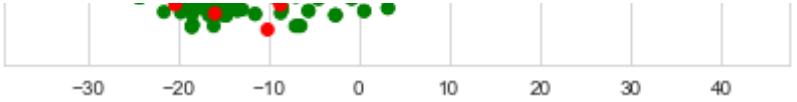





TSNE plot for merged data of all vectorized features and Categorical, Numerical features - Perplexity(100)






## Conclusion about data visualization using t-sne:

1. Bag of Words, Tf-Idf are better than word2vec vectorizations because of forming some small group of clusters with less overlap of overall data when compared to others.
2. Higher perplexity values seems better in data visualization because of less overlap of data than others.
3. None of the techniques are useful for classification because of huge overlap of data.
4. It is not seperable problem in 2-dimensions but it may be seperable in higher dimensions.

# Classification & Modelling using Logistic Regression

## Classification of data using Logistic Regression

### Splitting Data(Only training and test)

In [0]:
```
projectsData = projectsData.dropna(subset = ['teacher_prefix']);
projectsData.shape
```

Out[0]:
```
(109245, 22)
```

In [0]:
```
classesData = projectsData['project_is_approved']
print(classesData.shape)
```

```
(109245,)
```

In [0]:
```
trainingData, testData, classesTraining, classesTest = model_selection.
train_test_split(projectsData, classesData, test_size =  0.3, random_st
ate = 0, stratify = classesData);
trainingData, crossValidateData, classesTraining, classesCrossValidate
= model_selection.train_test_split(trainingData, classesTraining, test_
size = 0.3, random_state = 0, stratify = classesTraining);
```

In [0]:
```
print("Shapes of splitted data: ");
equalsBorder(70);

print("testData shape: ", testData.shape);
print("classesTest: ", classesTest.shape);
print("trainingData shape: ", trainingData.shape);
print("classesTraining shape: ", classesTraining.shape);
```

```
Shapes of splitted data:
======================================================================
testData shape:  (32774, 22)
classesTest:  (32774,)
trainingData shape:  (53529, 22)
classesTraining shape:  (53529,)
```

In [0]:
```
print("Number of negative points: ", trainingData[trainingData['project
_is_approved'] == 0].shape);
print("Number of positive points: ", trainingData[trainingData['project
_is_approved'] == 1].shape);
```

```
Number of negative points:  (8105, 22)
Number of positive points:  (45424, 22)
```

In [0]:
```
vectorizedFeatureNames = [];
```

## Balancing Data

***Note: Instead of displaying whole vectorization process for balanced and imbalanced data, we have simply disabled below cell while performing analysis on imbalanced data and enabled while performing analysis on balanced data***

In [0]:
```
negativeData = trainingData[trainingData['project_is_approved'] == 0];
positiveData = trainingData[trainingData['project_is_approved'] == 1];
negativeDataBalanced = resample(negativeData, replace = True, n_samples
 = trainingData[trainingData['project_is_approved'] == 1].shape[0], ran
dom_state = 44);
trainingData = pd.concat([positiveData, negativeDataBalanced]);
trainingData = shuffle(trainingData);
classesTraining = trainingData['project_is_approved'];
print("Testing whether data is balanced: ");
equalsBorder(60);
print("Number of positive points: ", trainingData[trainingData['project
_is_approved'] == 1].shape);
print("Number of negative points: ", trainingData[trainingData['project
_is_approved'] == 0].shape);
```

```
Testing whether data is balanced:
============================================================
Number of positive points:  (45424, 22)
Number of negative points:  (45424, 22)
```

# Vectorizing categorical data

## 1. Vectorizing cleaned_categories(project_subject_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_categories
subjectsCategoriesVectorizer = CountVectorizer(vocabulary = list(sorted
CategoriesDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with cleaned_categories values
subjectsCategoriesVectorizer.fit(trainingData['cleaned_categories'].val
ues);
```

```
# Vectorizing categories using one-hot-encoding
categoriesVectors = subjectsCategoriesVectorizer.transform(trainingData
['cleaned_categories'].values);
```

In [0]:
```
print("Features used in vectorizing categories: ");
equalsBorder(70);
print(subjectsCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", categoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(categoriesVectors[0:4])
```

```
Features used in vectorizing categories:
======================================================================
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearn
ing', 'SpecialNeeds', 'Health_Sports', 'Math_Science', 'Literacy_Langua
ge']
======================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (90848, 9)
======================================================================
Sample vectors of categories:
======================================================================
  (0, 7)        1
  (0, 8)        1
  (1, 7)        1
  (2, 7)        1
  (2, 8)        1
  (3, 6)        1
```

### 2. Vectorizing cleaned_sub_categories(project_subject_sub_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_sub_categories
subjectsSubCategoriesVectorizer = CountVectorizer(vocabulary = list(sor
tedDictionarySubCategories.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with cleaned_sub_categories values
subjectsSubCategoriesVectorizer.fit(trainingData['cleaned_sub_categorie
s'].values);
# Vectorizing sub categories using one-hot-encoding
subCategoriesVectors = subjectsSubCategoriesVectorizer.transform(traini
ngData['cleaned_sub_categories'].values);
```

In [0]:
```
print("Features used in vectorizing subject sub categories: ");
equalsBorder(70);
print(subjectsSubCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", subCategoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(subCategoriesVectors[0:4])
```

```
Features used in vectorizing subject sub categories:
======================================================================
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolveme
nt', 'Extracurricular', 'Civics_Government', 'ForeignLanguages', 'Nutri
tionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingA
rts', 'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPre
p', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopme
nt', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Healt
h_Wellness', 'AppliedSciences', 'SpecialNeeds', 'Literature_Writing',
'Mathematics', 'Literacy']
======================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (90848, 30)
======================================================================
Sample vectors of categories:
======================================================================
  (0, 27)       1
```

```
  (0, 28)       1
  (1, 28)       1
  (2, 28)       1
  (2, 29)       1
  (3, 21)       1
  (3, 24)       1
```

### 3. Vectorizing teacher_prefix - One Hot Encoding

In [0]:
```
def giveCounter(data):
    counter = Counter();
    for dataValue in data:
        counter.update(str(dataValue).split());
    return counter
```

In [0]:
```
giveCounter(trainingData['teacher_prefix'].values)
```

Out[0]:
```
Counter({'Dr': 7, 'Mr': 8836, 'Mrs': 46892, 'Ms': 32914, 'Teacher': 219
9})
```

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(trainingData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(trainingData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(trainingData[
'teacher_prefix'].values);
```

In [0]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
```

```
ing): ", teacherPrefixVectors.shape);
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:100]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs', 'Ms', 'Mr', 'Teacher', 'Dr']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(90848, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 1)        1
  (3, 1)        1
  (4, 0)        1
  (5, 0)        1
  (6, 1)        1
  (7, 0)        1
  (8, 1)        1
  (9, 1)        1
  (10, 1)       1
  (11, 0)       1
  (12, 0)       1
  (13, 0)       1
  (14, 0)       1
  (15, 0)       1
  (16, 1)       1
  (17, 0)       1
  (18, 1)       1
  (19, 0)       1
  (20, 1)       1
  (21, 0)       1
  (22, 2)       1
  (23, 1)       1
  (24, 1)       1
```

```
:     :
(75, 0)       1
(76, 1)       1
(77, 0)       1
(78, 1)       1
(79, 0)       1
(80, 0)       1
(81, 1)       1
(82, 1)       1
(83, 0)       1
(84, 0)       1
(85, 0)       1
(86, 0)       1
(87, 0)       1
(88, 0)       1
(89, 0)       1
(90, 0)       1
(91, 1)       1
(92, 0)       1
(93, 0)       1
(94, 0)       1
(95, 2)       1
(96, 1)       1
(97, 1)       1
(98, 1)       1
(99, 1)       1
```

In [0]:
```
teacherPrefixes = [prefix.replace('.', '') for prefix in trainingData[
'teacher_prefix'].values];
teacherPrefixes[0:5]
```

Out[0]:
```
['Mrs', 'Ms', 'Ms', 'Ms', 'Mrs']
```

In [0]:
```
trainingData['teacher_prefix'] = teacherPrefixes;
trainingData.head(3)
```

Out[0]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_ |
|---|---|---|---|---|---|
| **41663** | 99778 | p163668 | fa7ea9c64017bb18ec1994a8885c0c30 | Mrs | WI |
| **65749** | 96875 | p168614 | e748de284a1cdd2e249011ca83a50630 | Ms | FL |
| **78862** | 132015 | p194999 | c121651d9c749d0639efcfdba0c73ca9 | Ms | ME |

3 rows × 22 columns

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(trainingData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(trainingData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(trainingData[
'teacher_prefix'].values);
```

In [0]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
```

```
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
ing): ", teacherPrefixVectors.shape);
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:4]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs', 'Ms', 'Mr', 'Teacher', 'Dr']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(90848, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 1)        1
  (3, 1)        1
```

### 4. Vectorizing school_state - One Hot Encoding

In [0]:
```
schoolStateDictionary = dict(giveCounter(trainingData['school_state'].v
alues));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique school states
schoolStateVectorizer = CountVectorizer(vocabulary = list(schoolStateDi
ctionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with school_state values
schoolStateVectorizer.fit(trainingData['school_state'].values);
# Vectorizing school_state using one-hot-encoding
schoolStateVectors = schoolStateVectorizer.transform(trainingData['scho
ol_state'].values);
```

In [0]:
```
print("Features used in vectorizing school_state: ");
```

```
equalsBorder(70);
print(schoolStateVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of school_state matrix after vectorization(one-hot-encodin
g): ", schoolStateVectors.shape);
equalsBorder(70);
print("Sample vectors of school_state: ");
equalsBorder(70);
print(schoolStateVectors[0:4]);
```

```
Features used in vectorizing school_state:
======================================================================
['WI', 'FL', 'ME', 'MI', 'MT', 'CO', 'CA', 'MN', 'UT', 'IL', 'NC', 'M
A', 'GA', 'TX', 'NJ', 'ND', 'LA', 'KY', 'OH', 'MS', 'WA', 'AZ', 'OK',
'MD', 'NE', 'VT', 'OR', 'PA', 'NV', 'IN', 'TN', 'MO', 'AL', 'DC', 'IA',
'SD', 'VA', 'NY', 'NM', 'SC', 'CT', 'HI', 'NH', 'DE', 'AR', 'ID', 'WY',
'KS', 'WV', 'RI', 'AK']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (9
0848, 51)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 2)        1
  (3, 3)        1
```

### 5. Vectorizing project_grade_category - One Hot Encoding

In [0]:
```
giveCounter(trainingData['project_grade_category'])
```

Out[0]:
```
Counter({'Grades3to5': 30465,
         'Grades6to8': 14131,
         'Grades9to12': 9136,
         'GradesPreKto2': 37116})
```

In [0]:
```
cleanedGrades = []
```

```
for grade in trainingData['project_grade_category'].values:
    grade = grade.replace(' ', '');
    grade = grade.replace('-', 'to');
    cleanedGrades.append(grade);
cleanedGrades[0:4]
```

Out[0]: ['GradesPreKto2', 'GradesPreKto2', 'Grades3to5', 'Grades3to5']

In [0]:
```
trainingData['project_grade_category'] = cleanedGrades
trainingData.head(4)
```

Out[0]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_ |
|---|---|---|---|---|---|
| **41663** | 99778 | p163668 | fa7ea9c64017bb18ec1994a8885c0c30 | Mrs | WI |
| **65749** | 96875 | p168614 | e748de284a1cdd2e249011ca83a50630 | Ms | FL |
| **78862** | 132015 | p194999 | c121651d9c749d0639efcfdba0c73ca9 | Ms | ME |
| **40122** | 156126 | p122388 | aa96ef933643e81d26495f0cbaea40aa | Ms | MI |

4 rows × 22 columns

In [0]:
```
projectGradeDictionary = dict(giveCounter(trainingData['project_grade_category'].values));
# Using CountVectorizer for performing one-hot-encoding by setting vocabulary as list of all unique project grade categories
projectGradeVectorizer = CountVectorizer(vocabulary = list(projectGradeDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with project_grade_category values
projectGradeVectorizer.fit(trainingData['project_grade_category'].values);
# Vectorizing project_grade_category using one-hot-encoding
projectGradeVectors = projectGradeVectorizer.transform(trainingData['project_grade_category'].values);
```

In [0]:
```
print("Features used in vectorizing project_grade_category: ");
equalsBorder(70);
print(projectGradeVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of school_state matrix after vectorization(one-hot-encoding): ", projectGradeVectors.shape);
equalsBorder(70);
print("Sample vectors of school_state: ");
equalsBorder(70);
print(projectGradeVectors[0:4]);
```

```
Features used in vectorizing project_grade_category:
======================================================================
['GradesPreKto2', 'Grades3to5', 'Grades6to8', 'Grades9to12']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (90848, 4)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 0)        1
  (2, 1)        1
  (3, 1)        1
```

## Vectorizing Text Data

In [0]:
```
preProcessedEssaysWithStopWords, preProcessedEssaysWithoutStopWords = p
reProcessingWithAndWithoutStopWords(trainingData['project_essay']);
preProcessedProjectTitlesWithStopWords, preProcessedProjectTitlesWithou
tStopWords = preProcessingWithAndWithoutStopWords(trainingData['project
_title']);
```

In [0]:
```
bagOfWordsVectorizedFeatures = [];
```

### Bag of Words

#### 1. Vectorizing project_essay

In [0]:
```
# Initializing countvectorizer for bag of words vectorization of prepro
cessed project essays
bowEssayVectorizer = CountVectorizer(min_df = 10, max_features = 5000);
# Transforming the preprocessed essays to bag of words vectors
bowEssayModel = bowEssayVectorizer.fit_transform(preProcessedEssaysWith
outStopWords);
```

In [0]:
```
print("Some of the Features used in vectorizing preprocessed essays: "
);
equalsBorder(70);
print(bowEssayVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed essay matrix after vectorization: ", bowEs
sayModel.shape);
equalsBorder(70);
```

```
print("Sample bag-of-words vector of preprocessed essay: ");
equalsBorder(70);
print(bowEssayModel[0])
```

```
Some of the Features used in vectorizing preprocessed essays:
======================================================================
['worry', 'worrying', 'worse', 'worst', 'worth', 'worthy', 'would', 'wo
w', 'wrestling', 'write', 'writer', 'writers', 'writing', 'writings',
'written', 'wrong', 'wrote', 'xylophones', 'yard', 'year', 'yearbook',
'yearly', 'yearn', 'yearning', 'years', 'yes', 'yesterday', 'yet', 'yog
a', 'york', 'young', 'younger', 'youngest', 'youth', 'youtube', 'zest',
'zip', 'zone', 'zones', 'zoo']
======================================================================
Shape of preprocessed essay matrix after vectorization:  (90848, 5000)
======================================================================
Sample bag-of-words vector of preprocessed essay:
======================================================================
  (0, 3021)     1
  (0, 1091)     1
  (0, 818)      1
  (0, 4841)     1
  (0, 4186)     1
  (0, 4418)     1
  (0, 4684)     1
  (0, 4115)     1
  (0, 1849)     1
  (0, 3994)     2
  (0, 4754)     2
  (0, 328)      1
  (0, 2990)     1
  (0, 3168)     1
  (0, 2713)     1
  (0, 3322)     1
  (0, 1412)     1
  (0, 4790)     1
  (0, 2635)     2
  (0, 1555)     2
  (0, 893)      1
  (0, 2155)     2
  (0, 1466)     1
```

```
(0, 2211)     1
(0, 3910)     3
:     :
(0, 750)      1
(0, 2864)     1
(0, 1584)     1
(0, 2514)     1
(0, 22)       1
(0, 810)      1
(0, 2604)     1
(0, 4979)     2
(0, 2262)     1
(0, 2357)     1
(0, 2756)     1
(0, 890)      2
(0, 3319)     1
(0, 2206)     2
(0, 4591)     1
(0, 427)      1
(0, 3053)     1
(0, 808)      1
(0, 1339)     1
(0, 4363)     7
(0, 4921)     1
(0, 3960)     2
(0, 4751)     1
(0, 2555)     3
(0, 4478)     1
```

**2. Vectorizing project_title**

In [0]:
```
# Initializing countvectorizer for bag of words vectorization of prepro
cessed project titles
bowTitleVectorizer = CountVectorizer(min_df = 10);
# Transforming the preprocessed project titles to bag of words vectors
bowTitleModel = bowTitleVectorizer.fit_transform(preProcessedProjectTit
lesWithoutStopWords);
```

In [0]:
```
print("Some of the Features used in vectorizing preprocessed titles: "
);
equalsBorder(70);
print(bowTitleVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after vectorization: ", bowTi
tleModel.shape);
equalsBorder(70);
print("Sample bag-of-words vector of preprocessed title: ");
equalsBorder(70);
print(bowTitleModel[0])
```

```
Some of the Features used in vectorizing preprocessed titles:
======================================================================
['work', 'workers', 'working', 'workout', 'works', 'worksheets', 'works
hop', 'world', 'worlds', 'worldwide', 'worms', 'worth', 'would', 'wow',
'wrestling', 'write', 'writer', 'writers', 'writing', 'written', 'xylop
hone', 'ye', 'yeah', 'year', 'yearbook', 'yearbooks', 'years', 'yes',
'yet', 'yoga', 'yogi', 'yogis', 'young', 'youngest', 'youngsters', 'you
th', 'youtube', 'zearn', 'zone', 'zoom']
======================================================================
Shape of preprocessed title matrix after vectorization:  (90848, 3095)
======================================================================
Sample bag-of-words vector of preprocessed title:
======================================================================
  (0, 1621)     1
  (0, 2415)     1
```

## Tf-Idf Vectorization

#### 1. Vectorizing project_essay

In [0]:
```
# Intializing tfidf vectorizer for tf-idf vectorization of preprocessed
 project essays
tfIdfEssayVectorizer = TfidfVectorizer(min_df = 10, max_features = 5000
);
# Transforming the preprocessed project essays to tf-idf vectors
```

```
tfIdfEssayModel = tfIdfEssayVectorizer.fit_transform(preProcessedEssays
WithoutStopWords);
```

In [0]:
```
print("Some of the Features used in tf-idf vectorizing preprocessed ess
ays: ");
equalsBorder(70);
print(tfIdfEssayVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after tf-idf vectorization: "
, tfIdfEssayModel.shape);
equalsBorder(70);
print("Sample Tf-Idf vector of preprocessed essay: ");
equalsBorder(70);
print(tfIdfEssayModel[0])
```

```
Some of the Features used in tf-idf vectorizing preprocessed essays:
======================================================================
['worry', 'worrying', 'worse', 'worst', 'worth', 'worthy', 'would', 'wo
w', 'wrestling', 'write', 'writer', 'writers', 'writing', 'writings',
'written', 'wrong', 'wrote', 'xylophones', 'yard', 'year', 'yearbook',
'yearly', 'yearn', 'yearning', 'years', 'yes', 'yesterday', 'yet', 'yog
a', 'york', 'young', 'younger', 'youngest', 'youth', 'youtube', 'zest',
'zip', 'zone', 'zones', 'zoo']
======================================================================
Shape of preprocessed title matrix after tf-idf vectorization:  (90848,
5000)
======================================================================
Sample Tf-Idf vector of preprocessed essay:
======================================================================
  (0, 4478)     0.039247009382295285
  (0, 2555)     0.16725491374506513
  (0, 4751)     0.07093355619072357
  (0, 3960)     0.041104501971211536
  (0, 4921)     0.12844335612325308
  (0, 4363)     0.12446816643713174
  (0, 1339)     0.046930347184745536
  (0, 808)      0.06164951301917635
  (0, 3053)     0.0877027802798548
  (0, 427)      0.04944306092319574
  (0, 4591)     0.045820804169204
```

```
(0, 4591)     0.04502604041092047
(0, 2206)     0.08110478952671518
(0, 3319)     0.09306927912300185
(0, 890)      0.06107253063523166
(0, 2756)     0.04367549719385363
(0, 2357)     0.0480325394657383
(0, 2262)     0.07903913608758646
(0, 4979)     0.07058112210246457
(0, 2604)     0.06268344298443797
(0, 810)      0.033785212246706144
(0, 22)       0.08055485905304191
(0, 2514)     0.06783294831373751
(0, 1584)     0.07128119720923672
(0, 2864)     0.05709469837173233
(0, 750)      0.07000578463547066
:     :
(0, 3910)     0.2645594416818395
(0, 2211)     0.07835497232473385
(0, 1466)     0.08150631714389067
(0, 2155)     0.15300729014966274
(0, 893)      0.06170868089153746
(0, 1555)     0.1101762977545431
(0, 2635)     0.05133011837531919
(0, 4790)     0.124181066211245
(0, 1412)     0.11013842131478242
(0, 3322)     0.08102416372895034
(0, 2713)     0.05541345240532585
(0, 3168)     0.08114525709557913
(0, 2990)     0.05366279971339146
(0, 328)      0.04887803690981847
(0, 4754)     0.0645544752947594
(0, 3994)     0.09221565838635849
(0, 1849)     0.08148444354266587
(0, 4115)     0.05974978698250007
(0, 4684)     0.0755470839740163
(0, 4418)     0.1118913258309135
(0, 4186)     0.08872776140274803
(0, 4841)     0.0933100566322624
(0, 818)      0.024878385155018747

(0, 1091)     0.06065059144172435
```

```
  (0, 3021)     0.018425816756719806
```

**2. Vectorizing project_title**

In [0]:
```
# Intializing tfidf vectorizer for tf-idf vectorization of preprocessed
 project titles
tfIdfTitleVectorizer = TfidfVectorizer(min_df = 10);
# Transforming the preprocessed project titles to tf-idf vectors
tfIdfTitleModel = tfIdfTitleVectorizer.fit_transform(preProcessedProjec
tTitlesWithoutStopWords);
```

In [0]:
```
print("Some of the Features used in tf-idf vectorizing preprocessed tit
les: ");
equalsBorder(70);
print(tfIdfTitleVectorizer.get_feature_names()[-40:]);
equalsBorder(70);
print("Shape of preprocessed title matrix after tf-idf vectorization: "
, tfIdfTitleModel.shape);
equalsBorder(70);
print("Sample Tf-Idf vector of preprocessed title: ");
equalsBorder(70);
print(tfIdfTitleModel[0])
```

```
Some of the Features used in tf-idf vectorizing preprocessed titles:
======================================================================
['work', 'workers', 'working', 'workout', 'works', 'worksheets', 'works
hop', 'world', 'worlds', 'worldwide', 'worms', 'worth', 'would', 'wow',
'wrestling', 'write', 'writer', 'writers', 'writing', 'written', 'xylop
hone', 'ye', 'yeah', 'year', 'yearbook', 'yearbooks', 'years', 'yes',
'yet', 'yoga', 'yogi', 'yogis', 'young', 'youngest', 'youngsters', 'you
th', 'youtube', 'zearn', 'zone', 'zoom']
======================================================================
Shape of preprocessed title matrix after tf-idf vectorization:  (90848,
3095)
======================================================================
Sample Tf-Idf vector of preprocessed title:
======================================================================
```

```
  (0, 2415)     0.8239156099815332
  (0, 1621)     0.5667125087985602
```

## Average Word2Vector Vectorization

In [0]:
```
# stronging variables into pickle files python: http://www.jessicayung.
com/how-to-use-pickle-to-save-and-load-variables-in-python/
# We should have glove_vectors file for creating below model
with open('drive/My Drive/glove_vectors', 'rb') as f:
    gloveModel = pickle.load(f)
    gloveWords =  set(gloveModel.keys())
```

In [0]:
```
print("Glove vector of sample word: ");
equalsBorder(70);
print(gloveModel['technology']);
equalsBorder(70);
print("Shape of glove vector: ", gloveModel['technology'].shape);
```

```
Glove vector of sample word:
======================================================================
[-0.26078   -0.36898   -0.022831   0.21666    0.16672   -0.20268
 -3.1219     0.33057    0.71512    0.28874    0.074368  -0.033203
  0.23783    0.21052    0.076562   0.13007   -0.31706   -0.45888
 -0.45463   -0.13191    0.49761    0.072704   0.16811    0.18846
 -0.16688   -0.21973    0.08575   -0.19577   -0.2101    -0.32436
 -0.56336    0.077996  -0.22758   -0.66569    0.14824    0.038945
  0.50881   -0.1352     0.49966   -0.4401    -0.022335  -0.22744
  0.22086    0.21865    0.36647    0.30495   -0.16565    0.038759
  0.28108   -0.2167     0.12453    0.65401    0.34584   -0.2557
 -0.046363  -0.31111   -0.020936  -0.17122   -0.77114    0.29289
 -0.14625    0.39541   -0.078938   0.051127   0.15076    0.085126
  0.183     -0.06755    0.26312    0.0087276  0.0066415  0.37033
  0.03496   -0.12627   -0.052626  -0.34897    0.14672    0.14799
 -0.21821   -0.042785   0.2661    -1.1105     0.31789    0.27278
  0.054468  -0.27458    0.42732   -0.44101   -0.19302   -0.32948
  0.61501   -0.22301   -0.36354   -0.34983   -0.16125   -0.17195
 -3.363      0.45146   -0.13753    0.31107    0.2061     0.33063
```

```
 0.45879    0.24256    0.042342   0.074837  -0.12869    0.12066
 0.42843   -0.4704    -0.18937    0.32685    0.26079    0.20518
-0.18432   -0.47658    0.69193    0.18731   -0.12516    0.35447
-0.1969    -0.58981   -0.88914    0.5176     0.13177   -0.078557
 0.032963  -0.19411    0.15109    0.10547   -0.1113    -0.61533
 0.0948    -0.3393    -0.20071   -0.30197    0.29531    0.28017
 0.16049    0.25294   -0.44266   -0.39412    0.13486    0.25178
-0.044114   1.1519     0.32234   -0.34323   -0.10713   -0.15616
 0.031206   0.46636   -0.52761   -0.39296   -0.068424  -0.04072
 0.41508   -0.34564    0.71001   -0.364      0.2996     0.032281
 0.34035    0.23452    0.78342    0.48045   -0.1609     0.40102
-0.071795  -0.16531    0.082153   0.52065    0.24194    0.17113
 0.33552   -0.15725   -0.38984    0.59337   -0.19388   -0.39864
-0.47901    1.0835     0.24473    0.41309    0.64952    0.46846
 0.024386  -0.72087   -0.095061   0.10095   -0.025229   0.29435
-0.57696    0.53166   -0.0058338 -0.3304     0.19661   -0.085206
 0.34225    0.56262    0.19924   -0.027111  -0.44567    0.17266
 0.20887   -0.40702    0.63954    0.50708   -0.31862   -0.39602
-0.1714    -0.040006  -0.45077   -0.32482   -0.0316     0.54908
-0.1121     0.12951   -0.33577   -0.52768   -0.44592   -0.45388
 0.66145    0.33023   -1.9089     0.5318     0.21626   -0.13152
 0.48258    0.68028   -0.84115   -0.51165    0.40017    0.17233
-0.033749   0.045275   0.37398   -0.18252    0.19877    0.1511
 0.029803   0.16657   -0.12987   -0.50489    0.55311   -0.22504
 0.13085   -0.78459    0.36481   -0.27472    0.031805   0.53052
-0.20078    0.46392   -0.63554    0.040289  -0.19142   -0.0097011
 0.068084  -0.10602    0.25567    0.096125  -0.10046    0.15016
-0.26733   -0.26494    0.057888   0.062678  -0.11596    0.28115
 0.25375   -0.17954    0.20615    0.24189    0.062696   0.27719
-0.42601   -0.28619   -0.44697   -0.082253  -0.73415   -0.20675
-0.60289   -0.06728    0.15666   -0.042614   0.41368   -0.17367
-0.54012    0.23883    0.23075    0.13608   -0.058634  -0.089705
 0.18469    0.023634   0.16178    0.23384    0.24267    0.091846 ]
==========================================================================
Shape of glove vector:  (300,)
```

In [0]:
```
def getWord2VecVectors(texts):
    word2VecTextsVectors = [];
    for preProcessedText in tqdm(texts):
```

```
        word2VecTextVector = np.zeros(300);
        numberOfWordsInText = 0;
        for word in preProcessedText.split():
            if word in gloveWords:
                word2VecTextVector += gloveModel[word];
                numberOfWordsInText += 1;
        if numberOfWordsInText != 0:
            word2VecTextVector = word2VecTextVector / numberOfWordsInTe
xt;
        word2VecTextsVectors.append(word2VecTextVector);
    return word2VecTextsVectors;
```

**1. Vectorizing project_essay**

In [0]:
```
word2VecEssaysVectors = getWord2VecVectors(preProcessedEssaysWithoutSto
pWords);
```

In [0]:
```
print("Shape of Word2Vec vectorization matrix of essays: {},{}".format(
len(word2VecEssaysVectors), len(word2VecEssaysVectors[0])));
equalsBorder(70);
print("Sample essay: ");
equalsBorder(70);
print(preProcessedEssaysWithoutStopWords[0]);
equalsBorder(70);
print("Word2Vec vector of sample essay: ");
equalsBorder(70);
print(word2VecEssaysVectors[0]);
```

```
Shape of Word2Vec vectorization matrix of essays: 90848,300
======================================================================
Sample essay:
======================================================================
teach kindergarten urban midwestern school madison wisconsin students d
iverse city neighborhoods backgrounds title 1 school high percentage st
udents come low income households year large class 20 kindergarten stud
ents job ensure students meet challenging academic standards provide lo
```

```
ve learning last lifetime expectations standards kindergarten set high
day one kindergartners today learning read comprehend texts writing per
sonal narratives well informative opinion pieces adding subtracting sol
ving complex story problems learning amazing 5 6 year old children take
s great deal focus engagement differentiated instruction want provide f
un engaging learning environment gives students choice kind seating lea
rning space works needs flexible seating learning spaces classrooms pro
vides kids variety fun engaging choices independent choice helps young
children work collaboratively communicate engage critical thinking whet
her building community morning meetings day learning new concept whole
group kindergartners spend lot time sitting rug want learning time high
ly effective want students happy comfortable engaged come rug learn ben
ches versatile durable perfect little ones move around use see fit kids
sit learn bench turn use work surface someone walks classroom want see
kids creating learning environment works students happy engaged learnin
g rug nannan
==========================================================================
Word2Vec vector of sample essay:
==========================================================================
[ 2.17707912e-03  4.30569574e-02 -3.24074478e-02 -1.05464020e-01
  4.90164698e-02 -5.45596390e-02 -2.94726808e+00  2.04100131e-02
  6.63442242e-03 -2.87220978e-02 -3.04784066e-03 -2.95475026e-02
  3.77005835e-02 -1.06030827e-01 -3.50698791e-02 -3.90703385e-02
  2.29513397e-02 -2.62505742e-02 -4.16030604e-03 -2.55807313e-02
 -5.76231280e-02  4.44215440e-02 -6.32315538e-02 -5.36933473e-02
 -9.74767857e-03 -5.79343008e-02  1.26986475e-01  1.11236462e-02
 -1.81788423e-02 -9.52764643e-02 -2.82883044e-01 -4.16495348e-02
  6.80455181e-02  1.19962808e-01 -6.37358320e-02 -6.03349500e-02
 -4.49860341e-02  3.63737516e-02 -4.54585639e-02 -1.79519714e-02
 -1.19567167e-01  4.16129791e-02  3.78494885e-02 -1.25813758e-01
  7.65183297e-02 -9.53487698e-02  8.81898632e-02 -5.36944560e-02
 -1.50032918e-02 -1.62492030e-01  1.49798901e-02  2.54652835e-02
 -2.23424989e-02 -1.03970703e-02  4.98513275e-02 -6.70185962e-02
  2.64793725e-02 -4.17197308e-03 -6.39391758e-02  8.35040385e-02
 -5.82409221e-02 -7.74809280e-02  4.60176855e-02 -3.59251725e-02
 -9.89083871e-02 -4.18786808e-02  8.82192966e-02 -2.35899280e-02
  1.66398270e-01 -1.13791790e-01 -9.49176364e-02  3.82670555e-02
 -1.77482136e-02 -1.05524769e-01 -6.80576995e-02 -1.96818385e-01
  1.06937126e-02  4.85387967e-03  5.94052187e-02 -3.86374923e-02
```

```
 3.19225571e-02 -2.76290360e-01 -1.22392132e-02 -6.97901912e-02
-8.82657242e-02  7.73700348e-02  1.65614686e-01 -2.89925027e-02
 1.52248579e-01 -6.16283516e-04 -3.10989560e-03  3.25984890e-02
 1.38573819e-02  7.56231780e-02 -3.66657258e-02 -1.34343993e-01
-2.21863588e+00  1.04415730e-01  1.48238346e-01  1.54368930e-01
-5.81557220e-02 -1.48095426e-02  1.40180281e-01 -6.96105951e-02
 1.13854940e-01  7.15323445e-02 -1.29773566e-02 -1.05793423e-01
 2.17443886e-02  4.81857464e-02 -4.63787192e-02 -2.17846170e-02
 5.67572128e-02  2.40450369e-01  2.41049533e-02  7.64587088e-02
-2.45593561e-01  6.06703126e-02  1.37630687e-01  1.39214918e-02
 3.20865890e-02  3.60312104e-02  1.19900342e-01 -1.23319140e-01
 1.00099286e-01  7.25779670e-03  7.75520099e-02 -3.81707834e-02
-4.21087253e-02  1.37521414e-01  5.51964286e-02  4.71947538e-03
-3.09834176e-02 -1.08365562e-01  2.28289835e-03 -1.18564439e-01
 1.03417002e-01 -3.99223352e-02  6.74404253e-02  3.00580747e-01
 5.01157429e-02 -1.07933571e-03  8.26706841e-02  3.02092747e-03
-2.67087126e-02  1.31632056e-02  2.79614000e-02 -5.60139093e-02
 1.38741008e-01 -4.31703569e-02  3.99696544e-02 -1.95282363e-03
 5.48791518e-02 -1.27815368e-02  1.60073462e-03  6.46344725e-03
 2.25129863e-02  2.27978324e-02 -6.81861533e-02  4.15561802e-02
 5.75762247e-02  4.63964181e-02  4.20453368e-03 -9.39134199e-02
-5.86137714e-02  5.29939484e-02 -9.34186368e-02  3.36894285e-02
 4.63504591e-02 -7.92218390e-02 -7.84200330e-03 -5.07480701e-02
-8.59453594e-02 -1.71990119e-01  6.35529682e-02 -4.81186956e-02
-1.34189286e-02  1.33155868e-02 -1.81956391e-01 -5.81012000e-02
 5.92110659e-03  2.37462868e-01 -3.59759357e-02  9.89776500e-03
-6.32349802e-02 -1.80370396e-02 -5.15367000e-02 -3.52633604e-02
 2.94119593e-02  6.72623593e-02 -5.84514516e-02 -5.54773451e-02
-1.54325914e-01  1.61080209e-02  9.84139082e-03 -1.78869418e-02
-3.88647484e-02  1.26517833e-01  1.54825209e-02  7.03101386e-02
 2.03326989e-01 -2.73401429e-02 -6.92491319e-03  9.58065709e-02
-9.28693192e-02  1.08156555e-02  1.12520828e-01 -1.33446509e-01
 1.93755321e-01  4.81129758e-02 -2.13421819e-02 -3.52963797e-03
-4.20169615e-03 -1.80905802e-01 -1.30981424e-01 -9.20938077e-03
-9.19059164e-02 -1.64624203e-02 -5.23227440e-02  1.68430912e-02
-1.18487678e-01 -7.03127978e-02 -8.74793126e-02 -1.38571145e-01
-2.02733731e+00  5.21353401e-02 -5.50138395e-02  1.11370742e-02
-5.03246220e-02 -1.62937180e-01  5.72933346e-02 -6.17547637e-03
-9.13918680e-02 -3.88686819e-02 -1.02863861e-01 -8.08187209e-03
```

```
 7.88880396e-02 -6.03098676e-02  8.63339852e-02  1.05123590e-01
-5.53551854e-03 -3.41195355e-02 -2.13994574e-01  6.89724780e-02
-5.08795060e-02  5.28178022e-02 -2.52496198e-02 -1.08565680e-01
 2.32506154e-03  7.34798297e-03 -8.99563077e-03  1.56501927e-01
 7.53795791e-02 -2.04410967e-02  1.25703581e-01  4.53999132e-02
 1.32489533e-02  2.07894709e-02  1.03909782e-01 -9.06548033e-02
-3.34310632e-03  2.64184978e-02 -3.55721434e-02 -2.73384222e-02
 7.49900621e-02 -7.96857742e-02 -1.68004126e-01  3.88063599e-02
 9.45316859e-02  5.36107841e-02 -1.70442667e-02 -1.19391060e-02
-2.26127800e-01  9.90955934e-02 -6.70596824e-02  7.72681901e-02
 5.30918549e-02  1.84210659e-03 -1.23407607e-01  9.86481374e-02
 1.12378013e-01 -5.14339555e-02 -8.97400351e-02  7.05332297e-02
 7.53460099e-02  1.16287423e-01  1.15706045e-02  2.76136121e-02
 7.71022747e-03 -5.42705607e-02  3.24188714e-02 -1.18067424e-02
-9.76862224e-02 -8.63333500e-02  5.79583841e-02 -1.35902896e-02
 3.83599808e-02  2.36967346e-01  2.09166233e-01 -1.63274231e-02]
```

**2. Vectorizing project_title**

In [0]:
```
word2VecTitlesVectors = getWord2VecVectors(preProcessedProjectTitlesWit
houtStopWords);
```

In [0]:
```
print("Shape of Word2Vec vectorization matrix of project titles: {}, {}
".format(len(word2VecTitlesVectors), len(word2VecTitlesVectors[0])));
equalsBorder(70);
print("Sample title: ");
equalsBorder(70);
print(preProcessedProjectTitlesWithoutStopWords[0]);
equalsBorder(70);
print("Word2Vec vector of sample title: ");
equalsBorder(70);
print(word2VecTitlesVectors[0]);
```

```
Shape of Word2Vec vectorization matrix of project titles: 90848, 300
======================================================================
Sample title:
```

```
==========================================================================
seat learn
==========================================================================
Word2Vec vector of sample title:
==========================================================================
[-1.94029500e-01  1.53639500e-01  1.96764000e-01 -2.42755000e-01
  1.99955000e-01  5.66500000e-03 -2.91110000e+00  1.17545000e-01
 -2.61433650e-01  3.02509500e-01 -6.48190000e-02  7.18250000e-02
 -9.69110000e-02 -2.79840500e-01  1.32719000e-01  1.84890950e-01
 -3.46197000e-01  1.83825000e-01 -2.41170000e-02  1.79105000e-01
  1.91058000e-01  4.18500500e-01 -1.33910000e-01  1.49680500e-01
  2.56777000e-01  2.18790000e-02 -4.07600000e-02  3.38571000e-01
  1.15050000e-02 -3.87830000e-01 -3.84500000e-01  2.91000000e-03
  9.37150000e-02  3.30395000e-01  1.09005000e-01 -3.28325000e-01
 -2.77565000e-01  1.58050000e-02  1.38935500e-01  7.73545000e-01
 -2.86870000e-01  3.76200000e-01 -1.28790000e-01 -2.78925000e-01
  4.47260000e-01 -1.37370000e-01  1.49155500e-01 -1.67900000e-02
  4.79300000e-02 -6.25600000e-02 -2.26150000e-02 -1.32105000e-01
 -2.29925000e-01 -1.05225000e-01  2.99090000e-02 -8.75902000e-02
  1.11375500e-01 -7.13500000e-04 -1.56450000e-01 -3.96600000e-02
  3.70060000e-02 -3.79180000e-01 -8.65350000e-02 -2.04451435e-01
  5.60700000e-02  5.80495000e-02  4.59730000e-01  2.97560000e-01
  1.97924000e-01 -2.07145000e-01 -2.57435000e-01  3.55075000e-01
 -4.04490000e-01 -1.62050000e-01 -3.59380000e-01 -2.34731000e-01
 -2.39000000e-03 -3.33668500e-01  2.71865000e-01  1.61450000e-02
  1.97993000e-01  4.59475000e-01  1.55125000e-01 -1.30940000e-01
 -1.61332000e-01 -1.45664500e-01  6.49400000e-02 -7.38850000e-02
 -2.39850000e-02 -3.70780000e-01  2.48600000e-02  1.42678500e-01
 -1.95443500e-01  3.91604000e-01 -1.64499800e-01 -5.23800000e-01
 -2.07285000e+00  5.71250000e-02  3.37580000e-01  3.78615000e-01
 -1.26248300e-01  1.18550000e-02  1.35070750e-01 -3.02330000e-01
  2.16700000e-02 -3.30605000e-01 -3.93241000e-01  1.75895000e-01
  1.58093355e-01 -4.36513500e-01  2.49725000e-01  3.66550000e-02
 -5.11050000e-02  3.29880000e-01 -5.70290000e-01  1.03100000e-02
 -4.97290000e-01 -1.92905000e-01  4.26500000e-01 -2.38628000e-01
  1.30700000e-01  1.69157500e-01  2.06755000e-01 -4.47045500e-01
  1.37000000e-02 -3.80245000e-01  6.30850000e-02  5.93050000e-02
 -1.72400000e-01 -3.18475000e-02 -8.22035000e-02 -2.78950000e-02
 -1.88678500e-01 -9.28700000e-02 -1.75535000e-01 -8.22900000e-02
```

```
 8.46800000e-02  1.39000000e-01 -1.37650000e-02  1.82525000e-01
 7.94155000e-02 -1.18770000e-01  5.18396000e-01 -8.59850000e-03
-2.75267500e-01  2.40300000e-01 -5.79500000e-04 -9.88060000e-02
-2.86794000e-01  6.75800000e-02 -4.65721500e-01 -2.32835000e-01
-1.49785500e-01 -3.41320000e-01 -1.90295000e-02  9.04725000e-02
-4.37977450e-02 -4.67800000e-02  2.36540000e-01  1.20872500e-01
 1.86803000e-01 -7.64150000e-02  1.19803000e-01  2.97675000e-02
 3.83580000e-02  5.16015000e-01 -3.36687500e-01  2.59225000e-01
-4.12250000e-02 -1.09025000e-01 -7.66640000e-02 -2.66245000e-01
-2.81393000e-01 -3.13041650e-01  1.53622500e-01  5.55260000e-02
-5.70100000e-02  1.23600000e-01 -4.54035000e-01 -2.47883500e-02
 3.24850000e-02  2.98464000e-01  1.61750000e-02  1.31650000e-01
 1.83050000e-02  4.74435000e-01 -2.01968000e-01 -6.68500000e-03
-2.22114500e-01 -2.73210000e-01 -3.66085000e-01 -4.33000000e-01
-1.31961500e-01  3.31995000e-01  1.08335000e-01 -2.34615000e-01
 2.15135000e-01 -1.74940500e-01 -1.34250000e-02  2.52875000e-01
 2.26670000e-01 -2.12874500e-01 -1.12409600e-01  2.46575000e-01
 2.21705000e-01  1.71279000e-01  5.33000000e-02  1.27400000e-02
 5.98130000e-01  3.18650000e-02 -5.32320000e-02 -1.58460000e-01
-1.02787450e-01  1.85935000e-01  5.95770000e-01 -2.59910000e-01
-2.41685000e-02  4.36805000e-01  1.41919000e-01  7.70360000e-02
-2.49044500e-01 -1.45006500e-01 -1.68370000e-01 -2.04049000e-01
-2.52265000e+00 -3.16266500e-01 -2.74880000e-02  3.02005000e-01
-1.19002000e-01 -1.72350000e-02  1.42725000e-02  7.31540000e-02
 1.90308900e-01  2.73453500e-01  6.96650000e-03 -2.77250000e-02
 1.75240000e-01 -5.89000000e-03 -2.43545000e-01  8.47850000e-02
-2.56425000e-01 -1.17241000e-01 -1.90150000e-02  2.28540000e-01
 2.11200500e-01  1.57000000e-02  2.96500000e-02 -1.49181500e-01
 1.92495000e-01  2.58110000e-01 -6.69562000e-02  2.01113500e-01
 3.53010000e-01  5.12100000e-02  4.04907500e-01  1.59490000e-01
 1.59130000e-01  3.95420000e-01 -6.52550000e-02 -3.95100000e-02
 2.23300000e-01 -1.11599500e-01 -2.57310000e-01 -5.77575000e-01
-6.86230000e-02 -1.56995300e-01 -2.92940000e-02  4.43980000e-01
 2.52791000e-01  1.43095000e-01 -4.03450000e-02  1.47130000e-01
-3.83515000e-01  9.43455000e-02 -1.09405000e-01 -9.17070000e-02
-4.16450000e-02  2.93975000e-01 -5.42025000e-02 -2.19307500e-01
-2.24217000e-01 -6.07635000e-02 -3.50550000e-02  9.10920000e-02
-1.91171500e-01  3.03070000e-01 -3.10159000e-01 -2.25115000e-01
-8.90000000e-02  1.70165000e-01  2.96410000e-02 -2.00880000e-01
```

```
  1.32187500e-01  1.29750000e-02 -2.98455000e-01 -1.77665000e-01
  8.09430000e-02  1.84148000e-01  4.29280000e-01  6.01200000e-02]
```

## Tf-Idf Weighted Word2Vec Vectorization

### 1. Vectorizing project_essay

In [0]:
```
# Initializing tfidf vectorizer
tfIdfEssayTempVectorizer = TfidfVectorizer();
# Vectorizing preprocessed essays using tfidf vectorizer initialized ab
ove
tfIdfEssayTempVectorizer.fit(preProcessedEssaysWithoutStopWords);
# Saving dictionary in which each word is key and it's idf is value
tfIdfEssayDictionary = dict(zip(tfIdfEssayTempVectorizer.get_feature_na
mes(), list(tfIdfEssayTempVectorizer.idf_)));
# Creating set of all unique words used by tfidf vectorizer
tfIdfEssayWords = set(tfIdfEssayTempVectorizer.get_feature_names());
```

In [0]:
```
# Creating list to save tf-idf weighted vectors of essays
tfIdfWeightedWord2VecEssaysVectors = [];
# Iterating over each essay
for essay in tqdm(preProcessedEssaysWithoutStopWords):
    # Sum of tf-idf values of all words in a particular essay
    cumulativeSumTfIdfWeightOfEssay = 0;
    # Tf-Idf weighted word2vec vector of a particular essay
    tfIdfWeightedWord2VecEssayVector = np.zeros(300);
    # Splitting essay into list of words
    splittedEssay = essay.split();
    # Iterating over each word
    for word in splittedEssay:
        # Checking if word is in glove words and set of words used by t
fIdf essay vectorizer
        if (word in gloveWords) and (word in tfIdfEssayWords):
            # Tf-Idf value of particular word in essay
            tfIdfValueWord = tfIdfEssayDictionary[word] * (essay.count(
word) / len(splittedEssay));
```

```
            # Making tf-idf weighted word2vec
            tfIdfWeightedWord2VecEssayVector += tfIdfValueWord * gloveM
odel[word];
            # Summing tf-idf weight of word to cumulative sum
            cumulativeSumTfIdfWeightOfEssay += tfIdfValueWord;
    if cumulativeSumTfIdfWeightOfEssay != 0:
        # Taking average of sum of vectors with tf-idf cumulative sum
        tfIdfWeightedWord2VecEssayVector = tfIdfWeightedWord2VecEssayVe
ctor / cumulativeSumTfIdfWeightOfEssay;
    # Appending the above calculated tf-idf weighted vector of particul
ar essay to list of vectors of essays
    tfIdfWeightedWord2VecEssaysVectors.append(tfIdfWeightedWord2VecEssa
yVector);
```

In [0]:

```
print("Shape of Tf-Idf weighted Word2Vec vectorization matrix of projec
t essays: {}, {}".format(len(tfIdfWeightedWord2VecEssaysVectors), len(t
fIdfWeightedWord2VecEssaysVectors[0])));
equalsBorder(70);
print("Sample Essay: ");
equalsBorder(70);
print(preProcessedEssaysWithoutStopWords[0]);
equalsBorder(70);
print("Tf-Idf Weighted Word2Vec vector of sample essay: ");
equalsBorder(70);
print(tfIdfWeightedWord2VecEssaysVectors[0]);
```

```
Shape of Tf-Idf weighted Word2Vec vectorization matrix of project essay
s: 90848, 300
======================================================================
Sample Essay:
======================================================================
teach kindergarten urban midwestern school madison wisconsin students d
iverse city neighborhoods backgrounds title 1 school high percentage st
udents come low income households year large class 20 kindergarten stud
ents job ensure students meet challenging academic standards provide lo
ve learning last lifetime expectations standards kindergarten set high
day one kindergartners today learning read comprehend texts writing per
sonal narratives well informative opinion pieces adding subtracting sol
```

```
ving complex story problems learning amazing 5 6 year old children take
s great deal focus engagement differentiated instruction want provide f
un engaging learning environment gives students choice kind seating lea
rning space works needs flexible seating learning spaces classrooms pro
vides kids variety fun engaging choices independent choice helps young
children work collaboratively communicate engage critical thinking whet
her building community morning meetings day learning new concept whole
group kindergartners spend lot time sitting rug want learning time high
ly effective want students happy comfortable engaged come rug learn ben
ches versatile durable perfect little ones move around use see fit kids
sit learn bench turn use work surface someone walks classroom want see
kids creating learning environment works students happy engaged learnin
g rug nannan
==================================================================
Tf-Idf Weighted Word2Vec vector of sample essay:
==================================================================
[-3.42999576e-02  2.73270677e-02 -2.78123615e-02 -1.39259513e-01
  5.07528733e-02 -8.49010646e-02 -2.91529140e+00 -4.32406682e-02
  8.97247699e-03 -3.20040792e-02 -5.77932179e-03 -4.42248649e-02
  2.37533031e-02 -1.49568972e-01 -2.25534048e-02 -9.51151835e-03
  4.86021942e-02 -3.48334039e-02  2.95882854e-03 -1.67126632e-02
 -6.41209837e-02  4.78336344e-02 -5.61877522e-02 -3.51993825e-02
  8.56460341e-03 -5.23344276e-02  1.43896357e-01  4.05005552e-02
 -2.97551459e-02 -1.38899331e-01 -3.19248097e-01 -7.98267849e-02
  5.83561442e-02  1.49663668e-01 -1.00283341e-01 -1.01205508e-01
 -2.45034790e-02  6.21653094e-02 -4.56017157e-02  3.47307890e-03
 -1.42760775e-01  3.59747486e-02  5.99093871e-03 -1.30315678e-01
  1.24028954e-01 -1.15618631e-01  1.16820557e-01 -3.74212133e-02
  3.35545862e-02 -2.01699738e-01  5.14154160e-02  1.09381536e-02
  8.89991444e-04 -2.32917350e-02  2.58713839e-02 -1.10068413e-01
 -1.98477303e-02  1.10062559e-02 -6.08299558e-02  6.80726469e-02
 -1.08879367e-01 -1.06024384e-01  8.24703027e-02 -4.44437836e-02
 -1.10317888e-01 -2.01471079e-02  1.33474753e-01 -3.08145955e-02
  1.83497085e-01 -1.16252147e-01 -1.13035068e-01  5.14475733e-02
 -2.21902324e-02 -1.48974461e-01 -1.23159631e-01 -1.77429940e-01
  3.56024206e-02 -5.15888327e-03  6.47954873e-02 -4.44658375e-02
  1.92210792e-02 -1.97899896e-01 -3.30007434e-02 -2.64188751e-02
 -1.23653855e-01  9.33551335e-02  1.73016889e-01 -3.55417926e-02
  1.81798656e-01 -3.51616610e-02  1.52972874e-02  6.05876386e-02
```

```
-5.15603255e-03  6.43257354e-02 -5.24859144e-02 -1.48840893e-01
-2.23696522e+00  1.69581284e-01  1.78354670e-01  1.90264715e-01
-5.27492474e-02 -3.20716975e-02  1.44066596e-01 -9.20790455e-02
 1.31726409e-01  1.56987117e-02 -2.81311998e-02 -9.02223663e-02
 2.76860057e-02  3.63596255e-02 -3.32317197e-02 -3.32247275e-02
 4.50170487e-02  3.30140158e-01  7.36301747e-03  8.78431558e-02
-3.05995440e-01  7.07275043e-02  1.44587576e-01 -3.30209013e-02
 5.02179899e-03  7.28690041e-02  1.37278114e-01 -1.02941603e-01
 1.05824073e-01  2.17123458e-03  7.63048655e-02 -2.96452901e-02
-1.93814100e-02  1.57312065e-01  8.58160392e-02 -7.80811077e-03
-4.34794029e-02 -1.41081723e-01  2.76627409e-02 -1.60300868e-01
 1.21032934e-01 -4.19907739e-02  3.28821902e-02  3.00880947e-01
 4.89612304e-02  1.94723233e-03  1.04301264e-01  4.38723526e-03
-1.30920157e-02  4.67545964e-02  3.67692703e-02 -7.97684738e-02
 1.28110935e-01 -7.11008827e-02  2.58526634e-02  1.87191322e-02
 1.93469791e-02 -3.27573662e-02 -1.88011274e-02  2.33793034e-02
 1.65554759e-02  5.70609376e-02 -7.58573325e-02  6.00533628e-02
 3.45684305e-02  7.07010167e-02  1.52412937e-02 -6.66713932e-02
-9.16870674e-02  7.20469257e-02 -1.29168323e-01  4.74791612e-02
 4.39814699e-02 -5.25245967e-02 -2.26582962e-03 -6.65586391e-02
-9.80908955e-02 -1.85116549e-01  6.65251210e-02 -4.97456056e-02
-2.48820610e-02  3.43397180e-02 -2.48354693e-01 -5.84337906e-02
-5.92108957e-02  2.62328902e-01 -5.15194240e-02  4.50703144e-02
-7.23880163e-02 -2.46468019e-02 -5.33878168e-02 -2.90574513e-02
-7.69431540e-03  6.22765101e-02 -5.23762999e-02 -6.83526496e-02
-1.68750576e-01  3.87141932e-02  3.49403304e-02 -1.40863581e-02
-3.32514517e-02  1.10184779e-01  4.44687740e-02  1.04505308e-01
 2.44617896e-01 -1.49790258e-02 -7.82310674e-03  1.15684217e-01
-1.12771527e-01 -2.93169481e-02  1.58392422e-01 -1.87893609e-01
 2.67490480e-01  5.46745816e-02 -1.61810859e-02  2.14394119e-02
-1.42824274e-02 -1.70079100e-01 -1.29275778e-01 -2.93106726e-02
-1.05685726e-01  1.07030142e-02 -4.98104096e-02  3.06392930e-02
-1.54855783e-01 -9.04866704e-02 -1.24584212e-01 -1.76764248e-01
-1.90448451e+00  4.13671847e-02 -1.00852495e-01  5.61933686e-02
-4.04125968e-02 -1.80411491e-01  8.13714093e-02 -1.28955349e-02
-7.36473286e-02 -1.19279706e-02 -1.15178470e-01 -5.88091454e-02
 1.15936780e-01 -8.24687283e-02  6.21737783e-02  1.46217227e-01
-1.38152910e-02 -6.06084683e-02 -2.33794121e-01  3.19363796e-02
-3.63234088e-02  8.26871696e-02 -3.02437184e-02 -1.03143806e-01
```

```
 4.33859805e-02  4.78017233e-03 -2.17299672e-02  2.15516585e-01
 1.04599269e-01 -4.60156472e-03  1.90314436e-01  5.45380260e-02
 6.73526613e-03  2.97718701e-02  1.28675246e-01 -1.30891879e-01
 3.26965058e-02  4.11610092e-02 -1.72395779e-02 -4.21028918e-04
 4.08393521e-02 -1.08062795e-01 -1.82068406e-01  3.95993120e-02
 1.10639991e-01  7.35224680e-02 -5.44884201e-02 -5.07513023e-02
-2.51169581e-01  1.08845425e-01 -1.02257576e-01  1.02094804e-01
 6.17550335e-02  3.16492381e-02 -1.23372669e-01  1.02711903e-01
 5.70866173e-02 -5.62677571e-02 -1.48408778e-01  8.21982740e-02
 6.27577593e-02  1.29731525e-01 -1.14933307e-02  1.95529920e-02
-1.09940138e-03 -6.22525728e-02  1.04202523e-03 -1.95784974e-02
-7.99433412e-02 -4.03035160e-02  4.82527556e-02 -7.52904008e-03
 8.56012309e-02  2.70221122e-01  2.79138826e-01 -2.63031072e-02]
```

**2. Vectorizing project_title**

In [0]:
```
# Initializing tfidf vectorizer
tfIdfTitleTempVectorizer = TfidfVectorizer();
# Vectorizing preprocessed titles using tfidf vectorizer initialized ab
ove
tfIdfTitleTempVectorizer.fit(preProcessedProjectTitlesWithoutStopWords
);
# Saving dictionary in which each word is key and it's idf is value
tfIdfTitleDictionary = dict(zip(tfIdfTitleTempVectorizer.get_feature_na
mes(), list(tfIdfTitleTempVectorizer.idf_)));
# Creating set of all unique words used by tfidf vectorizer
tfIdfTitleWords = set(tfIdfTitleTempVectorizer.get_feature_names());
```

In [0]:
```
# Creating list to save tf-idf weighted vectors of project titles
tfIdfWeightedWord2VecTitlesVectors = [];
# Iterating over each title
for title in tqdm(preProcessedProjectTitlesWithoutStopWords):
    # Sum of tf-idf values of all words in a particular project title
    cumulativeSumTfIdfWeightOfTitle = 0;
    # Tf-Idf weighted word2vec vector of a particular project title
    tfIdfWeightedWord2VecTitleVector = np.zeros(300);
    # Splitting title into list of words
```

```
    splittedTitle = title.split();
    # Iterating over each word
    for word in splittedTitle:
        # Checking if word is in glove words and set of words used by t
fIdf title vectorizer
        if (word in gloveWords) and (word in tfIdfTitleWords):
            # Tf-Idf value of particular word in title
            tfIdfValueWord = tfIdfTitleDictionary[word] * (title.count(
word) / len(splittedTitle));
            # Making tf-idf weighted word2vec
            tfIdfWeightedWord2VecTitleVector += tfIdfValueWord * gloveM
odel[word];
            # Summing tf-idf weight of word to cumulative sum
            cumulativeSumTfIdfWeightOfTitle += tfIdfValueWord;
    if cumulativeSumTfIdfWeightOfTitle != 0:
        # Taking average of sum of vectors with tf-idf cumulative sum
        tfIdfWeightedWord2VecTitleVector = tfIdfWeightedWord2VecTitleVe
ctor / cumulativeSumTfIdfWeightOfTitle;
    # Appending the above calculated tf-idf weighted vector of particul
ar title to list of vectors of project titles
    tfIdfWeightedWord2VecTitlesVectors.append(tfIdfWeightedWord2VecTitl
eVector);
```

In [0]:

```
print("Shape of Tf-Idf weighted Word2Vec vectorization matrix of projec
t titles: {}, {}".format(len(tfIdfWeightedWord2VecTitlesVectors), len(t
fIdfWeightedWord2VecTitlesVectors[0])));
equalsBorder(70);
print("Sample Title: ");
equalsBorder(70);
print(preProcessedProjectTitlesWithoutStopWords[0]);
equalsBorder(70);
print("Tf-Idf Weighted Word2Vec vector of sample title: ");
equalsBorder(70);
print(tfIdfWeightedWord2VecTitlesVectors[0]);
```

```
Shape of Tf-Idf weighted Word2Vec vectorization matrix of project title
s: 90848, 300
======================================================================
```

```
Sample Title:
==========================================================================
seat learn
==========================================================================
Tf-Idf Weighted Word2Vec vector of sample title:
==========================================================================
[-2.26862737e-01  1.32556430e-01  2.22777128e-01 -2.94672688e-01
  2.78569038e-01 -3.66590914e-02 -2.75433246e+00  2.25396665e-01
 -3.11450254e-01  3.52682233e-01 -8.58152339e-02  1.04261417e-01
 -1.06178891e-01 -2.40792972e-01  1.03823354e-01  2.18315774e-01
 -4.04818923e-01  2.45752432e-01 -1.75035775e-02  2.31914169e-01
  2.24501865e-01  4.79884998e-01 -1.35876068e-01  1.66616702e-01
  2.21457510e-01  2.20704121e-03 -6.83219381e-02  3.93779770e-01
  6.77506260e-02 -4.18445539e-01 -3.59150119e-01  6.56743517e-02
  6.42619009e-02  3.26604355e-01  1.50069550e-01 -3.21586178e-01
 -2.79409922e-01  6.52294995e-02  1.47664341e-01  8.40767683e-01
 -3.18954079e-01  4.26405933e-01 -4.91207956e-02 -2.86523861e-01
  4.54462133e-01 -1.62435051e-01  1.62428676e-01 -4.09728171e-02
 -1.10538792e-02  6.07695930e-02 -5.98602375e-02 -8.04203550e-02
 -2.52538477e-01 -6.62448882e-02  4.16828415e-02 -7.22630473e-02
  1.17766515e-01 -1.07309199e-02 -1.02679992e-01 -1.25770944e-02
  3.80530281e-02 -3.70760866e-01 -1.55070861e-01 -1.66748700e-01
  1.43235415e-01  6.46659743e-02  4.59247268e-01  3.73395095e-01
  2.26822420e-01 -1.88125191e-01 -2.42534131e-01  3.50429865e-01
 -4.36287399e-01 -1.34353045e-01 -3.39859889e-01 -2.63729851e-01
  2.81238139e-02 -4.08843584e-01  2.60629931e-01  4.65025273e-02
  2.17484813e-01  5.23866028e-01  1.56542677e-01 -2.33125580e-01
 -1.73296345e-01 -1.88820980e-01 -4.50306322e-03 -2.62887043e-04
 -6.23547616e-02 -4.04663687e-01  4.85027496e-02  1.27358431e-01
 -1.66987029e-01  4.47171027e-01 -1.33146255e-01 -5.39158632e-01
 -1.99127576e+00 -2.19532744e-02  2.99007713e-01  4.23743004e-01
 -1.03391293e-01 -1.82010512e-02  1.11034186e-01 -3.06003199e-01
 -1.09976379e-03 -3.34854332e-01 -4.83088072e-01  1.93496207e-01
  1.87465378e-01 -5.07664897e-01  2.76255816e-01  6.86530748e-02
 -7.72296811e-03  2.49091820e-01 -5.75964408e-01  2.33827778e-02
 -5.12108564e-01 -2.00783142e-01  5.33237314e-01 -2.73342503e-01
  1.29481149e-01  1.53469186e-01  1.92986053e-01 -5.24653293e-01
  5.73788841e-02 -4.16075335e-01  1.18775762e-01  9.67499885e-02
 -2.24407391e-01 -4.53884904e-02 -6.90760681e-02 -5.75441510e-02
```

```
-2.32041389e-01 -1.33556318e-01 -3.12591000e-01  1.63555488e-02
 3.27962032e-03  1.35093758e-01  2.86885597e-02  4.69689816e-02
 7.62010811e-02 -1.74977710e-01  5.99120186e-01  3.91007372e-03
-2.38586837e-01  3.54886788e-01 -5.29297612e-03 -7.36532808e-02
-3.33457312e-01  1.14490041e-01 -5.34903500e-01 -2.83274900e-01
-1.74192945e-01 -3.48133728e-01 -3.41847744e-02  7.79126939e-02
-3.58625896e-02 -7.67981355e-02  2.90968447e-01  1.57857414e-01
 1.54965466e-01 -1.29607025e-01  1.58783482e-01  8.07001064e-03
 2.61117844e-02  5.34838757e-01 -2.70308673e-01  2.31419847e-01
-9.31741301e-02 -1.44589000e-01 -4.66891439e-02 -2.85386879e-01
-3.43165440e-01 -2.54816390e-01  1.92930444e-01  8.12835208e-02
-1.25857488e-02  8.16467426e-02 -3.92270328e-01 -2.78361525e-02
 8.08497095e-02  2.33819400e-01  5.40564814e-02  2.00490111e-01
 4.49708330e-02  5.02537930e-01 -1.55325773e-01 -4.80806192e-02
-2.51645373e-01 -2.89504502e-01 -3.74590138e-01 -4.85876678e-01
-1.72479232e-01  3.65707604e-01  1.23849919e-01 -2.58275320e-01
 1.97576333e-01 -1.36986054e-01 -7.24116535e-02  2.53669380e-01
 2.22197797e-01 -2.63364431e-01 -9.04998012e-02  2.34093337e-01
 3.21522236e-01  2.14887786e-01 -5.13641353e-03  4.30152224e-02
 6.44048685e-01  5.75957956e-03 -6.72937302e-02 -1.56549419e-01
-1.20219895e-01  2.91463635e-01  6.34886053e-01 -2.66551721e-01
-4.11272669e-02  4.82214135e-01  1.60205649e-01  1.08715768e-01
-2.20712133e-01 -1.20105412e-01 -5.14028461e-02 -2.35086420e-01
-2.66524077e+00 -3.79758371e-01 -2.26026086e-02  2.82400735e-01
-1.44456935e-01 -9.95962190e-02  3.34647790e-02  6.81017795e-02
 2.26784005e-01  2.38825093e-01  2.34578866e-02  5.16085697e-03
 1.20245592e-01  3.30096566e-02 -2.24821119e-01  4.95927589e-02
-2.84459497e-01 -1.28308500e-01  5.72797073e-02  3.19815107e-01
 2.36039814e-01 -8.42548119e-03  1.00846433e-01 -1.72331993e-01
 1.81872131e-01  2.61043380e-01 -8.04419445e-02  2.20834564e-01
 3.85715526e-01  8.78458120e-02  3.20343934e-01  2.26930005e-01
 1.63367310e-01  4.33820279e-01 -1.18040125e-01  8.39694677e-03
 2.23653263e-01 -9.28875482e-02 -3.07282890e-01 -6.07383212e-01
-4.47570102e-02 -1.28846186e-01 -1.62743042e-02  4.95750649e-01
 2.10188737e-01  7.47071029e-02 -3.27292724e-04  1.55758133e-01
-4.01471320e-01  1.08475016e-01 -7.65968240e-02 -1.12263412e-01
-1.21975417e-01  3.17331995e-01 -5.80252347e-02 -1.73723121e-01
-1.97179778e-01 -8.39561625e-02  2.66800793e-02  9.08889198e-02
-2.09910825e-01  2.94796980e-01 -2.67426899e-01 -2.34957360e-01
```

```
 -4.76330487e-02  1.88903678e-01  2.71084163e-02 -1.37631067e-01
  1.05187362e-01 -2.42517420e-02 -2.86949898e-01 -1.80686234e-01
  5.74783617e-02  1.34150696e-01  3.83686836e-01  2.53836720e-02]
```

## Method for vectorizing unknown essays using our training data tf-idf weighted model

In [0]:

```
def getAvgTfIdfEssayVectors(arrayOfTexts):
    # Creating list to save tf-idf weighted vectors of essays
    tfIdfWeightedWord2VecEssaysVectors = [];
    # Iterating over each essay
    for essay in tqdm(arrayOfTexts):
        # Sum of tf-idf values of all words in a particular essay
        cumulativeSumTfIdfWeightOfEssay = 0;
        # Tf-Idf weighted word2vec vector of a particular essay
        tfIdfWeightedWord2VecEssayVector = np.zeros(300);
        # Splitting essay into list of words
        splittedEssay = essay.split();
        # Iterating over each word
        for word in splittedEssay:
            # Checking if word is in glove words and set of words used
 by tfIdf essay vectorizer
            if (word in gloveWords) and (word in tfIdfEssayWords):
                # Tf-Idf value of particular word in essay
                tfIdfValueWord = tfIdfEssayDictionary[word] * (essay.co
unt(word) / len(splittedEssay));
                # Making tf-idf weighted word2vec
                tfIdfWeightedWord2VecEssayVector += tfIdfValueWord * gl
oveModel[word];
                # Summing tf-idf weight of word to cumulative sum
                cumulativeSumTfIdfWeightOfEssay += tfIdfValueWord;
        if cumulativeSumTfIdfWeightOfEssay != 0:
            # Taking average of sum of vectors with tf-idf cumulative s
um
            tfIdfWeightedWord2VecEssayVector = tfIdfWeightedWord2VecEss
ayVector / cumulativeSumTfIdfWeightOfEssay;
        # Appending the above calculated tf-idf weighted vector of part
icular essay to list of vectors of essays
```

```
        tfIdfWeightedWord2VecEssaysVectors.append(tfIdfWeightedWord2Vec
EssayVector);
    return tfIdfWeightedWord2VecEssaysVectors;
```

## Method for vectorizing unknown titles using our training data tf-idf weighted model

In [0]:

```
def getAvgTfIdfTitleVectors(arrayOfTexts):
    # Creating list to save tf-idf weighted vectors of project titles
    tfIdfWeightedWord2VecTitlesVectors = [];
    # Iterating over each title
    for title in tqdm(arrayOfTexts):
        # Sum of tf-idf values of all words in a particular project tit
le
        cumulativeSumTfIdfWeightOfTitle = 0;
        # Tf-Idf weighted word2vec vector of a particular project title
        tfIdfWeightedWord2VecTitleVector = np.zeros(300);
        # Splitting title into list of words
        splittedTitle = title.split();
        # Iterating over each word
        for word in splittedTitle:
            # Checking if word is in glove words and set of words used
 by tfIdf title vectorizer
            if (word in gloveWords) and (word in tfIdfTitleWords):
                # Tf-Idf value of particular word in title
                tfIdfValueWord = tfIdfTitleDictionary[word] * (title.co
unt(word) / len(splittedTitle));
                # Making tf-idf weighted word2vec
                tfIdfWeightedWord2VecTitleVector += tfIdfValueWord * gl
oveModel[word];
                # Summing tf-idf weight of word to cumulative sum
                cumulativeSumTfIdfWeightOfTitle += tfIdfValueWord;
        if cumulativeSumTfIdfWeightOfTitle != 0:
            # Taking average of sum of vectors with tf-idf cumulative s
um
            tfIdfWeightedWord2VecTitleVector = tfIdfWeightedWord2VecTit
leVector / cumulativeSumTfIdfWeightOfTitle;
```

```
        # Appending the above calculated tf-idf weighted vector of part
icular title to list of vectors of project titles
        tfIdfWeightedWord2VecTitlesVectors.append(tfIdfWeightedWord2Vec
TitleVector);
    return tfIdfWeightedWord2VecTitlesVectors;
```

## Vectorizing numerical features

### 1. Vectorizing price

In [0]:
```
# Standardizing the price data using StandardScaler(Uses mean and std f
or standardization)
priceScaler = MinMaxScaler();
priceScaler.fit(trainingData['price'].values.reshape(-1, 1));
priceStandardized = priceScaler.transform(trainingData['price'].values.
reshape(-1, 1));
```

In [0]:
```
print("Shape of standardized matrix of prices: ", priceStandardized.sha
pe);
equalsBorder(70);
print("Sample original prices: ");
equalsBorder(70);
print(trainingData['price'].values[0:5]);
print("Sample standardized prices: ");
equalsBorder(70);
print(priceStandardized[0:5]);
```

```
Shape of standardized matrix of prices:  (90848, 1)
======================================================================
Sample original prices:
======================================================================
[ 87.07 177.41 257.94 356.73 806.65]
Sample standardized prices:
======================================================================
[[0.00864243]
 [0.01767793]
```

```
 [0.02573227]
 [0.03561291]
 [0.08061238]]
```

#### 2. Vectorizing quantity

In [0]:
```
# Standardizing the quantity data using StandardScaler(Uses mean and st
d for standardization)
quantityScaler = MinMaxScaler();
quantityScaler.fit(trainingData['quantity'].values.reshape(-1, 1));
quantityStandardized = quantityScaler.transform(trainingData['quantity'
].values.reshape(-1, 1));
```

In [0]:
```
print("Shape of standardized matrix of quantities: ", quantityStandardi
zed.shape);
equalsBorder(70);
print("Sample original quantities: ");
equalsBorder(70);
print(trainingData['quantity'].values[0:5]);
print("Sample standardized quantities: ");
equalsBorder(70);
print(quantityStandardized[0:5]);
```

```
Shape of standardized matrix of quantities:  (90848, 1)
======================================================================
Sample original quantities:
======================================================================
[ 4  2  6 22 14]
Sample standardized quantities:
======================================================================
[[0.00322928]
 [0.00107643]
 [0.00538213]
 [0.02260495]
 [0.01399354]]
```

#### 3. Vectorizing teacher_number_of_previously_posted_projects

In [0]:
```
# Standardizing the teacher_number_of_previously_posted_projects data u
sing StandardScaler(Uses mean and std for standardization)
previouslyPostedScaler = MinMaxScaler();
previouslyPostedScaler.fit(trainingData['teacher_number_of_previously_p
osted_projects'].values.reshape(-1, 1));
previouslyPostedStandardized = previouslyPostedScaler.transform(trainin
gData['teacher_number_of_previously_posted_projects'].values.reshape(-1
, 1));
```

In [0]:
```
print("Shape of standardized matrix of teacher_number_of_previously_pos
ted_projects: ", previouslyPostedStandardized.shape);
equalsBorder(70);
print("Sample original quantities: ");
equalsBorder(70);
print(trainingData['teacher_number_of_previously_posted_projects'].valu
es[0:5]);
print("Sample standardized teacher_number_of_previously_posted_project
s: ");
equalsBorder(70);
print(previouslyPostedStandardized[0:5]);
```

```
Shape of standardized matrix of teacher_number_of_previously_posted_pro
jects:  (90848, 1)
======================================================================
Sample original quantities:
======================================================================
[ 2  2  0  3 27]
Sample standardized teacher_number_of_previously_posted_projects:
======================================================================
[[0.00443459]
 [0.00443459]
 [0.        ]
 [0.00665188]
 [0.05986696]]
```

In [0]:
```
numberOfPoints = previouslyPostedStandardized.shape[0];
# Categorical data
```

```
categoriesVectorsSub = categoriesVectors[0:numberOfPoints];
subCategoriesVectorsSub = subCategoriesVectors[0:numberOfPoints];
teacherPrefixVectorsSub = teacherPrefixVectors[0:numberOfPoints];
schoolStateVectorsSub = schoolStateVectors[0:numberOfPoints];
projectGradeVectorsSub = projectGradeVectors[0:numberOfPoints];

# Text data
bowEssayModelSub = bowEssayModel[0:numberOfPoints];
bowTitleModelSub = bowTitleModel[0:numberOfPoints];
tfIdfEssayModelSub = tfIdfEssayModel[0:numberOfPoints];
tfIdfTitleModelSub = tfIdfTitleModel[0:numberOfPoints];

# Numerical data
priceStandardizedSub = priceStandardized[0:numberOfPoints];
quantityStandardizedSub = quantityStandardized[0:numberOfPoints];
previouslyPostedStandardizedSub = previouslyPostedStandardized[0:number
OfPoints];
```

In [11]:
```
logisticResultsDataFrame = pd.DataFrame(columns =  ['Vectorizer', 'Mode
l', 'Hyper Parameter - C', 'AUC', 'Data']);
logisticResultsDataFrame
```

Out[11]:

| | Vectorizer | Model | Hyper Parameter - C | AUC | Data |
|---|---|---|---|---|---|

## Preparing cross validate data for analysis

In [0]:
```
# Test data categorical features transformation
categoriesTransformedCrossValidateData = subjectsCategoriesVectorizer.t
ransform(crossValidateData['cleaned_categories']);
subCategoriesTransformedCrossValidateData = subjectsSubCategoriesVector
izer.transform(crossValidateData['cleaned_sub_categories']);
teacherPrefixTransformedCrossValidateData = teacherPrefixVectorizer.tra
nsform(crossValidateData['teacher_prefix']);
schoolStateTransformedCrossValidateData = schoolStateVectorizer.transfo
rm(crossValidateData['school_state']);
projectGradeTransformedCrossValidateData = projectGradeVectorizer.trans
```

```
form(crossValidateData['project_grade_category']);

# Test data text features transformation
preProcessedEssaysTemp = preProcessingWithAndWithoutStopWords(crossVali
dateData['project_essay'])[1];
preProcessedTitlesTemp = preProcessingWithAndWithoutStopWords(crossVali
dateData['project_title'])[1];
bowEssayTransformedCrossValidateData = bowEssayVectorizer.transform(pre
ProcessedEssaysTemp);
bowTitleTransformedCrossValidateData = bowTitleVectorizer.transform(pre
ProcessedTitlesTemp);
tfIdfEssayTransformedCrossValidateData = tfIdfEssayVectorizer.transform
(preProcessedEssaysTemp);
tfIdfTitleTransformedCrossValidateData = tfIdfTitleVectorizer.transform
(preProcessedTitlesTemp);
avgWord2VecEssayTransformedCrossValidateData = getWord2VecVectors(prePr
ocessedEssaysTemp);
avgWord2VecTitleTransformedCrossValidateData = getWord2VecVectors(prePr
ocessedTitlesTemp);
tfIdfWeightedWord2VecEssayTransformedCrossValidateData = getAvgTfIdfEss
ayVectors(preProcessedEssaysTemp);
tfIdfWeightedWord2VecTitleTransformedCrossValidateData = getAvgTfIdfTit
leVectors(preProcessedTitlesTemp);


# Test data numerical features transformation
priceTransformedCrossValidateData = priceScaler.transform(crossValidate
Data['price'].values.reshape(-1, 1));
quantityTransformedCrossValidateData = quantityScaler.transform(crossVa
lidateData['quantity'].values.reshape(-1, 1));
previouslyPostedTransformedCrossValidateData = previouslyPostedScaler.t
ransform(crossValidateData['teacher_number_of_previously_posted_project
s'].values.reshape(-1, 1));
```

## Preparing Test data for analysis

In [0]:
```
# Test data categorical features transformation
categoriesTransformedTestData = subjectsCategoriesVectorizer.transform(
testData['cleaned_categories']);
subCategoriesTransformedTestData = subjectsSubCategoriesVectorizer.tran
sform(testData['cleaned_sub_categories']);
teacherPrefixTransformedTestData = teacherPrefixVectorizer.transform(te
stData['teacher_prefix']);
schoolStateTransformedTestData = schoolStateVectorizer.transform(testDa
ta['school_state']);
projectGradeTransformedTestData = projectGradeVectorizer.transform(test
Data['project_grade_category']);

# Test data text features transformation
preProcessedEssaysTemp = preProcessingWithAndWithoutStopWords(testData[
'project_essay'])[1];
preProcessedTitlesTemp = preProcessingWithAndWithoutStopWords(testData[
'project_title'])[1];
bowEssayTransformedTestData = bowEssayVectorizer.transform(preProcessed
EssaysTemp);
bowTitleTransformedTestData = bowTitleVectorizer.transform(preProcessed
TitlesTemp);
tfIdfEssayTransformedTestData = tfIdfEssayVectorizer.transform(preProce
ssedEssaysTemp);
tfIdfTitleTransformedTestData = tfIdfTitleVectorizer.transform(preProce
ssedTitlesTemp);
avgWord2VecEssayTransformedTestData = getWord2VecVectors(preProcessedEs
saysTemp);
avgWord2VecTitleTransformedTestData = getWord2VecVectors(preProcessedTi
tlesTemp);
tfIdfWeightedWord2VecEssayTransformedTestData = getAvgTfIdfEssayVectors
(preProcessedEssaysTemp);
tfIdfWeightedWord2VecTitleTransformedTestData = getAvgTfIdfTitleVectors
(preProcessedTitlesTemp);
```

```
# Test data numerical features transformation
priceTransformedTestData = priceScaler.transform(testData['price'].valu
es.reshape(-1, 1));
quantityTransformedTestData = quantityScaler.transform(testData['quanti
ty'].values.reshape(-1, 1));
previouslyPostedTransformedTestData = previouslyPostedScaler.transform(
testData['teacher_number_of_previously_posted_projects'].values.reshape
(-1, 1));
```

## Classification using imbalanced data containing all features by logistic regression(l1 regularizer)

In [0]:
```
techniques = ['Bag of words', 'Tf-Idf', 'Average Word2Vector', 'Tf-Idf
 Weighted Word2Vector'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
                                 teacherPrefixVectorsSub,\
                                 schoolStateVectorsSub,\
                                 projectGradeVectorsSub,\
                                 priceStandardizedSub,\
                                 previouslyPostedStandardizedSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                      subCategoriesTransformedCross
ValidateData,\
                                      teacherPrefixTransformedCross
ValidateData,\
```

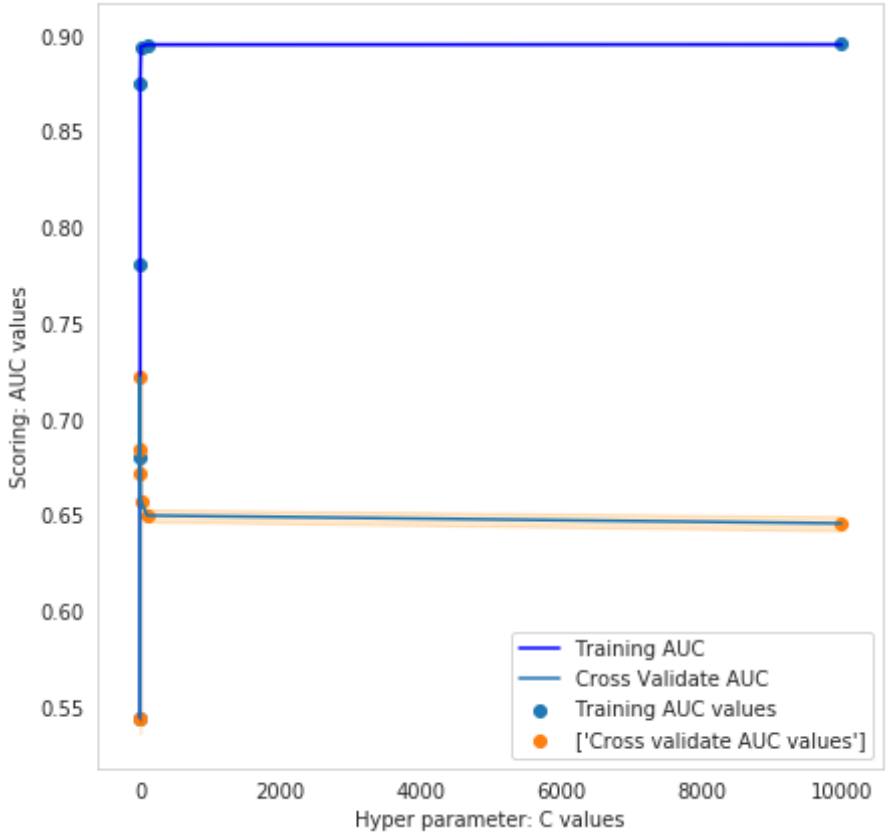
```
                                                schoolStateTransformedCrossVa
lidateData,\
                                                projectGradeTransformedCrossV
alidateData,\
                                                priceTransformedCrossValidate
Data,\
                                                previouslyPostedTransformedCr
ossValidateData));
    testMergedData = hstack((categoriesTransformedTestData,\
                                                subCategoriesTransformedTestD
ata,\
                                                teacherPrefixTransformedTestD
ata,\
                                                schoolStateTransformedTestDat
a,\
                                                projectGradeTransformedTestDa
ta,\
                                                priceTransformedTestData,\
                                                previouslyPostedTransformedTe
stData));
    if(index == 0):
        trainingMergedData = hstack((trainingMergedData,\
                                     bowTitleModelSub,\
                                     bowEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    bowTitleTransformedCrossValidateData,\
                                    bowEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    bowTitleTransformedTestData,\
                                    bowEssayTransformedTestData));
    elif(index == 1):
        trainingMergedData = hstack((trainingMergedData,\
                                     tfIdfTitleModelSub,\
                                     tfIdfEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    tfIdfTitleTransformedCrossValidateData
,\
                                    tfIdfEssayTransformedCrossValidateData
```

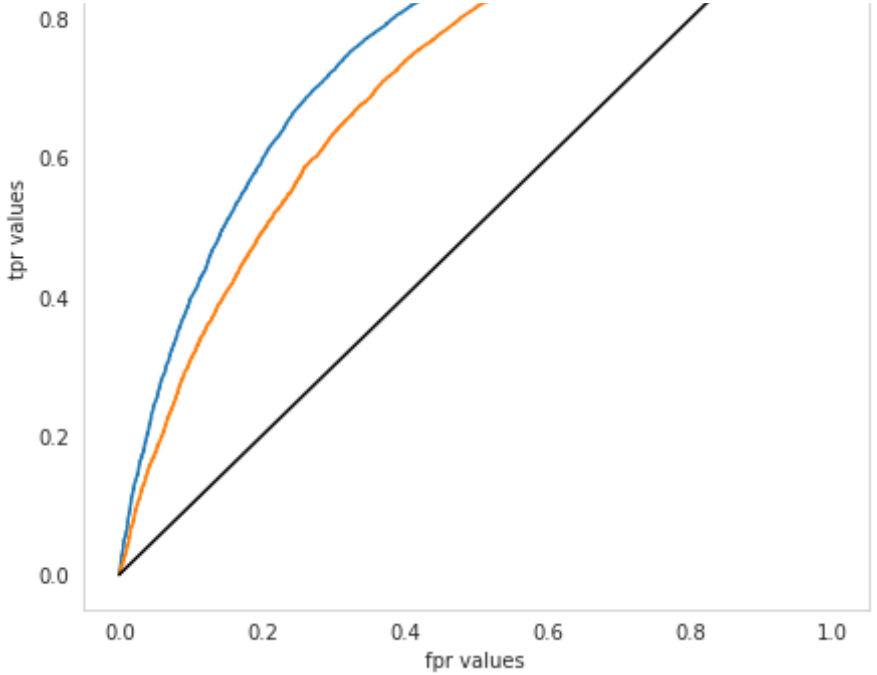
```
));
        testMergedData = hstack((testMergedData,\
                                 tfIdfTitleTransformedTestData,\
                                 tfIdfEssayTransformedTestData));
    elif(index == 2):
        trainingMergedData = hstack((trainingMergedData,\
                                     word2VecTitlesVectors,\
                                     word2VecEssaysVectors));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                 avgWord2VecTitleTransformedCrossValida
teData,\
                                 avgWord2VecEssayTransformedCrossValida
teData));
        testMergedData = hstack((testMergedData,\
                                 avgWord2VecTitleTransformedTestData,\
                                 avgWord2VecEssayTransformedTestData));
    elif(index == 3):
        trainingMergedData = hstack((trainingMergedData,\
                                     tfIdfWeightedWord2VecTitlesVectors
,\
                                     tfIdfWeightedWord2VecEssaysVectors
));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                 tfIdfWeightedWord2VecTitleTransformedC
rossValidateData,\
                                 tfIdfWeightedWord2VecEssayTransformedC
rossValidateData));
        testMergedData = hstack((testMergedData,\
                                 tfIdfWeightedWord2VecTitleTransformedT
estData,\
                                 tfIdfWeightedWord2VecEssayTransformedT
estData));

    lrClassifier = LogisticRegression(penalty = 'l1');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);
```

```
    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l1', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);
```

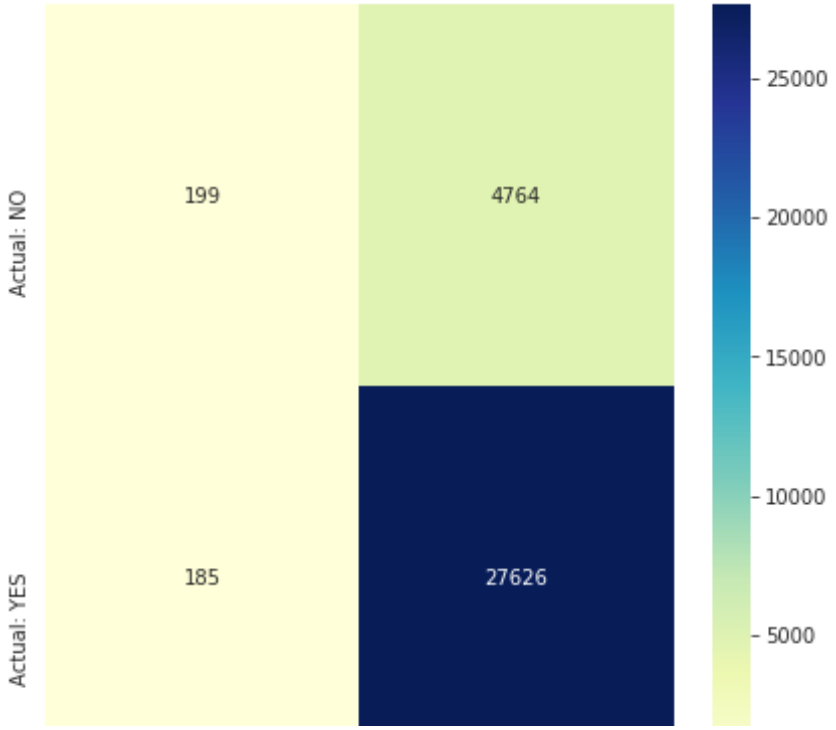
```
    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using {} vectorized text features merged
 with other features using logistic regression classifier: ".format(tec
hnique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
```

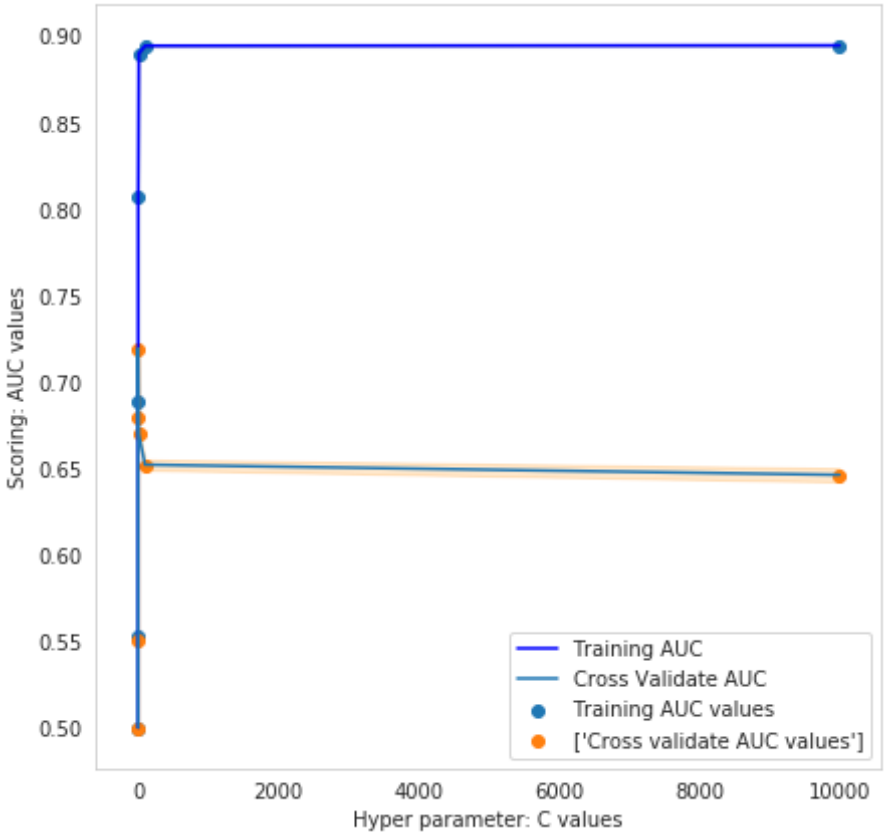
```
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l1)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```

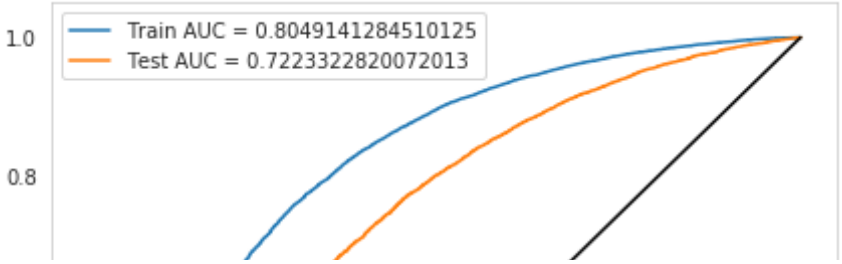


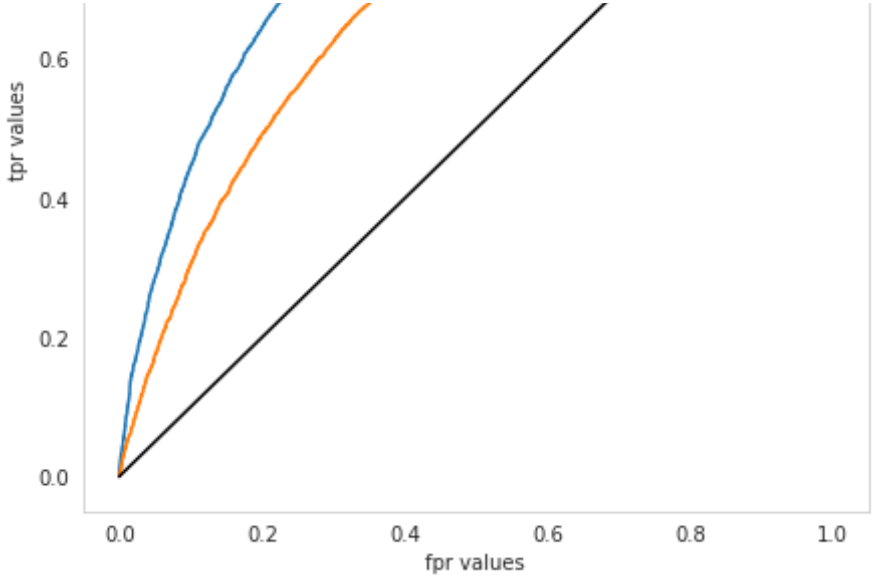




```
Results of analysis using Bag of words vectorized text features merged
with other features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.54467875 0.68069999 0.78104633 0.87543839 0.89427016 0.89531594
 0.89540994]
========================================
Optimal Hyper parameter Value:  0.1
========================================
AUC value of test data:  0.7247367204233771
========================================
Confusion Matrix :
============================================================
```

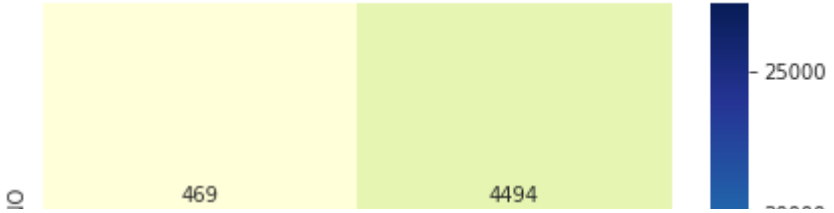


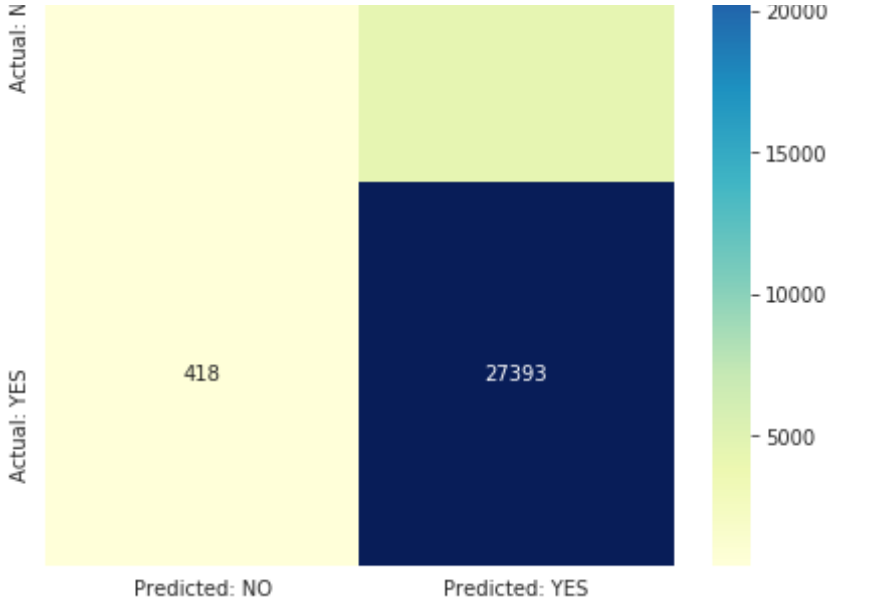



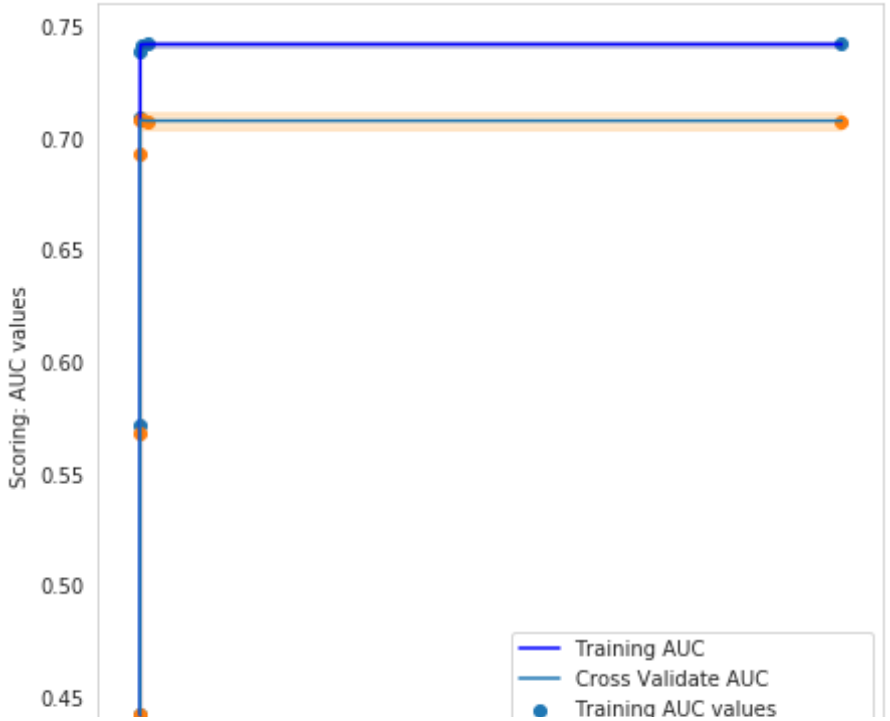



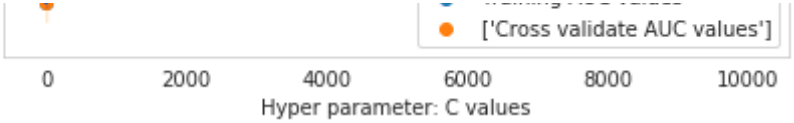


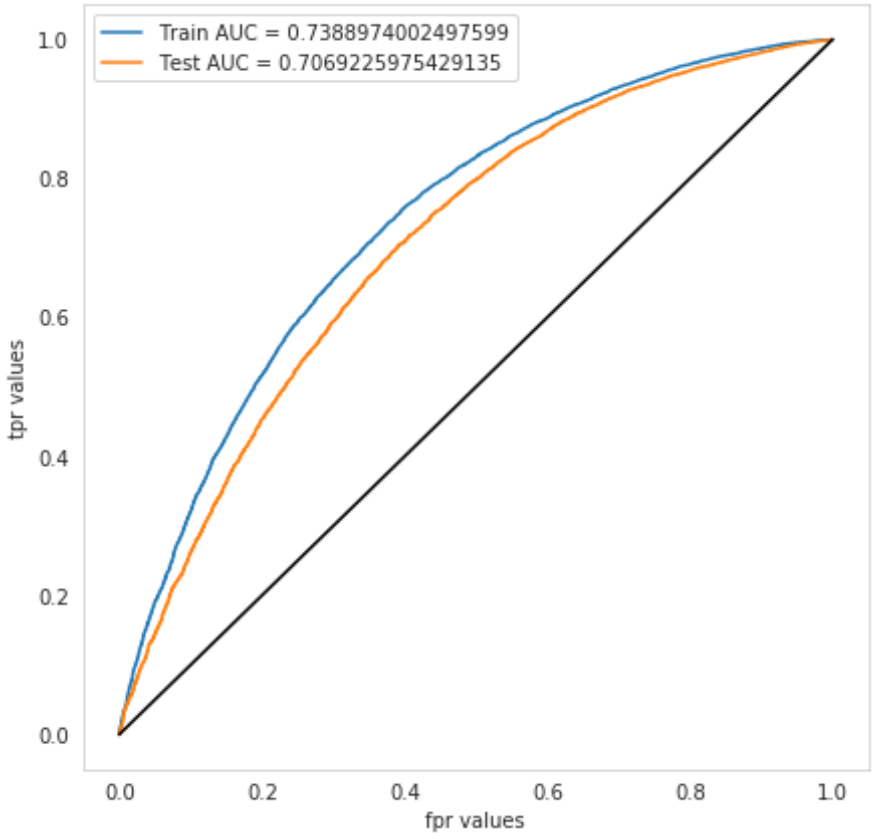

```
Results of analysis using Tf-Idf vectorized text features merged with o
ther features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.5        0.55263619 0.6892003  0.8075599  0.8900225  0.89452826
 0.89476864]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.7223322820072013
========================================
Confusion Matrix :
============================================================
```





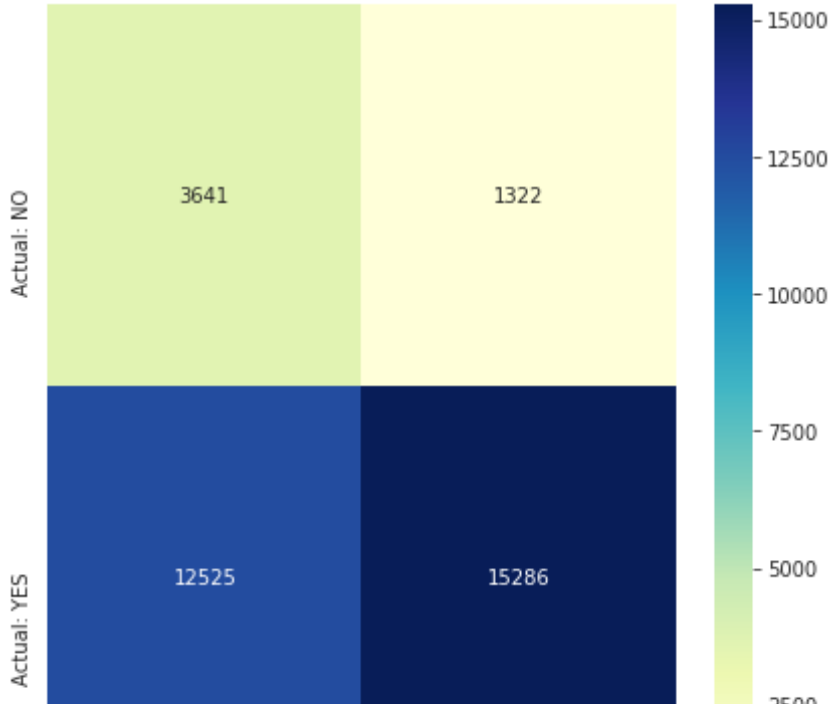



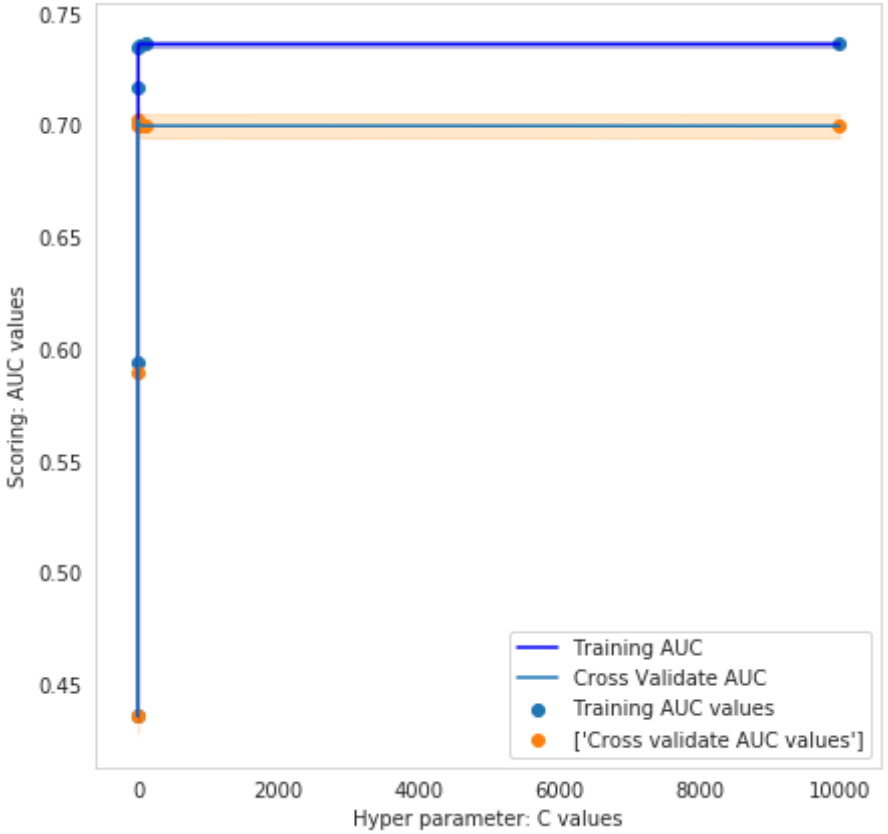


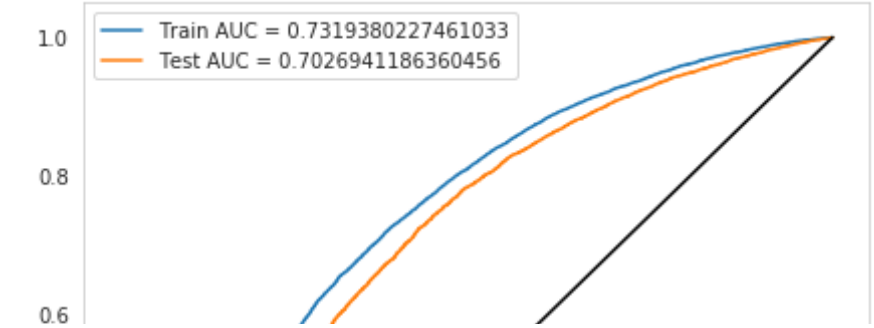



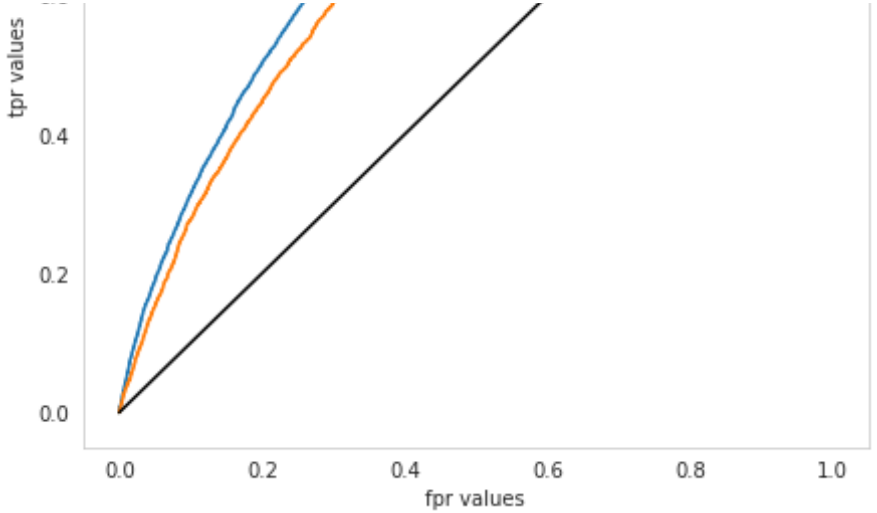

```
Results of analysis using Average Word2Vector vectorized text features
merged with other features using logistic regression classifier:
========================================================================
AUC values of train data:
=========================================
[0.44206766 0.57186875 0.70912202 0.73882933 0.74219065 0.74230992
 0.742313  ]
=========================================
Optimal Hyper parameter Value:  10
=========================================
```

```
----------------------------------------
AUC value of test data:  0.7069225975429135
========================================
Confusion Matrix :
============================================================
```

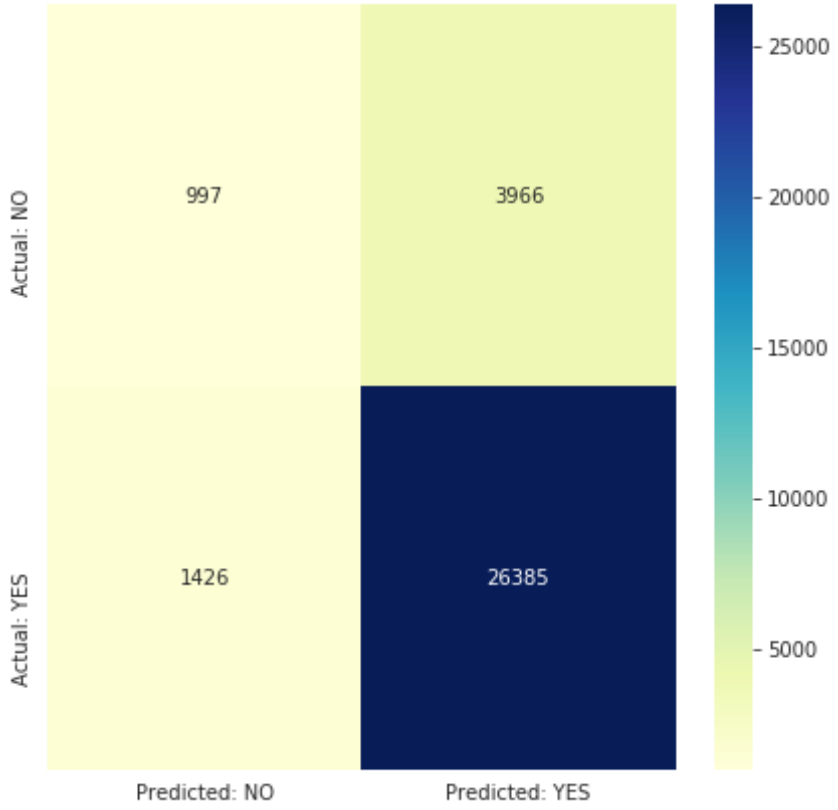





```
Results of analysis using Tf-Idf Weighted Word2Vector vectorized text f
eatures merged with other features using logistic regression classifie
r:
========================================================================
AUC values of train data:
========================================
[0.43585946 0.59385553 0.71699242 0.7351215  0.73640921 0.73642674
 0.7364281 ]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.7026941186360456
========================================
Confusion Matrix :
============================================================
```




## Classification using imbalanced data containing all features by logistic regression(l2 regularizer)

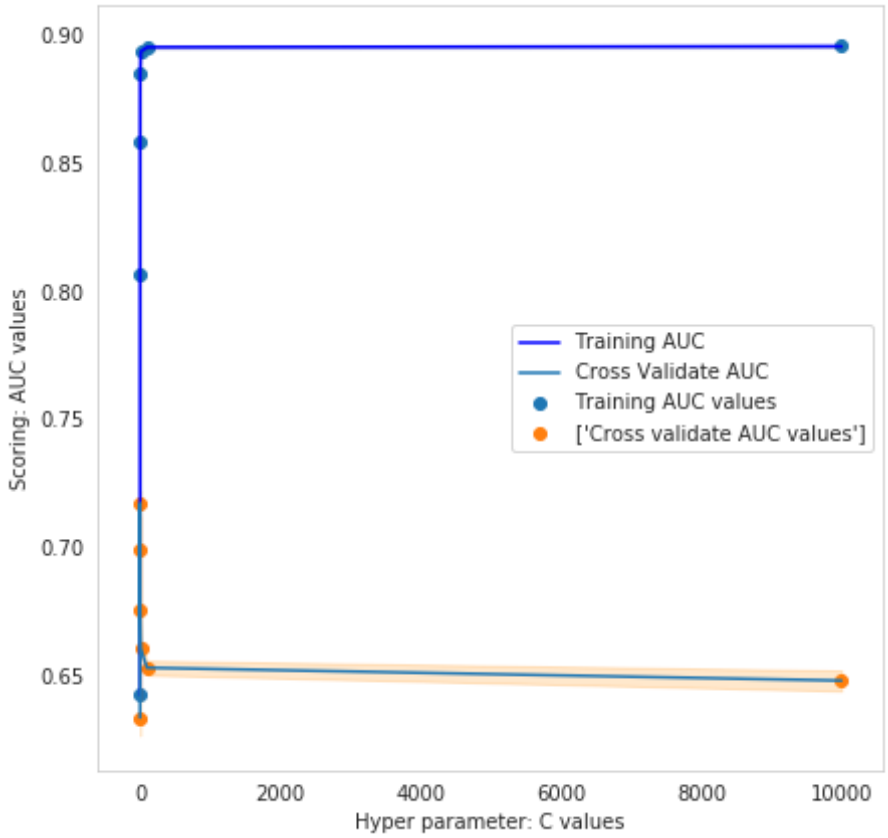
In [0]:
```
techniques = ['Bag of words', 'Tf-Idf', 'Average Word2Vector', 'Tf-Idf
 Weighted Word2Vector'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                        subCategoriesVectorsSub,\
                                        teacherPrefixVectorsSub,\
                                        schoolStateVectorsSub,\
                                        projectGradeVectorsSub,\
                                        priceStandardizedSub,\
                                        previouslyPostedStandardizedSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                                subCategoriesTransformedCross
ValidateData,\
                                                teacherPrefixTransformedCross
ValidateData,\
                                                schoolStateTransformedCrossVa
lidateData,\
                                                projectGradeTransformedCrossV
alidateData,\
                                                priceTransformedCrossValidate
Data,\
                                                previouslyPostedTransformedCr
ossValidateData));
    testMergedData = hstack((categoriesTransformedTestData,\
                                                subCategoriesTransformedTestD
ata,\
                                                teacherPrefixTransformedTestD
ata,\
                                                schoolStateTransformedTestDat
a,\
                                                projectGradeTransformedTestDa
ta,\
                                                priceTransformedTestData,\
                                                previouslyPostedTransformedTe
stData));
    if(index == 0):
        trainingMergedData = hstack((trainingMergedData,\
                                        bowTitleModelSub,\
```

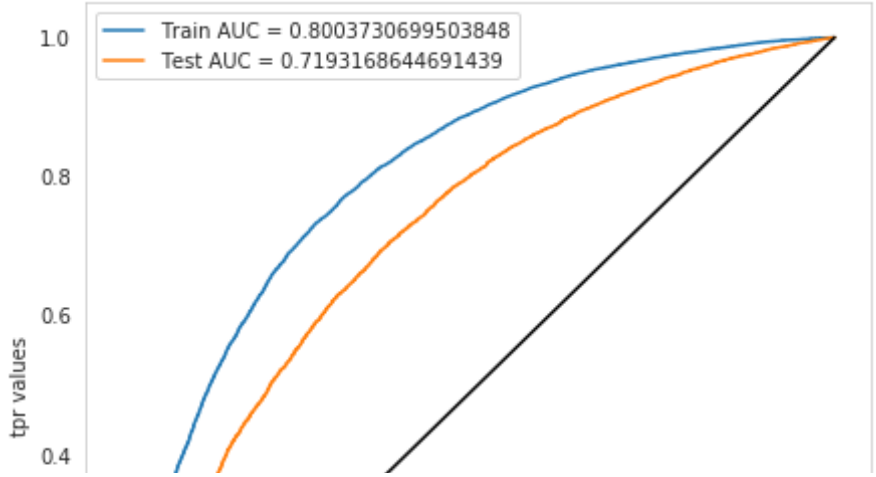
```
                                        bowEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    bowTitleTransformedCrossValidateData,\
                                    bowEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    bowTitleTransformedTestData,\
                                    bowEssayTransformedTestData));
    elif(index == 1):
        trainingMergedData = hstack((trainingMergedData,\
                                        tfIdfTitleModelSub,\
                                        tfIdfEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    tfIdfTitleTransformedCrossValidateData
,\
                                    tfIdfEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    tfIdfTitleTransformedTestData,\
                                    tfIdfEssayTransformedTestData));
    elif(index == 2):
        trainingMergedData = hstack((trainingMergedData,\
                                        word2VecTitlesVectors,\
                                        word2VecEssaysVectors));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    avgWord2VecTitleTransformedCrossValida
teData,\
                                    avgWord2VecEssayTransformedCrossValida
teData));
        testMergedData = hstack((testMergedData,\
                                    avgWord2VecTitleTransformedTestData,\
                                    avgWord2VecEssayTransformedTestData));
    elif(index == 3):
        trainingMergedData = hstack((trainingMergedData,\
                                        tfIdfWeightedWord2VecTitlesVectors
,\
                                        tfIdfWeightedWord2VecEssaysVectors
));
        crossValidateMergedData = hstack((crossValidateMergedData,\
```

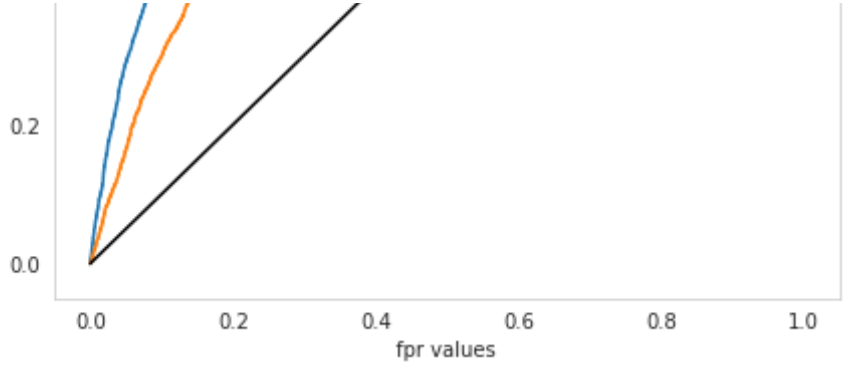
```
                                        tfIdfWeightedWord2VecTitleTransformedC
rossValidateData,\
                                        tfIdfWeightedWord2VecEssayTransformedC
rossValidateData));
        testMergedData = hstack((testMergedData,\
                                        tfIdfWeightedWord2VecTitleTransformedT
estData,\
                                        tfIdfWeightedWord2VecEssayTransformedT
estData));

    lrClassifier = LogisticRegression(penalty = 'l2');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
```

```
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l2', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using {} vectorized text features merged
 with other features using logistic regression classifier: ".format(tec
hnique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
```

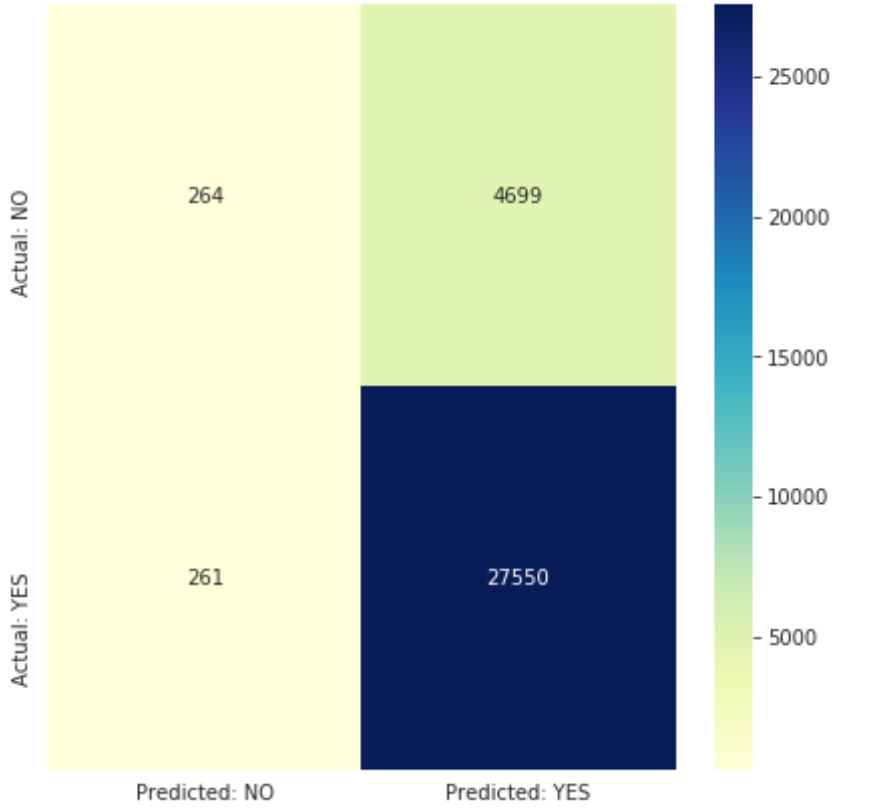
```
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l2)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```

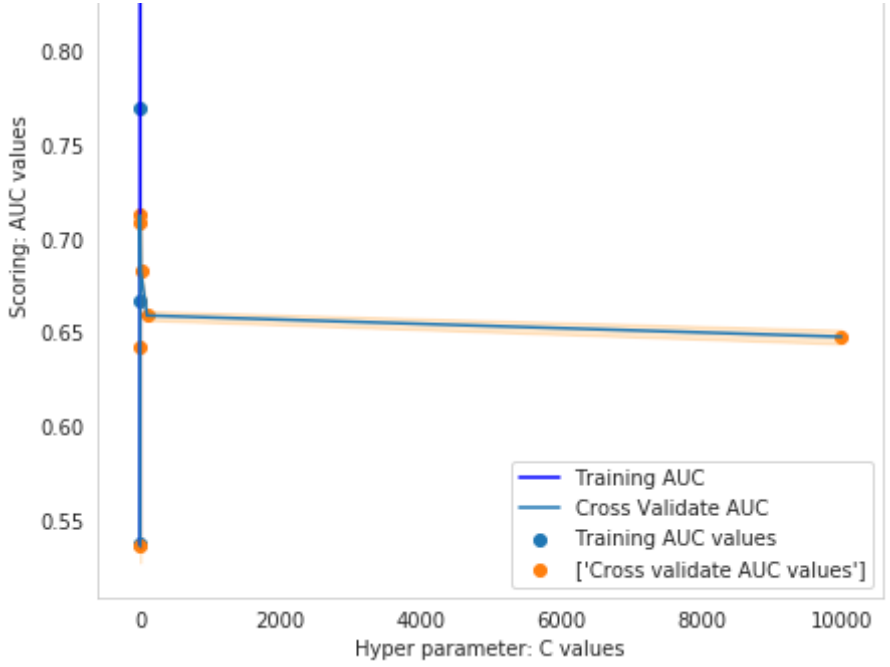


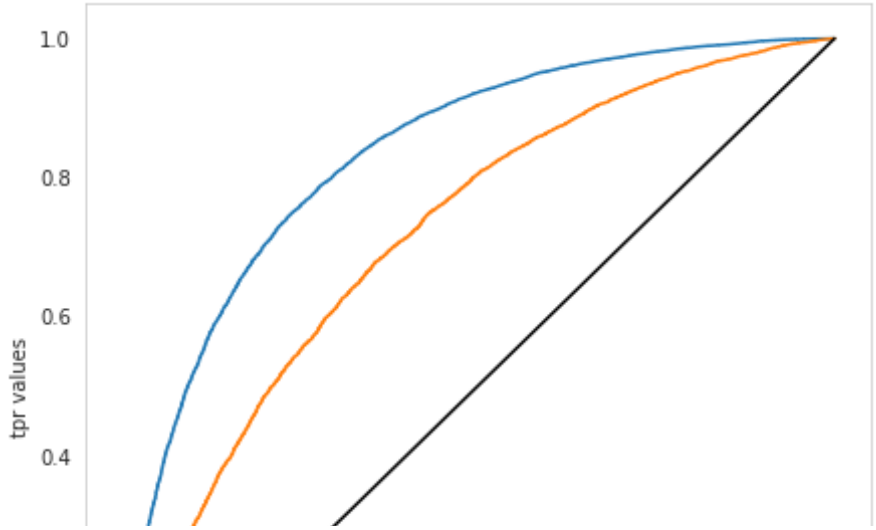



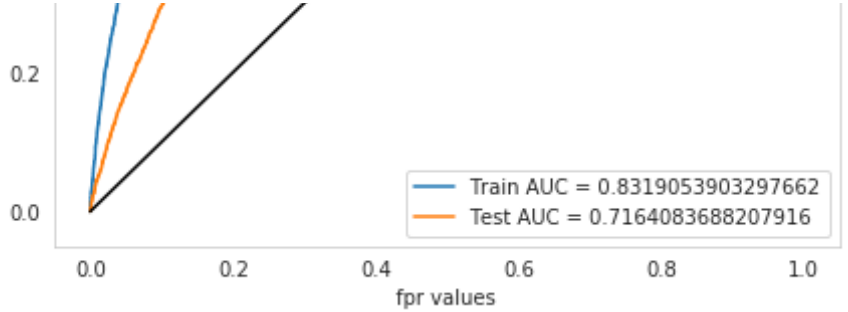



```
Results of analysis using Bag of words vectorized text features merged
with other features using logistic regression classifier:
==========================================================================
AUC values of train data:
========================================
[0.64223041 0.80632339 0.85780703 0.88452522 0.89333059 0.89505771
 0.89538859]
========================================
Optimal Hyper parameter Value:  0.01
========================================
AUC value of test data:  0.7193168644691439
========================================
Confusion Matrix :
============================================================
```

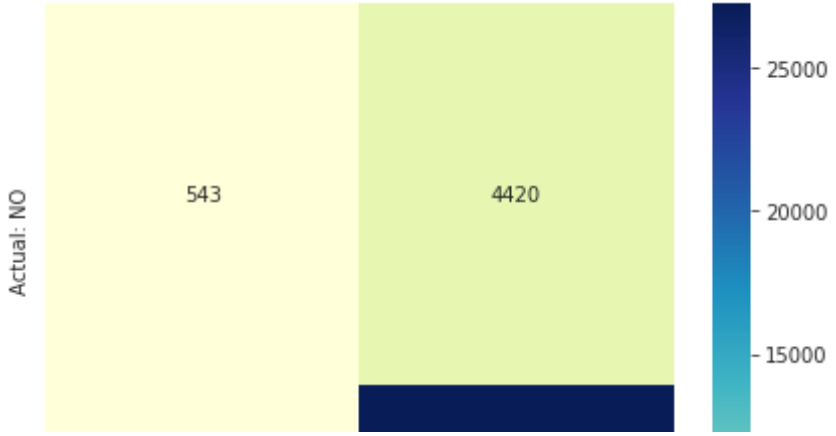


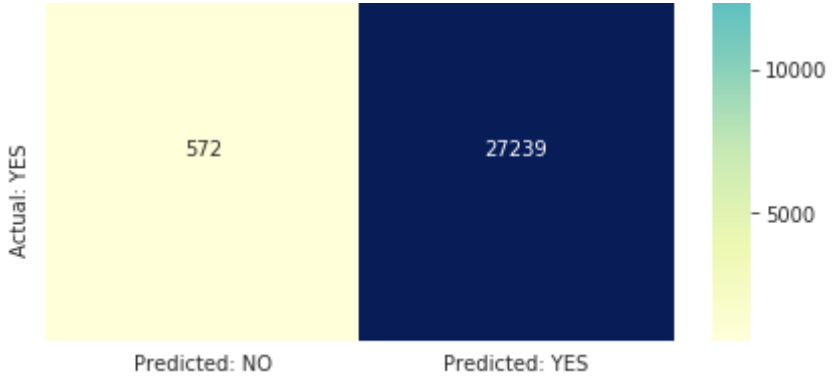



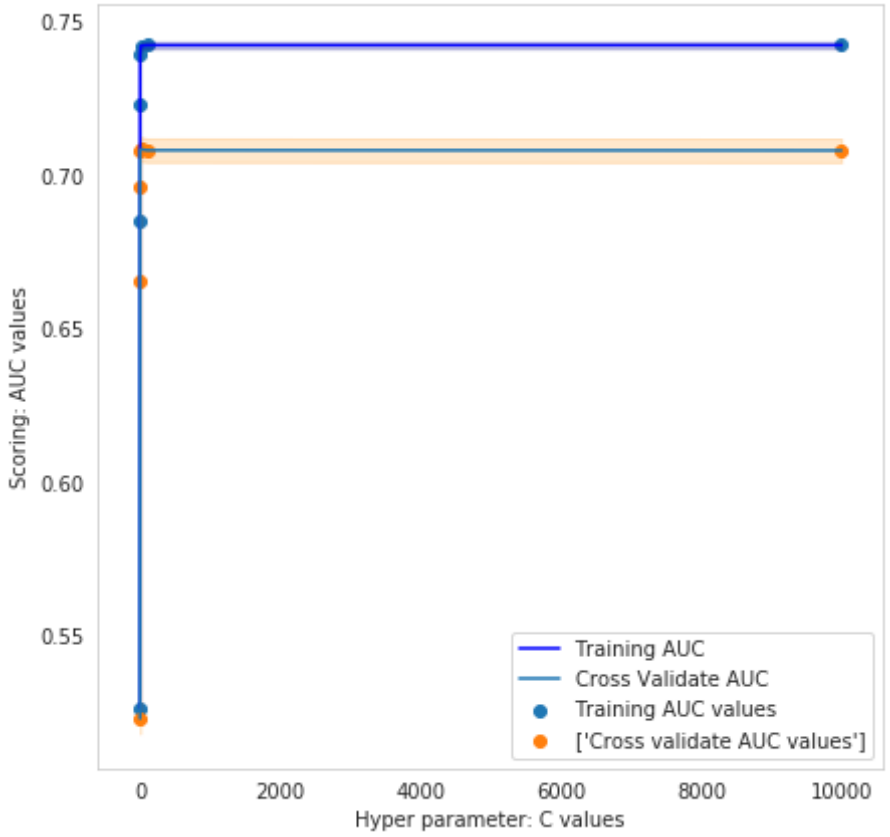



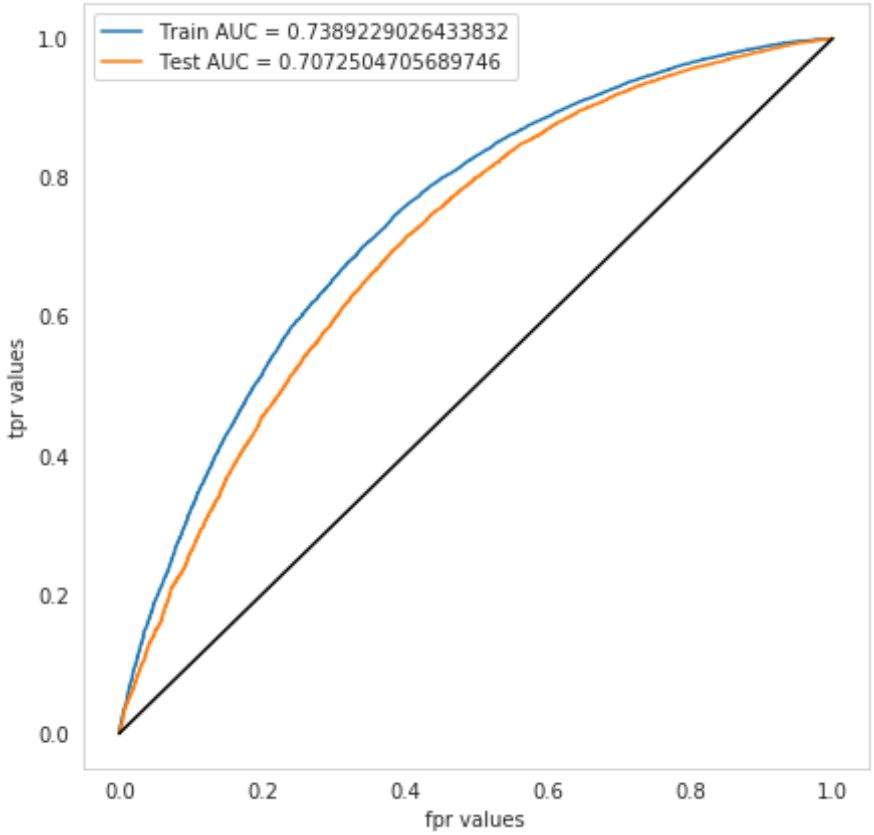



```
Results of analysis using Tf-Idf vectorized text features merged with o
ther features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.53695771 0.66677739 0.76971945 0.84212811 0.88313124 0.89338676
 0.89475287]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.7164083688207916
========================================
Confusion Matrix :
============================================================
```



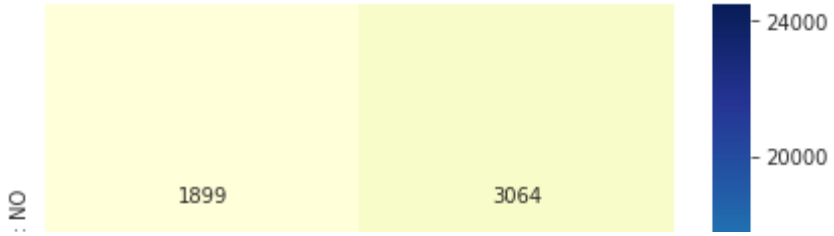


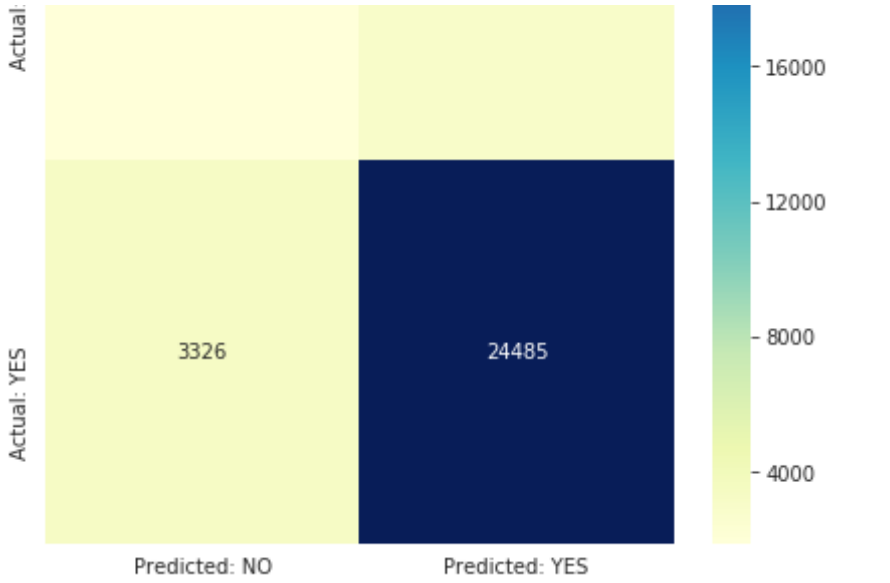



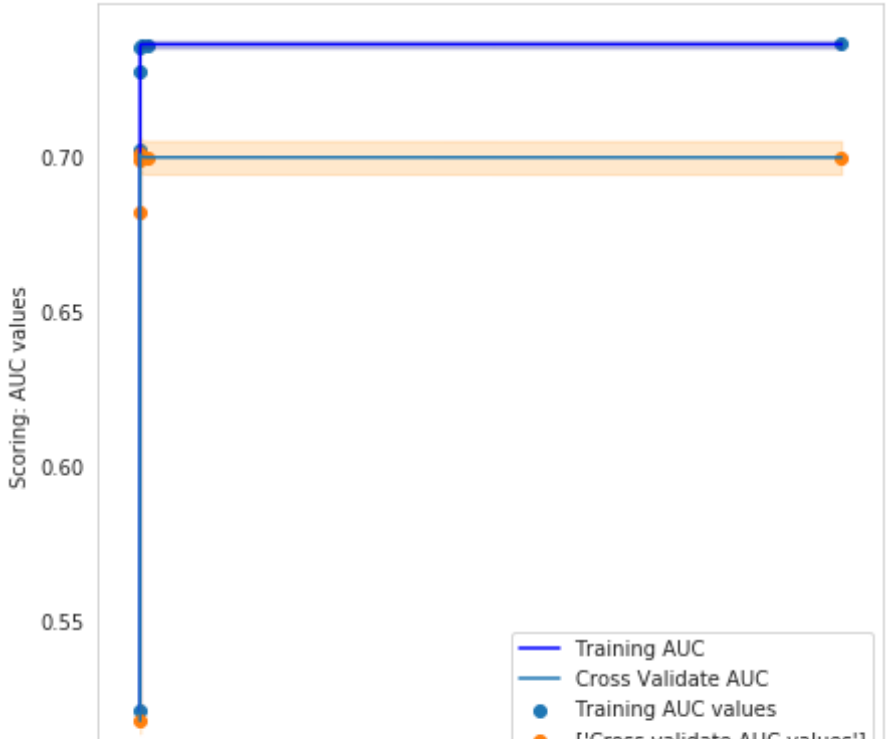


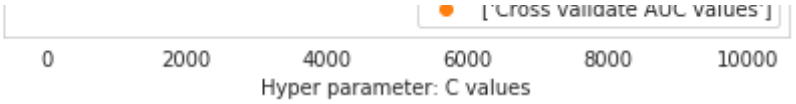

```
Results of analysis using Average Word2Vector vectorized text features
merged with other features using logistic regression classifier:
========================================================================
AUC values of train data:
=========================================
[0.52631381 0.68545447 0.72282828 0.73917894 0.74220102 0.74240965
```

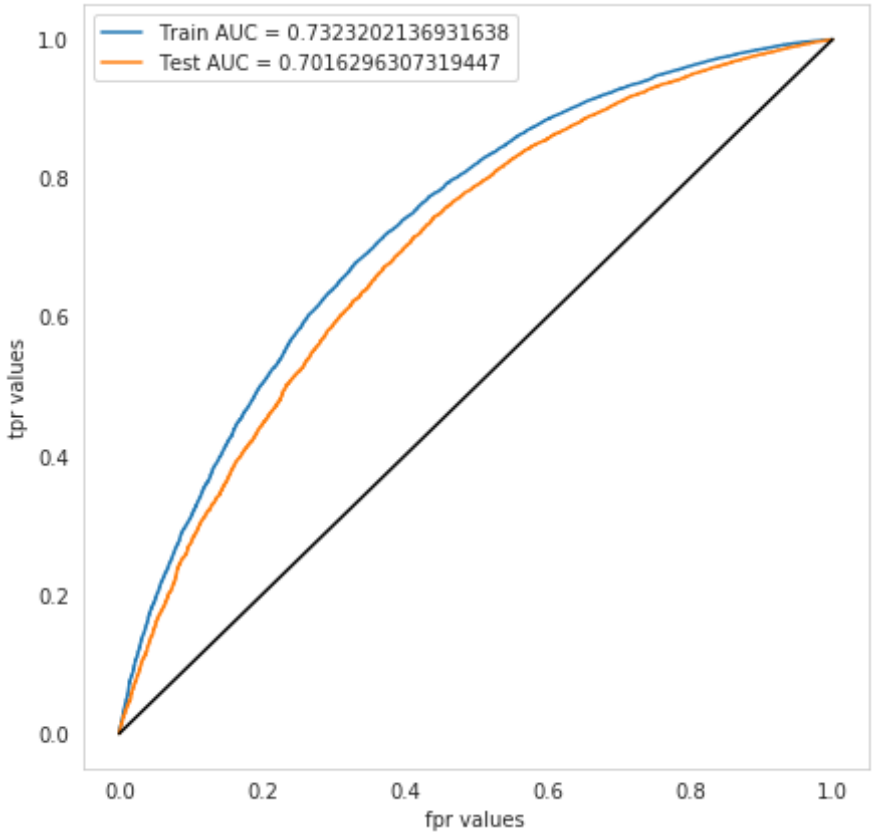
```
 0.74241624]
========================================
Optimal Hyper parameter Value:  10
========================================
AUC value of test data:  0.7072504705689746
========================================
Confusion Matrix :
============================================================
```





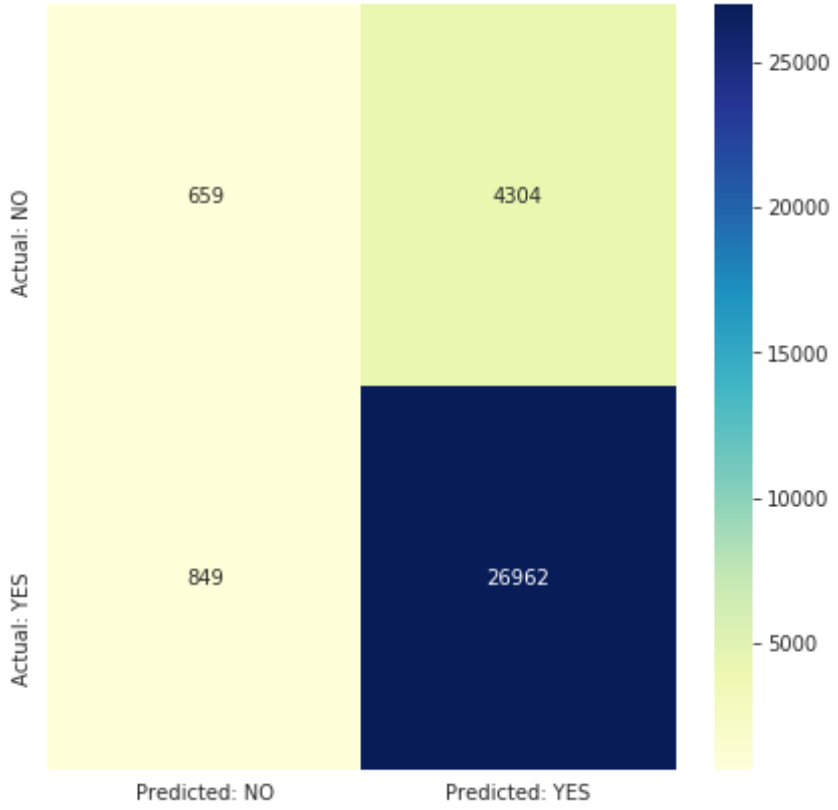






```
Results of analysis using Tf-Idf Weighted Word2Vector vectorized text f
eatures merged with other features using logistic regression classifie
r:
========================================================================
AUC values of train data:
========================================
[0.52093619 0.70249046 0.72789466 0.73560832 0.73641489 0.73645139
 0.73645735]
========================================
Optimal Hyper parameter Value:  1
```

```
========================================
AUC value of test data:  0.7016296307319447
========================================
Confusion Matrix :
============================================================
```




## Classification using balanced data containing all features by logistic regression(l1 regularizer)

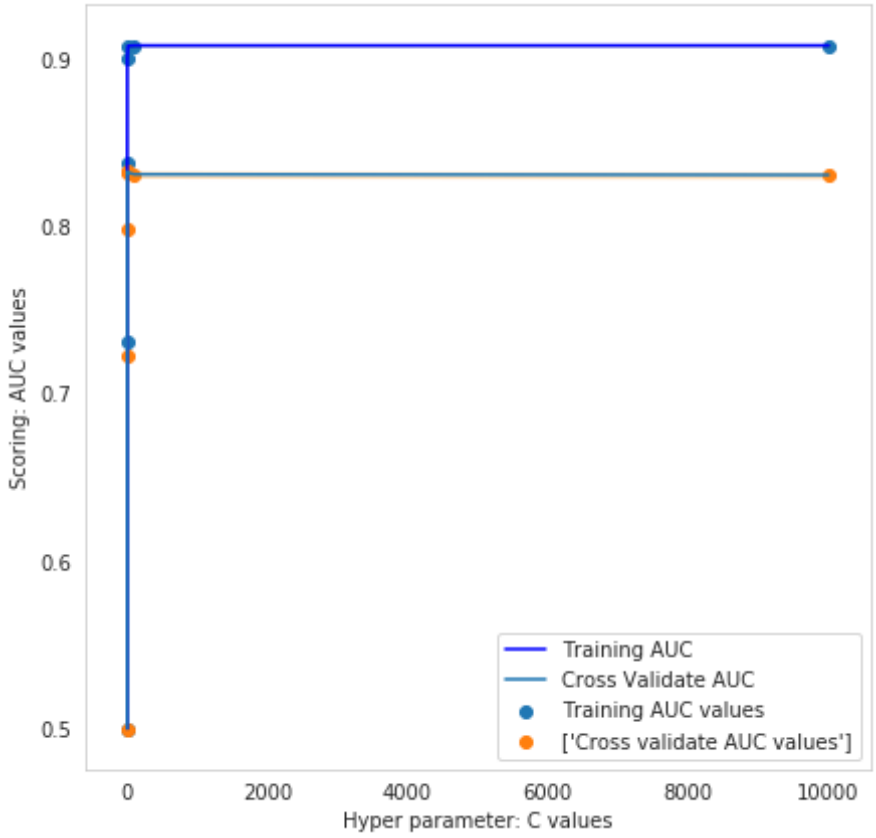
In [0]:
```
techniques = ['Bag of words', 'Tf-Idf'];
for index, technique in enumerate(techniques):
```

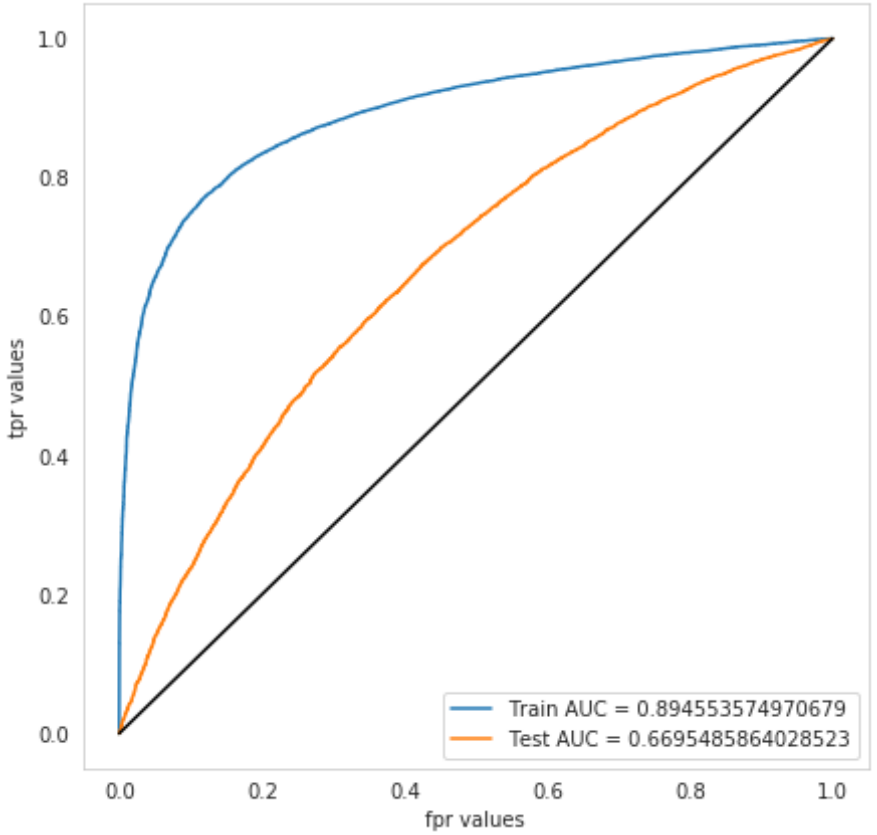
```
    trainingMergedData = hstack((categoriesVectorsSub,\
                                        subCategoriesVectorsSub,\
                                        teacherPrefixVectorsSub,\
                                        schoolStateVectorsSub,\
                                        projectGradeVectorsSub,\
                                        priceStandardizedSub,\
                                        previouslyPostedStandardizedSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                            subCategoriesTransformedCross
ValidateData,\
                                            teacherPrefixTransformedCross
ValidateData,\
                                            schoolStateTransformedCrossVa
lidateData,\
                                            projectGradeTransformedCrossV
alidateData,\
                                            priceTransformedCrossValidate
Data,\
                                            previouslyPostedTransformedCr
ossValidateData));
    testMergedData = hstack((categoriesTransformedTestData,\
                                            subCategoriesTransformedTestD
ata,\
                                            teacherPrefixTransformedTestD
ata,\
                                            schoolStateTransformedTestDat
a,\
                                            projectGradeTransformedTestDa
ta,\
                                            priceTransformedTestData,\
                                            previouslyPostedTransformedTe
stData));
    if(index == 0):
        trainingMergedData = hstack((trainingMergedData,\
                                        bowTitleModelSub,\
                                        bowEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    bowTitleTransformedCrossValidateData,\
```

```
                                    bowEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    bowTitleTransformedTestData,\
                                    bowEssayTransformedTestData));
    elif(index == 1):
        trainingMergedData = hstack((trainingMergedData,\
                                        tfIdfTitleModelSub,\
                                        tfIdfEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    tfIdfTitleTransformedCrossValidateData
,\
                                    tfIdfEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    tfIdfTitleTransformedTestData,\
                                    tfIdfEssayTransformedTestData));

    lrClassifier = LogisticRegression(penalty = 'l1');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
```

```
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l1', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using {} vectorized text features merged
 with other features using logistic regression classifier: ".format(tec
```

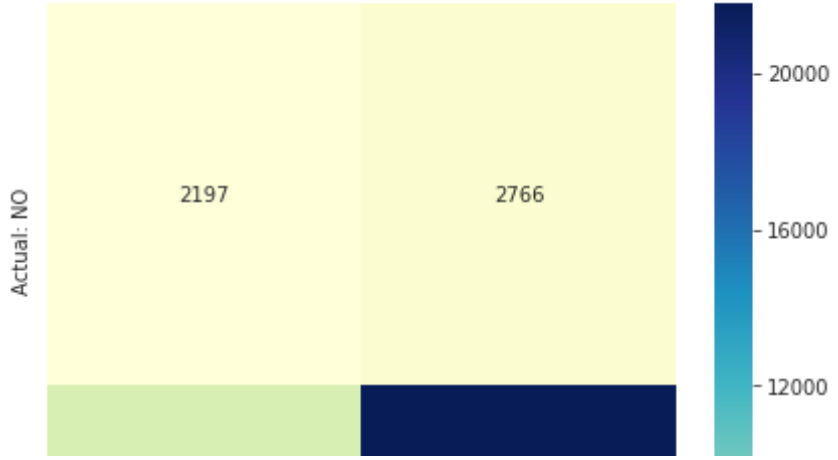
```
hnique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l2)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```

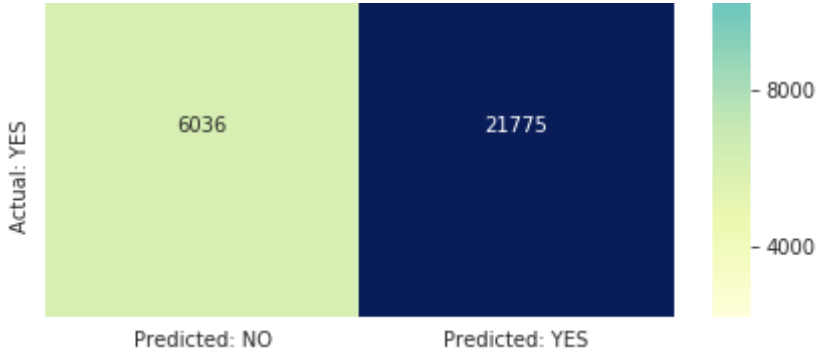


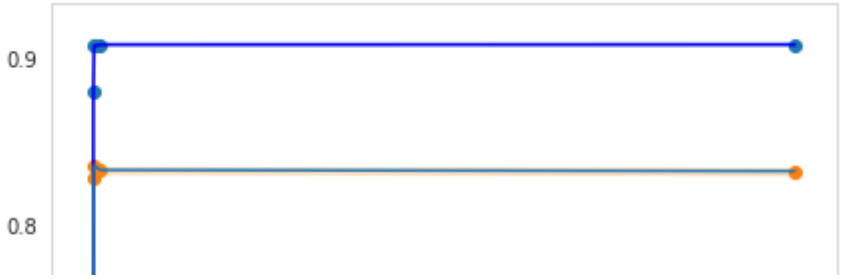


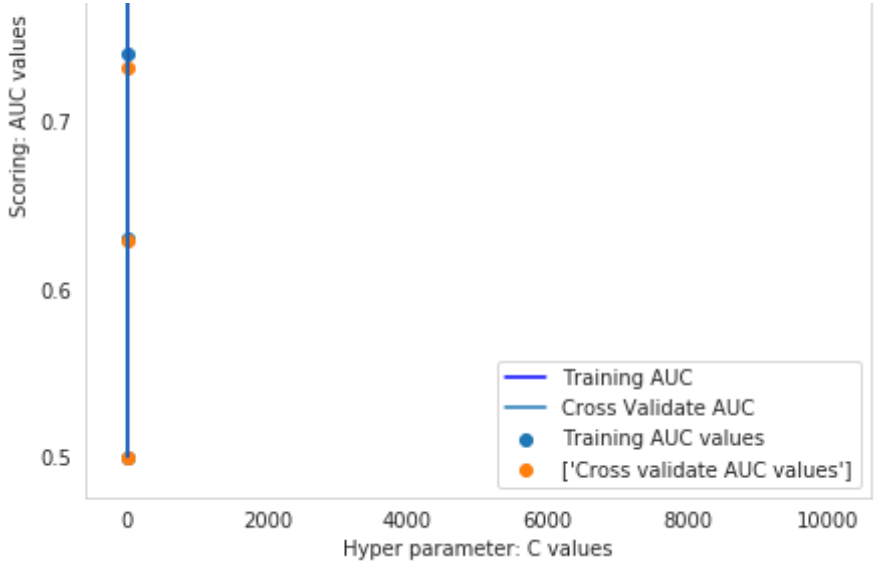

```
Results of analysis using Bag of words vectorized text features merged
with other features using logistic regression classifier:
__________________________________________________________________
```

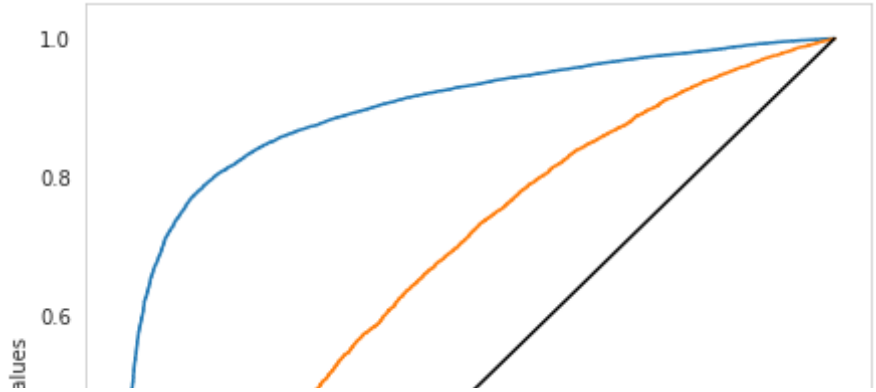
```
==========================================================================
AUC values of train data:
========================================
[0.5        0.73075458 0.83781763 0.90002753 0.90777405 0.90805276
 0.90807229]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.6695485864028523
========================================
Confusion Matrix :
============================================================
```

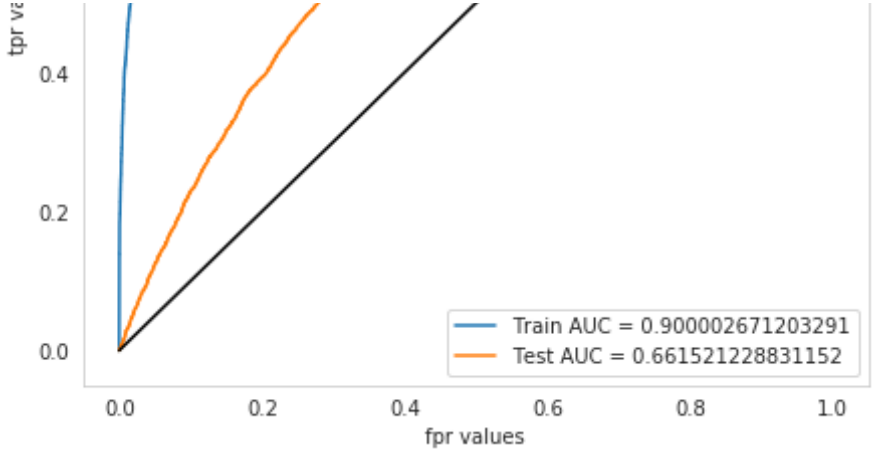




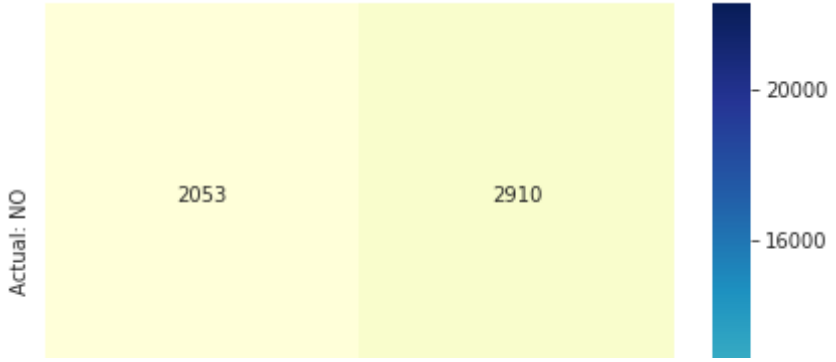



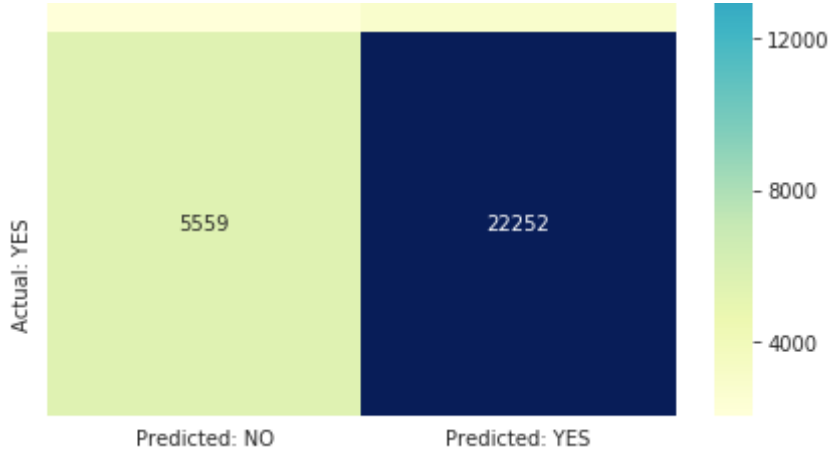





```
Results of analysis using Tf-Idf vectorized text features merged with o
ther features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.5        0.63000047 0.7408425  0.88013257 0.90744882 0.90829361
 0.90832484]
========================================
Optimal Hyper parameter Value:  10
========================================
AUC value of test data:  0.661521228831152
========================================
Confusion Matrix :
============================================================
```






## Classification using balanced data containing all features by logistic regression(l2 regularizer)

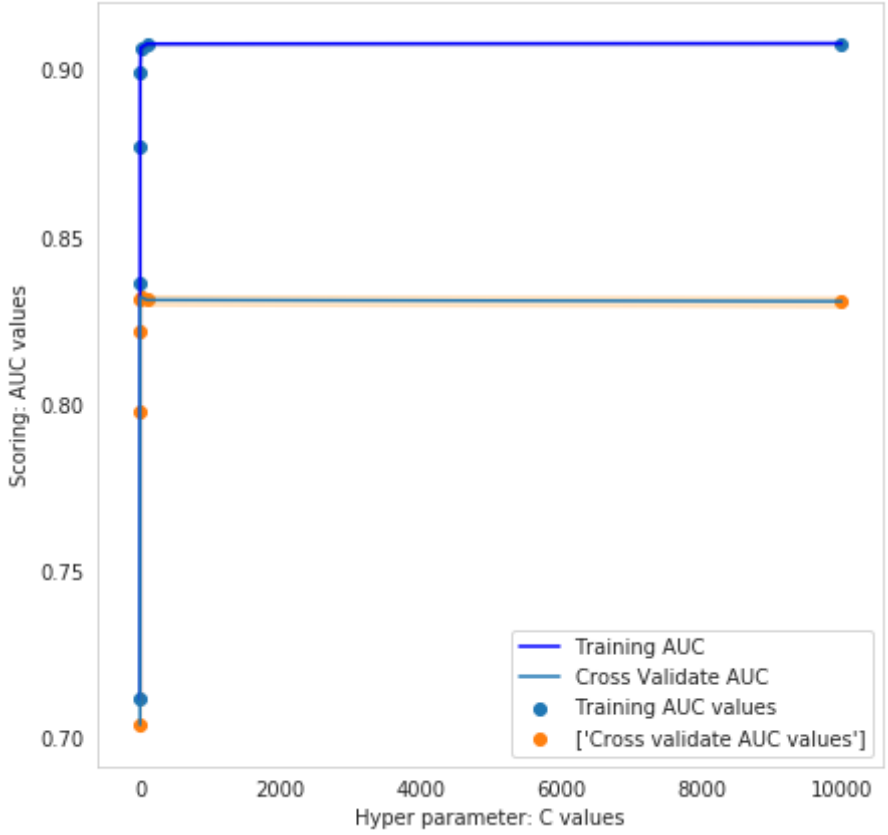
In [0]:
```
techniques = ['Bag of words', 'Tf-Idf'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
                                 teacherPrefixVectorsSub,\
                                 schoolStateVectorsSub,\
                                 projectGradeVectorsSub,\
                                 priceStandardizedSub,\
                                 previouslyPostedStandardizedSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                      subCategoriesTransformedCross
ValidateData,\
                                      teacherPrefixTransformedCross
ValidateData,\
                                      schoolStateTransformedCrossVa
lidateData,\
                                      projectGradeTransformedCrossV
alidateData,\
```

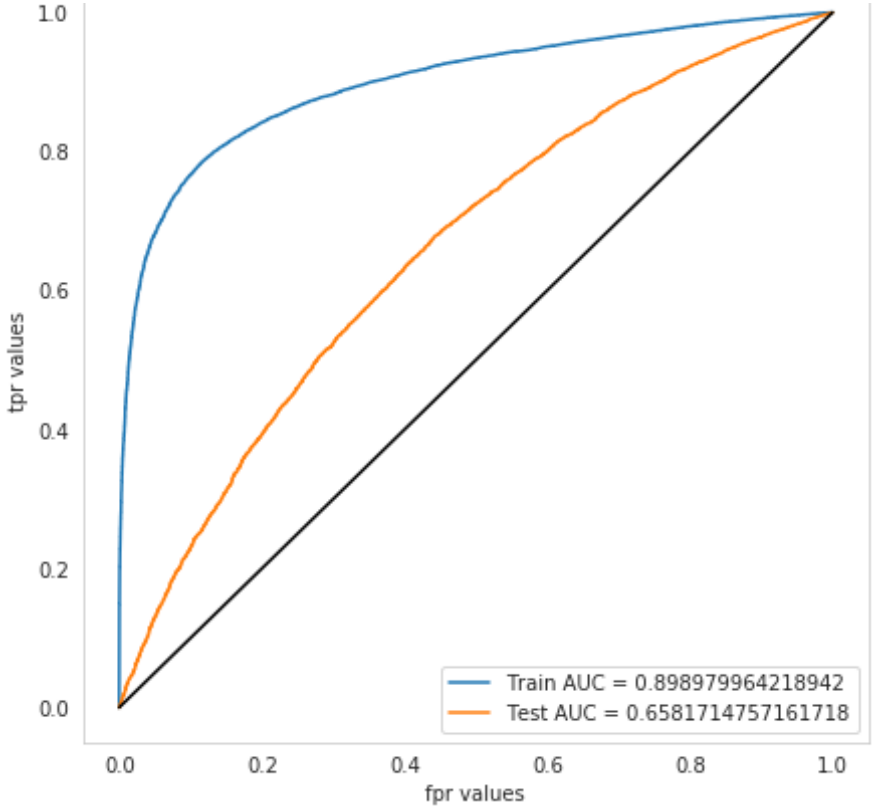
```
                                                priceTransformedCrossValidate
Data,\
                                                previouslyPostedTransformedCr
ossValidateData));
    testMergedData = hstack((categoriesTransformedTestData,\
                                                subCategoriesTransformedTestD
ata,\
                                                teacherPrefixTransformedTestD
ata,\
                                                schoolStateTransformedTestDat
a,\
                                                projectGradeTransformedTestDa
ta,\
                                                priceTransformedTestData,\
                                                previouslyPostedTransformedTe
stData));
    if(index == 0):
        trainingMergedData = hstack((trainingMergedData,\
                                     bowTitleModelSub,\
                                     bowEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    bowTitleTransformedCrossValidateData,\
                                    bowEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    bowTitleTransformedTestData,\
                                    bowEssayTransformedTestData));
    elif(index == 1):
        trainingMergedData = hstack((trainingMergedData,\
                                     tfIdfTitleModelSub,\
                                     tfIdfEssayModelSub));
        crossValidateMergedData = hstack((crossValidateMergedData,\
                                    tfIdfTitleTransformedCrossValidateData
,\
                                    tfIdfEssayTransformedCrossValidateData
));
        testMergedData = hstack((testMergedData,\
                                    tfIdfTitleTransformedTestData,\
                                    tfIdfEssayTransformedTestData));
```

```
    lrClassifier = LogisticRegression(penalty = 'l2');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l2', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
```

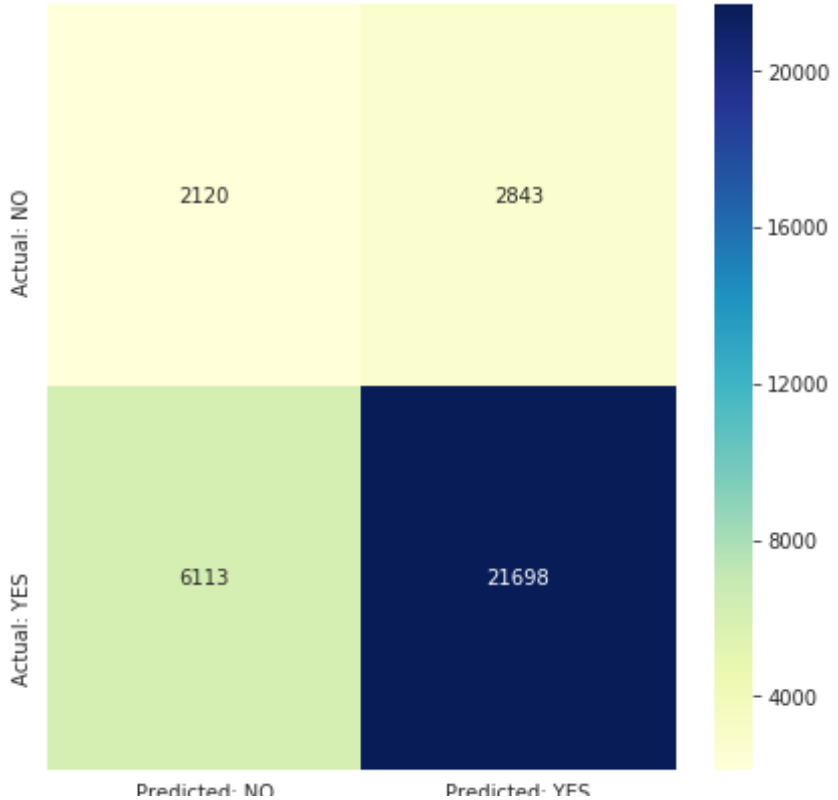
```
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using {} vectorized text features merged
 with other features using logistic regression classifier: ".format(tec
hnique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
```

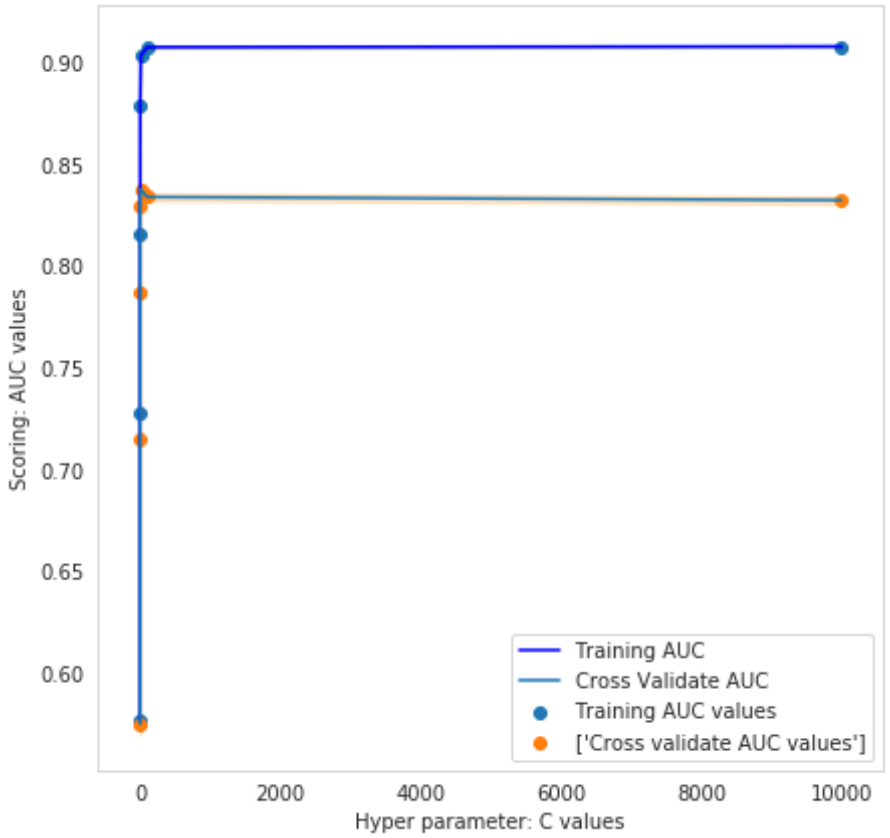
```
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l2)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```

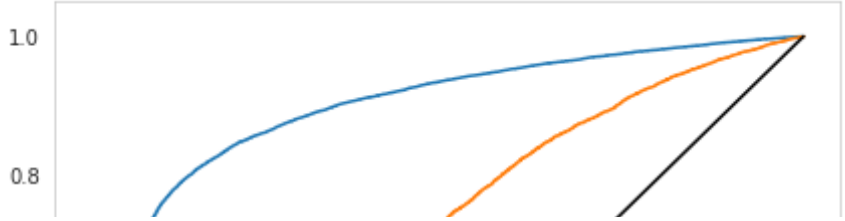


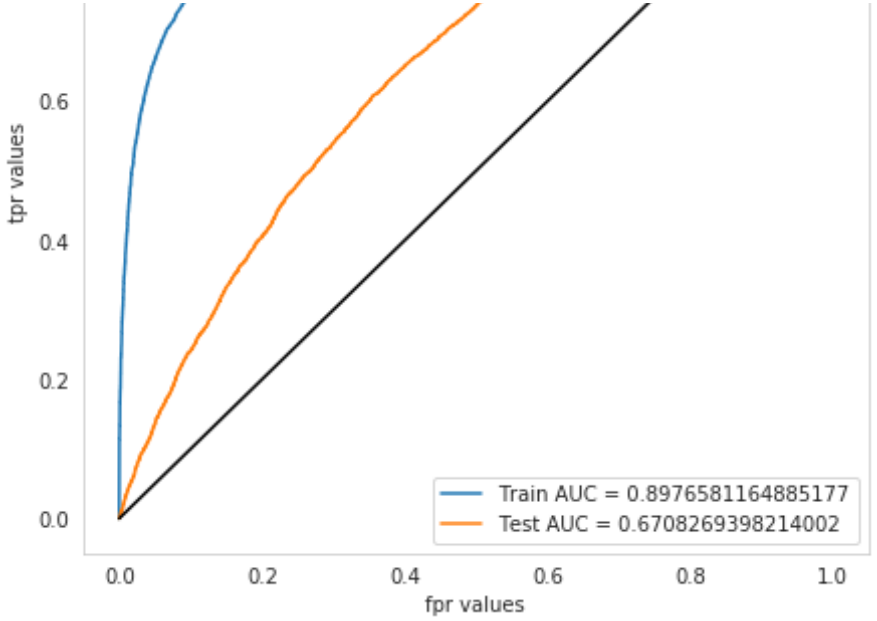



```
Results of analysis using Bag of words vectorized text features merged
with other features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.71179694 0.83616414 0.87716454 0.89924396 0.90675979 0.90792005
 0.90804964]
========================================
Optimal Hyper parameter Value:  10
========================================
AUC value of test data:  0.6581714757161718
========================================
Confusion Matrix :
============================================================
```

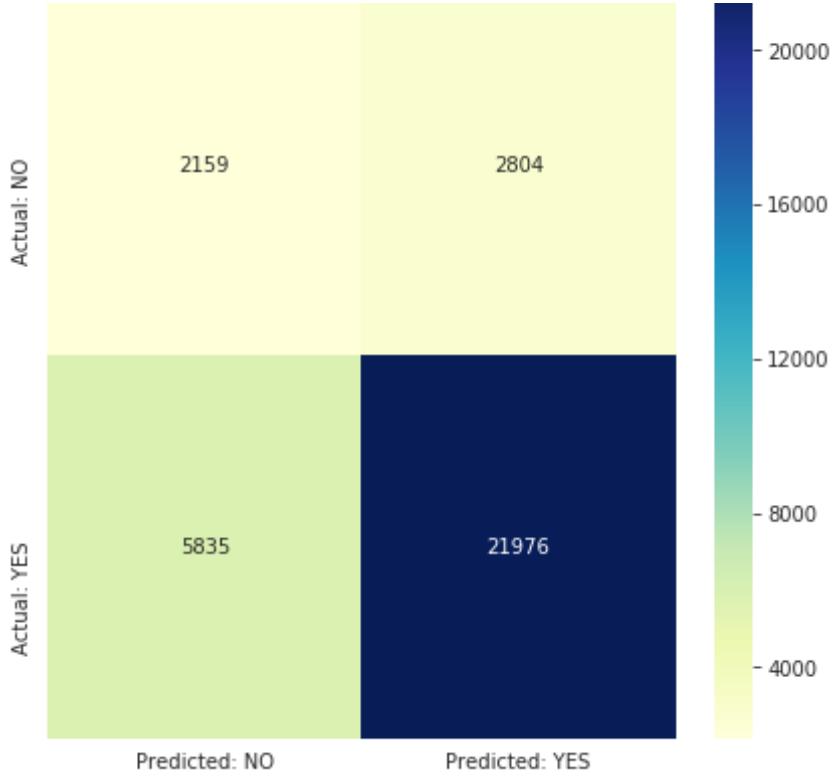







```
Results of analysis using Tf-Idf vectorized text features merged with o
ther features using logistic regression classifier:
========================================================================
AUC values of train data:
========================================
[0.57648166 0.72784659 0.81585789 0.87964134 0.90396523 0.90794688
 0.90832096]
========================================
Optimal Hyper parameter Value:  10
========================================
AUC value of test data:  0.6708269398214002
========================================
Confusion Matrix :
============================================================
```




## Classification using data with no text features by logistic regression

In [145]:
```
projectsData['number_of_words_in_title'] = [len(projectTitle.split(" ")) for projectTitle in projectsData['preprocessed_titles']];
projectsData.shape
```

Out[145]: (109248, 25)

In [146]:
```
projectsData['number_of_words_in_essay'] = [len(projectEssay.split(" ")) for projectEssay in projectsData['preprocessed_essays']];
projectsData.shape
```

Out[146]: (109248, 26)

## Calculating sentiment score of each essay

In [147]:
```
sentimentAnalyzer = SentimentIntensityAnalyzer();
positiveSentimentScores = [];
negativeSentimentScores = [];
neutralSentimentScores = [];
compoundSentimentScores = [];
for projectEssay in tqdm(projectsData['preprocessed_essays'].values):
  sentimentScore = sentimentAnalyzer.polarity_scores(projectEssay);
  positiveSentimentScores.append(sentimentScore['pos']);
  negativeSentimentScores.append(sentimentScore['neg']);
  neutralSentimentScores.append(sentimentScore['neu']);
  compoundSentimentScores.append(sentimentScore['compound']);
print(len(positiveSentimentScores), len(negativeSentimentScores), len(n
eutralSentimentScores), len(compoundSentimentScores));
print(positiveSentimentScores[0:5])
```

```
109248 109248 109248 109248
[0.154, 0.305, 0.23, 0.256, 0.151]
```

In [148]:
```
projectsData['positive_sentiment_score'] = positiveSentimentScores;
projectsData['negative_sentiment_score'] = negativeSentimentScores;
projectsData['neutral_sentiment_score'] = neutralSentimentScores;
projectsData['compound_sentiment_score'] = compoundSentimentScores;
projectsData.shape
```

Out[148]: (109248, 30)

## Splitting Data(Only training and test)

In [150]:
```
projectsData = projectsData.dropna(subset = ['teacher_prefix']);
projectsData.shape
```

Out[150]: (109245, 30)

In [151]:
```
classesData = projectsData['project_is_approved']
print(classesData.shape)
```

```
(109245,)
```

In [0]:
```
trainingData, testData, classesTraining, classesTest = model_selection.
train_test_split(projectsData, classesData, test_size =  0.3, random_st
ate = 0, stratify = classesData);
trainingData, crossValidateData, classesTraining, classesCrossValidate
= model_selection.train_test_split(trainingData, classesTraining, test_
size = 0.3, random_state = 0, stratify = classesTraining);
```

In [153]:
```
print("Shapes of splitted data: ");
equalsBorder(70);

print("testData shape: ", testData.shape);
print("classesTest: ", classesTest.shape);
print("trainingData shape: ", trainingData.shape);
print("classesTraining shape: ", classesTraining.shape);
```

```
Shapes of splitted data:
======================================================================
testData shape:  (32774, 30)
classesTest:  (32774,)
trainingData shape:  (53529, 30)
classesTraining shape:  (53529,)
```

In [154]:
```
print("Number of negative points: ", trainingData[trainingData['project
_is_approved'] == 0].shape);
print("Number of positive points: ", trainingData[trainingData['project
_is_approved'] == 1].shape);
```

```
Number of negative points:  (8105, 30)
Number of positive points:  (45424, 30)
```

In [0]:
```
vectorizedFeatureNames = [];
```

## Balancing Data

***Note: Instead of displaying whole vectorization process for balanced and imbalanced data, we have simply disabled below cell while performing analysis on imbalanced data and enabled while performing analysis on balanced data***

In [212]:
```
negativeData = trainingData[trainingData['project_is_approved'] == 0];
positiveData = trainingData[trainingData['project_is_approved'] == 1];
negativeDataBalanced = resample(negativeData, replace = True, n_samples
 = trainingData[trainingData['project_is_approved'] == 1].shape[0], ran
dom_state = 44);
trainingData = pd.concat([positiveData, negativeDataBalanced]);
trainingData = shuffle(trainingData);
classesTraining = trainingData['project_is_approved'];
print("Testing whether data is balanced: ");
equalsBorder(60);
print("Number of positive points: ", trainingData[trainingData['project
_is_approved'] == 1].shape);
print("Number of negative points: ", trainingData[trainingData['project
_is_approved'] == 0].shape);
```

```
Testing whether data is balanced:
============================================================
Number of positive points:  (45424, 30)
Number of negative points:  (45424, 30)
```

# Vectorizing categorical data

### 1. Vectorizing cleaned_categories(project_subject_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_categories
subjectsCategoriesVectorizer = CountVectorizer(vocabulary = list(sorted
CategoriesDictionary.keys()), lowercase = False, binary = True);
```

```
# Fitting CountVectorizer with cleaned_categories values
subjectsCategoriesVectorizer.fit(trainingData['cleaned_categories'].val
ues);
# Vectorizing categories using one-hot-encoding
categoriesVectors = subjectsCategoriesVectorizer.transform(trainingData
['cleaned_categories'].values);
```

In [217]:
```
print("Features used in vectorizing categories: ");
equalsBorder(70);
print(subjectsCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", categoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(categoriesVectors[0:4])
```

```
Features used in vectorizing categories:
======================================================================
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearn
ing', 'SpecialNeeds', 'Health_Sports', 'Math_Science', 'Literacy_Langua
ge']
======================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (90848, 9)
======================================================================
Sample vectors of categories:
======================================================================
  (0, 6)        1
  (1, 8)        1
  (2, 3)        1
  (3, 8)        1
```

### 2. Vectorizing cleaned_sub_categories(project_subject_sub_categories cleaned) - One Hot Encoding

In [0]:
```
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique cleaned_sub_categories
subjectsSubCategoriesVectorizer = CountVectorizer(vocabulary = list(sor
tedDictionarySubCategories.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with cleaned_sub_categories values
subjectsSubCategoriesVectorizer.fit(trainingData['cleaned_sub_categorie
s'].values);
# Vectorizing sub categories using one-hot-encoding
subCategoriesVectors = subjectsSubCategoriesVectorizer.transform(traini
ngData['cleaned_sub_categories'].values);
```

In [219]:
```
print("Features used in vectorizing subject sub categories: ");
equalsBorder(70);
print(subjectsSubCategoriesVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of cleaned_categories matrix after vectorization(one-hot-e
ncoding): ", subCategoriesVectors.shape);
equalsBorder(70);
print("Sample vectors of categories: ");
equalsBorder(70);
print(subCategoriesVectors[0:4])
```

```
Features used in vectorizing subject sub categories:
======================================================================
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolveme
nt', 'Extracurricular', 'Civics_Government', 'ForeignLanguages', 'Nutri
tionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingA
rts', 'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPre
p', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopme
nt', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Healt
h_Wellness', 'AppliedSciences', 'SpecialNeeds', 'Literature_Writing',
'Mathematics', 'Literacy']
======================================================================
Shape of cleaned_categories matrix after vectorization(one-hot-encodin
g):  (90848, 30)
======================================================================
Sample vectors of categories:
======================================================================
  (0, 13)       1
```

```
(1, 27)       1
(1, 29)       1
(2, 23)       1
(3, 27)       1
(3, 29)       1
```

### 3. Vectorizing teacher_prefix - One Hot Encoding

In [0]:
```
def giveCounter(data):
    counter = Counter();
    for dataValue in data:
        counter.update(str(dataValue).split());
    return counter
```

In [221]:
```
giveCounter(trainingData['teacher_prefix'].values)
```

Out[221]:
```
Counter({'Dr': 10, 'Mr': 8755, 'Mrs': 47006, 'Ms': 32849, 'Teacher': 22
28})
```

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(trainingData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(trainingData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(trainingData[
'teacher_prefix'].values);
```

In [223]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
ing): ", teacherPrefixVectors.shape);
```

```
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:100]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs', 'Ms', 'Mr', 'Teacher', 'Dr']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(90848, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (0, 0)        1
  (1, 0)        1
  (2, 0)        1
  (3, 1)        1
  (4, 1)        1
  (5, 1)        1
  (6, 1)        1
  (7, 1)        1
  (8, 0)        1
  (9, 1)        1
  (10, 2)       1
  (11, 0)       1
  (12, 0)       1
  (13, 1)       1
  (14, 1)       1
  (15, 0)       1
  (16, 3)       1
  (17, 1)       1
  (18, 0)       1
  (19, 0)       1
  (20, 1)       1
  (21, 2)       1
  (22, 0)       1
  (23, 0)       1
  (24, 2)       1
  :     :
```

```
(75, 0)       1
(76, 2)       1
(77, 0)       1
(78, 0)       1
(79, 0)       1
(80, 1)       1
(81, 1)       1
(82, 0)       1
(83, 0)       1
(84, 1)       1
(85, 1)       1
(86, 0)       1
(87, 1)       1
(88, 0)       1
(89, 0)       1
(90, 1)       1
(91, 0)       1
(92, 0)       1
(93, 1)       1
(94, 0)       1
(95, 2)       1
(96, 0)       1
(97, 0)       1
(98, 1)       1
(99, 0)       1
```

In [224]:
```
teacherPrefixes = [prefix.replace('.', '') for prefix in trainingData[
'teacher_prefix'].values];
teacherPrefixes[0:5]
```

Out[224]:
```
['Mrs', 'Mrs', 'Mrs', 'Ms', 'Ms']
```

In [225]:
```
trainingData['teacher_prefix'] = teacherPrefixes;
trainingData.head(3)
```

Out[225]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_s |
|---|---|---|---|---|---|

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_s |
|---|---|---|---|---|---|
| **52363** | 105141 | p145920 | ec62734af8778ec78a9a252a6cc745c7 | Mrs | TX |
| **88449** | 175840 | p193423 | aea0b2eef51510ba43089b1175f2d352 | Mrs | TN |
| **5880** | 117780 | p244178 | f13d781491aab67687a8c5df86bec29f | Mrs | UT |

3 rows × 30 columns

In [0]:
```
teacherPrefixDictionary = dict(giveCounter(trainingData['teacher_prefi
x'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique teacher_prefix
teacherPrefixVectorizer = CountVectorizer(vocabulary = list(teacherPref
ixDictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with teacher_prefix values
teacherPrefixVectorizer.fit(trainingData['teacher_prefix'].values);
# Vectorizing teacher_prefix using one-hot-encoding
teacherPrefixVectors = teacherPrefixVectorizer.transform(trainingData[
'teacher_prefix'].values);
```

In [227]:
```
print("Features used in vectorizing teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of teacher_prefix matrix after vectorization(one-hot-encod
ing): ", teacherPrefixVectors.shape);
```

```
equalsBorder(70);
print("Sample vectors of teacher_prefix: ");
equalsBorder(70);
print(teacherPrefixVectors[0:4]);
```

```
Features used in vectorizing teacher_prefix:
======================================================================
['Mrs', 'Ms', 'Mr', 'Teacher', 'Dr']
======================================================================
Shape of teacher_prefix matrix after vectorization(one-hot-encoding):
(90848, 5)
======================================================================
Sample vectors of teacher_prefix:
======================================================================
  (0, 0)        1
  (1, 0)        1
  (2, 0)        1
  (3, 1)        1
```

### 4. Vectorizing school_state - One Hot Encoding

In [0]:
```
schoolStateDictionary = dict(giveCounter(trainingData['school_state'].v
alues));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique school states
schoolStateVectorizer = CountVectorizer(vocabulary = list(schoolStateDi
ctionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with school_state values
schoolStateVectorizer.fit(trainingData['school_state'].values);
# Vectorizing school_state using one-hot-encoding
schoolStateVectors = schoolStateVectorizer.transform(trainingData['scho
ol_state'].values);
```

In [229]:
```
print("Features used in vectorizing school_state: ");
equalsBorder(70);
print(schoolStateVectorizer.get_feature_names());
equalsBorder(70);
```

```
print("Shape of school_state matrix after vectorization(one-hot-encodin
g): ", schoolStateVectors.shape);
equalsBorder(70);
print("Sample vectors of school_state: ");
equalsBorder(70);
print(schoolStateVectors[0:4]);
```

```
Features used in vectorizing school_state:
======================================================================
['TX', 'TN', 'UT', 'CA', 'IL', 'OR', 'AZ', 'MA', 'KY', 'DC', 'NC', 'I
D', 'KS', 'NJ', 'WI', 'IN', 'RI', 'SD', 'MN', 'NY', 'OH', 'AL', 'MD',
'CT', 'LA', 'SC', 'GA', 'AR', 'OK', 'MS', 'MO', 'FL', 'NV', 'MI', 'NM',
'VA', 'CO', 'WA', 'WV', 'HI', 'MT', 'NH', 'ME', 'PA', 'AK', 'DE', 'NE',
'VT', 'IA', 'WY', 'ND']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (9
0848, 51)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 1)        1
  (2, 2)        1
  (3, 3)        1
```

### 5. Vectorizing project_grade_category - One Hot Encoding

In [230]:
```
giveCounter(trainingData['project_grade_category'])
```

Out[230]:
```
Counter({'Grades3to5': 30387,
         'Grades6to8': 14171,
         'Grades9to12': 9036,
         'GradesPreKto2': 37254})
```

In [231]:
```
cleanedGrades = []
for grade in trainingData['project_grade_category'].values:
    grade = grade.replace(' ', '');
    grade = grade.replace('-', 'to');
```

```
    cleanedGrades.append(grade);
cleanedGrades[0:4]
```

Out[231]: ['Grades3to5', 'Grades3to5', 'Grades3to5', 'GradesPreKto2']

In [232]:
```
trainingData['project_grade_category'] = cleanedGrades
trainingData.head(4)
```

Out[232]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_ |
|---|---|---|---|---|---|
| **52363** | 105141 | p145920 | ec62734af8778ec78a9a252a6cc745c7 | Mrs | TX |
| **88449** | 175840 | p193423 | aea0b2eef51510ba43089b1175f2d352 | Mrs | TN |
| **5880** | 117780 | p244178 | f13d781491aab67687a8c5df86bec29f | Mrs | UT |
| **68315** | 85645 | p223648 | 3a0569a63553c413a319b41ba686c884 | Ms | CA |

4 rows × 30 columns

In [0]:
```
projectGradeDictionary = dict(giveCounter(trainingData['project_grade_c
ategory'].values));
# Using CountVectorizer for performing one-hot-encoding by setting voca
bulary as list of all unique project grade categories
projectGradeVectorizer = CountVectorizer(vocabulary = list(projectGrade
```

```
Dictionary.keys()), lowercase = False, binary = True);
# Fitting CountVectorizer with project_grade_category values
projectGradeVectorizer.fit(trainingData['project_grade_category'].value
s);
# Vectorizing project_grade_category using one-hot-encoding
projectGradeVectors = projectGradeVectorizer.transform(trainingData['pr
oject_grade_category'].values);
```

In [234]:

```
print("Features used in vectorizing project_grade_category: ");
equalsBorder(70);
print(projectGradeVectorizer.get_feature_names());
equalsBorder(70);
print("Shape of school_state matrix after vectorization(one-hot-encodin
g): ", projectGradeVectors.shape);
equalsBorder(70);
print("Sample vectors of school_state: ");
equalsBorder(70);
print(projectGradeVectors[0:4]);
```

```
Features used in vectorizing project_grade_category:
======================================================================
['Grades3to5', 'GradesPreKto2', 'Grades6to8', 'Grades9to12']
======================================================================
Shape of school_state matrix after vectorization(one-hot-encoding):  (9
0848, 4)
======================================================================
Sample vectors of school_state:
======================================================================
  (0, 0)        1
  (1, 0)        1
  (2, 0)        1
  (3, 1)        1
```

## Vectorizing numerical features

### 1. Vectorizing price

In [0]:
```
# Standardizing the price data using StandardScaler(Uses mean and std f
or standardization)
priceScaler = MinMaxScaler();
priceScaler.fit(trainingData['price'].values.reshape(-1, 1));
priceStandardized = priceScaler.transform(trainingData['price'].values.
reshape(-1, 1));
```

In [236]:
```
print("Shape of standardized matrix of prices: ", priceStandardized.sha
pe);
equalsBorder(70);
print("Sample original prices: ");
equalsBorder(70);
print(trainingData['price'].values[0:5]);
print("Sample standardized prices: ");
equalsBorder(70);
print(priceStandardized[0:5]);
```

```
Shape of standardized matrix of prices:  (90848, 1)
======================================================================
Sample original prices:
======================================================================
[ 59.98  55.24 977.55  99.42 189.57]
Sample standardized prices:
======================================================================
[[0.00593298]
 [0.00545891]
 [0.09770522]
 [0.00987764]
 [0.01889414]]
```

**2. Vectorizing quantity**

In [0]:
```
# Standardizing the quantity data using StandardScaler(Uses mean and st
d for standardization)
quantityScaler = MinMaxScaler();
quantityScaler.fit(trainingData['quantity'].values.reshape(-1, 1));
quantityStandardized = quantityScaler.transform(trainingData['quantity'
].values.reshape(-1, 1));
```

In [238]:
```
print("Shape of standardized matrix of quantities: ", quantityStandardized.shape);
equalsBorder(70);
print("Sample original quantities: ");
equalsBorder(70);
print(trainingData['quantity'].values[0:5]);
print("Sample standardized quantities: ");
equalsBorder(70);
print(quantityStandardized[0:5]);
```

```
Shape of standardized matrix of quantities:  (90848, 1)
======================================================================
Sample original quantities:
======================================================================
[ 33 100  60  12  18]
Sample standardized quantities:
======================================================================
[[0.03444564]
 [0.1065662 ]
 [0.06350915]
 [0.01184069]
 [0.01829925]]
```

#### 3. Vectorizing teacher_number_of_previously_posted_projects

In [0]:
```
# Standardizing the teacher_number_of_previously_posted_projects data using StandardScaler(Uses mean and std for standardization)
previouslyPostedScaler = MinMaxScaler();
previouslyPostedScaler.fit(trainingData['teacher_number_of_previously_posted_projects'].values.reshape(-1, 1));
previouslyPostedStandardized = previouslyPostedScaler.transform(trainingData['teacher_number_of_previously_posted_projects'].values.reshape(-1, 1));
```

In [240]:
```
print("Shape of standardized matrix of teacher_number_of_previously_posted_projects: ", previouslyPostedStandardized.shape);
```

```
equalsBorder(70);
print("Sample original number of previously posted projects: ");
equalsBorder(70);
print(trainingData['teacher_number_of_previously_posted_projects'].valu
es[0:5]);
print("Sample standardized teacher_number_of_previously_posted_project
s: ");
equalsBorder(70);
print(previouslyPostedStandardized[0:5]);
```

```
Shape of standardized matrix of teacher_number_of_previously_posted_pro
jects:  (90848, 1)
======================================================================
Sample original number of previously posted projects:
======================================================================
[1 0 0 0 2]
Sample standardized teacher_number_of_previously_posted_projects:
======================================================================
[[0.00221729]
 [0.        ]
 [0.        ]
 [0.        ]
 [0.00443459]]
```

#### 4. Vectorizing number_of_words_in_title

In [0]:
```
numberOfWordsInTitleScaler = MinMaxScaler();
numberOfWordsInTitleScaler.fit(trainingData['number_of_words_in_title']
.values.reshape(-1, 1));
numberOfWordsInTitleStandardized = numberOfWordsInTitleScaler.transform
(trainingData['number_of_words_in_title'].values.reshape(-1, 1));
```

In [242]:
```
print("Shape of standardized matrix of number_of_words_in_title: ", num
berOfWordsInTitleStandardized.shape);
equalsBorder(70);
print("Sample original number of words in title: ");
equalsBorder(70);
print(trainingData['number_of_words_in_title'].values[0:5]);
```

```
print("Sample standardized number_of_words_in_title: ");
equalsBorder(70);
print(numberOfWordsInTitleStandardized[0:5]);
```

```
Shape of standardized matrix of number_of_words_in_title:  (90848, 1)
======================================================================
Sample original number of words in title:
======================================================================
[2 4 4 5 2]
Sample standardized number_of_words_in_title:
======================================================================
[[0.1]
 [0.3]
 [0.3]
 [0.4]
 [0.1]]
```

#### 5. Vectorizing number_of_words_in_essay

In [0]:
```
numberOfWordsInEssayScaler = MinMaxScaler();
numberOfWordsInEssayScaler.fit(trainingData['number_of_words_in_essay']
.values.reshape(-1, 1));
numberOfWordsInEssayStandardized = numberOfWordsInEssayScaler.transform
(trainingData['number_of_words_in_essay'].values.reshape(-1, 1));
```

In [244]:
```
print("Shape of standardized matrix of number_of_words_in_essay: ", num
berOfWordsInEssayStandardized.shape);
equalsBorder(70);
print("Sample original number of words in essay: ");
equalsBorder(70);
print(trainingData['number_of_words_in_essay'].values[0:5]);
print("Sample standardized number_of_words_in_essay: ");
equalsBorder(70);
print(numberOfWordsInEssayStandardized[0:5]);
```

```
Shape of standardized matrix of number_of_words_in_essay:  (90848, 1)
======================================================================
Sample original number of words in essay:
```

```
==========================================================================
[225 112 100 140 159]
Sample standardized number_of_words_in_essay:
==========================================================================
[[0.63786008]
 [0.17283951]
 [0.12345679]
 [0.28806584]
 [0.36625514]]
```

In [0]:

```
numberOfPoints = previouslyPostedStandardized.shape[0];
# Categorical data
categoriesVectorsSub = categoriesVectors[0:numberOfPoints];
subCategoriesVectorsSub = subCategoriesVectors[0:numberOfPoints];
teacherPrefixVectorsSub = teacherPrefixVectors[0:numberOfPoints];
schoolStateVectorsSub = schoolStateVectors[0:numberOfPoints];
projectGradeVectorsSub = projectGradeVectors[0:numberOfPoints];


# Numerical data
priceStandardizedSub = priceStandardized[0:numberOfPoints];
quantityStandardizedSub = quantityStandardized[0:numberOfPoints];
previouslyPostedStandardizedSub = previouslyPostedStandardized[0:number
OfPoints];
numberOfWordsInTitleStandardizedSub = numberOfWordsInTitleStandardized[
0:numberOfPoints];
numberOfWordsInEssayStandardizedSub = numberOfWordsInEssayStandardized[
0:numberOfPoints];
positiveSentimentScoreSub = trainingData['positive_sentiment_score'].va
lues[0:numberOfPoints].reshape(-1, 1);
negativeSentimentScoreSub = trainingData['negative_sentiment_score'].va
lues[0:numberOfPoints].reshape(-1, 1);
neutralSentimentScoreSub = trainingData['neutral_sentiment_score'].valu
es[0:numberOfPoints].reshape(-1, 1);
compoundSentimentScoreSub = trainingData['compound_sentiment_score'].va
lues[0:numberOfPoints].reshape(-1, 1);
```

## Preparing cross validate data for analysis

In [0]:
```
# Test data categorical features transformation
categoriesTransformedCrossValidateData = subjectsCategoriesVectorizer.t
ransform(crossValidateData['cleaned_categories']);
subCategoriesTransformedCrossValidateData = subjectsSubCategoriesVector
izer.transform(crossValidateData['cleaned_sub_categories']);
teacherPrefixTransformedCrossValidateData = teacherPrefixVectorizer.tra
nsform(crossValidateData['teacher_prefix']);
schoolStateTransformedCrossValidateData = schoolStateVectorizer.transfo
rm(crossValidateData['school_state']);
projectGradeTransformedCrossValidateData = projectGradeVectorizer.trans
form(crossValidateData['project_grade_category']);


# Test data numerical features transformation
priceTransformedCrossValidateData = priceScaler.transform(crossValidate
Data['price'].values.reshape(-1, 1));
quantityTransformedCrossValidateData = quantityScaler.transform(crossVa
lidateData['quantity'].values.reshape(-1, 1));
previouslyPostedTransformedCrossValidateData = previouslyPostedScaler.t
ransform(crossValidateData['teacher_number_of_previously_posted_project
s'].values.reshape(-1, 1));
numberOfWordsInTitleTransformedCrossValidateData = numberOfWordsInTitle
Scaler.transform(crossValidateData['number_of_words_in_title'].values.r
eshape(-1, 1));
numberOfWordsInEssayTransformedCrossValidateData = numberOfWordsInEssay
Scaler.transform(crossValidateData['number_of_words_in_essay'].values.r
eshape(-1, 1));
positiveSentimentScoreCrossValidateData = crossValidateData['positive_s
entiment_score'].values.reshape(-1, 1);
negativeSentimentScoreCrossValidateData = crossValidateData['negative_s
entiment_score'].values.reshape(-1, 1);
neutralSentimentScoreCrossValidateData = crossValidateData['neutral_sen
timent_score'].values.reshape(-1, 1);
compoundSentimentScoreCrossValidateData = crossValidateData['compound_s
entiment_score'].values.reshape(-1, 1);
```

## Preparing Test data for analysis

In [0]:
```
# Test data categorical features transformation
categoriesTransformedTestData = subjectsCategoriesVectorizer.transform(
testData['cleaned_categories']);
subCategoriesTransformedTestData = subjectsSubCategoriesVectorizer.tran
sform(testData['cleaned_sub_categories']);
teacherPrefixTransformedTestData = teacherPrefixVectorizer.transform(te
stData['teacher_prefix']);
schoolStateTransformedTestData = schoolStateVectorizer.transform(testDa
ta['school_state']);
projectGradeTransformedTestData = projectGradeVectorizer.transform(test
Data['project_grade_category']);

# Test data numerical features transformation
priceTransformedTestData = priceScaler.transform(testData['price'].valu
es.reshape(-1, 1));
quantityTransformedTestData = quantityScaler.transform(testData['quanti
ty'].values.reshape(-1, 1));
previouslyPostedTransformedTestData = previouslyPostedScaler.transform(
testData['teacher_number_of_previously_posted_projects'].values.reshape
(-1, 1));
numberOfWordsInTitleTransformedTestData = numberOfWordsInTitleScaler.tr
ansform(testData['number_of_words_in_title'].values.reshape(-1, 1));
numberOfWordsInEssayTransformedTestData = numberOfWordsInEssayScaler.tr
ansform(testData['number_of_words_in_essay'].values.reshape(-1, 1));
positiveSentimentScoreTestData = testData['positive_sentiment_score'].v
alues.reshape(-1, 1);
negativeSentimentScoreTestData = testData['negative_sentiment_score'].v
alues.reshape(-1, 1);
neutralSentimentScoreTestData = testData['neutral_sentiment_score'].val
ues.reshape(-1, 1);
compoundSentimentScoreTestData = testData['compound_sentiment_score'].v
alues.reshape(-1, 1);
```

## Classification using data with no text features by logistic regression(l1 regularizer)

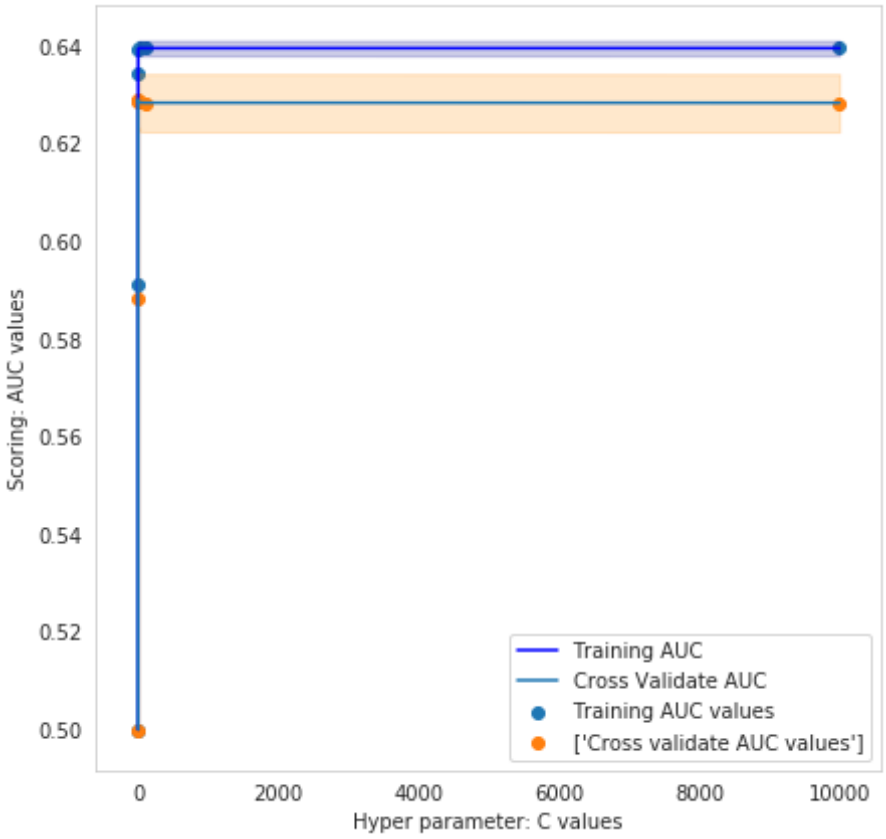
In [207]:

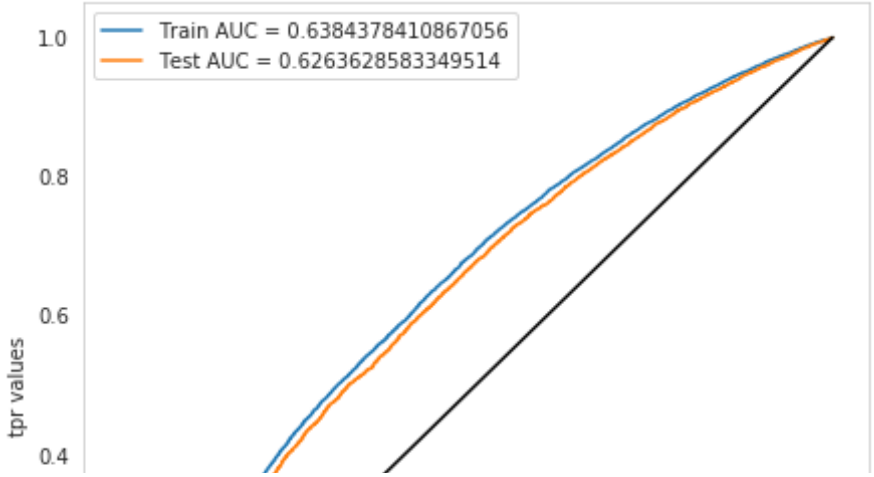
```
techniques = ['without text features'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
                                 teacherPrefixVectorsSub,\
                                 schoolStateVectorsSub,\
                                 projectGradeVectorsSub,\
                                 priceStandardizedSub,\
                                 previouslyPostedStandardizedSub,\
                                 numberOfWordsInTitleStandardizedSub,\
                                 numberOfWordsInEssayStandardizedSub,\
                                 positiveSentimentScoreSub,\
                                 negativeSentimentScoreSub,\
                                 neutralSentimentScoreSub,\
                                 compoundSentimentScoreSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                      subCategoriesTransformedCrossVali
dateData,\
                                      teacherPrefixTransformedCrossVali
dateData,\
                                      schoolStateTransformedCrossValida
teData,\
                                      projectGradeTransformedCrossValid
ateData,\
                                      priceTransformedCrossValidateData
,\
                                      previouslyPostedTransformedCrossV
alidateData,\
                                      numberOfWordsInTitleTransformedCr
ossValidateData,\
                                      numberOfWordsInEssayTransformedCr
ossValidateData,\
                                      positiveSentimentScoreCrossValida
teData,\
                                      negativeSentimentScoreCrossValida
teData,\
                                      neutralSentimentScoreCrossValidat
eData,\
```

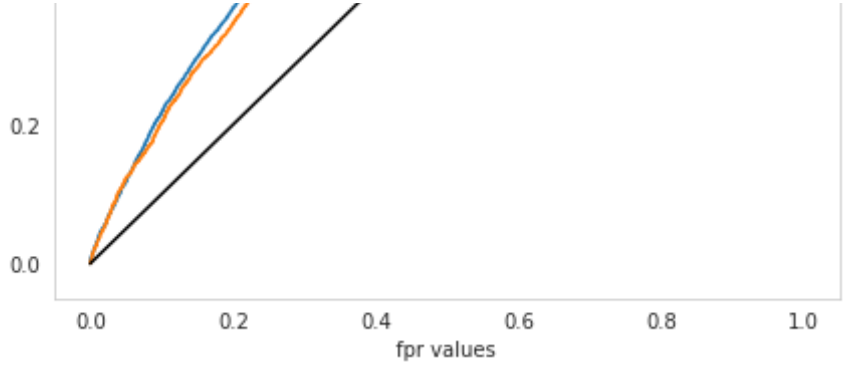
```
                                                compoundSentimentScoreCrossValida
teData));
    testMergedData = hstack((categoriesTransformedTestData,\
                             subCategoriesTransformedTestData,\
                             teacherPrefixTransformedTestData,\
                             schoolStateTransformedTestData,\
                             projectGradeTransformedTestData,\
                             priceTransformedTestData,\
                             previouslyPostedTransformedTestData,\
                             numberOfWordsInTitleTransformedTestData,\
                             numberOfWordsInEssayTransformedTestData,\
                             positiveSentimentScoreTestData,\
                             negativeSentimentScoreTestData,\
                             neutralSentimentScoreTestData,\
                             compoundSentimentScoreTestData));

    lrClassifier = LogisticRegression(penalty = 'l1');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
```

```
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l1', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using data {}  using logistic regression
 classifier: ".format(technique));
    equalsBorder(70);
    print("AUC values of train data: ");
```

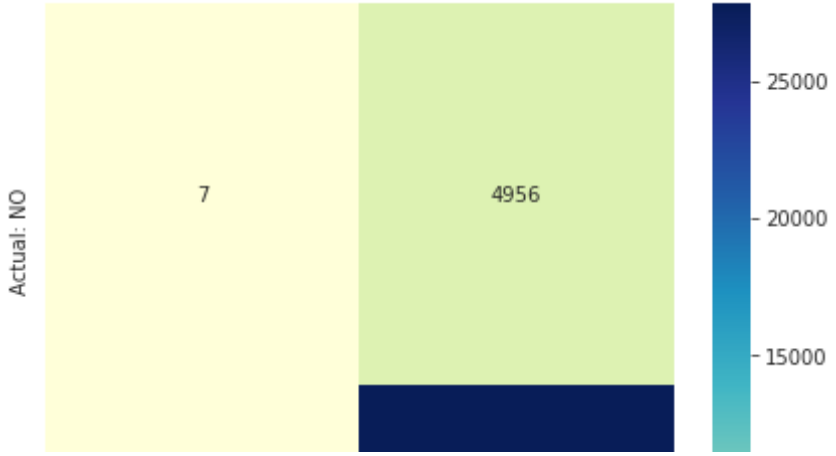
```
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l1)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```

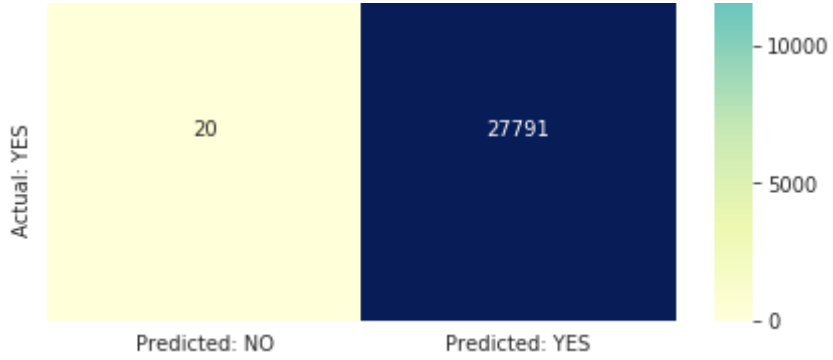





```
Results of analysis using data without text features  using logistic re
gression classifier:
========================================================================
AUC values of train data:
========================================
[0.5        0.5909744  0.6343117  0.63934334 0.63953178 0.63952865
 0.63952882]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.6263628583349514
========================================
Confusion Matrix :
============================================================
```






## Classification using data with no text features by logistic regression(l2 regularizer)

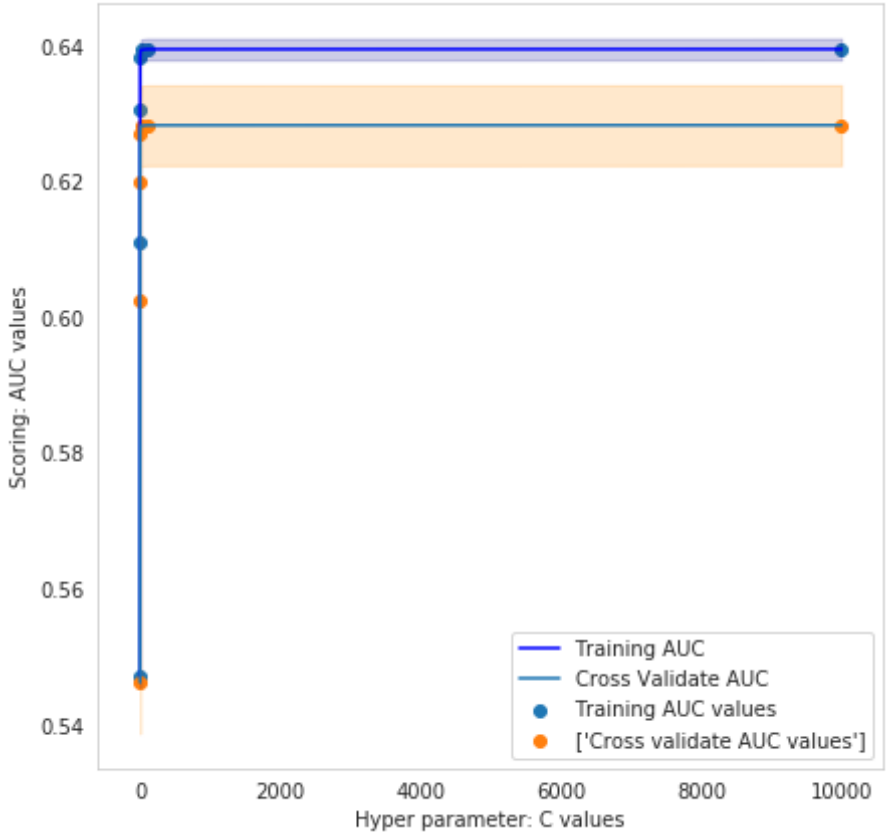
In [208]:
```
techniques = ['without text features'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
                                 teacherPrefixVectorsSub,\
                                 schoolStateVectorsSub,\
                                 projectGradeVectorsSub,\
                                 priceStandardizedSub,\
                                 previouslyPostedStandardizedSub,\
                                 numberOfWordsInTitleStandardizedSub,\
                                 numberOfWordsInEssayStandardizedSub,\
                                 positiveSentimentScoreSub,\
                                 negativeSentimentScoreSub,\
                                 neutralSentimentScoreSub,\
                                 compoundSentimentScoreSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                      subCategoriesTransformedCrossVali
dateData,\
                                      teacherPrefixTransformedCrossVali
dateData,\
                                      schoolStateTransformedCrossValida
```

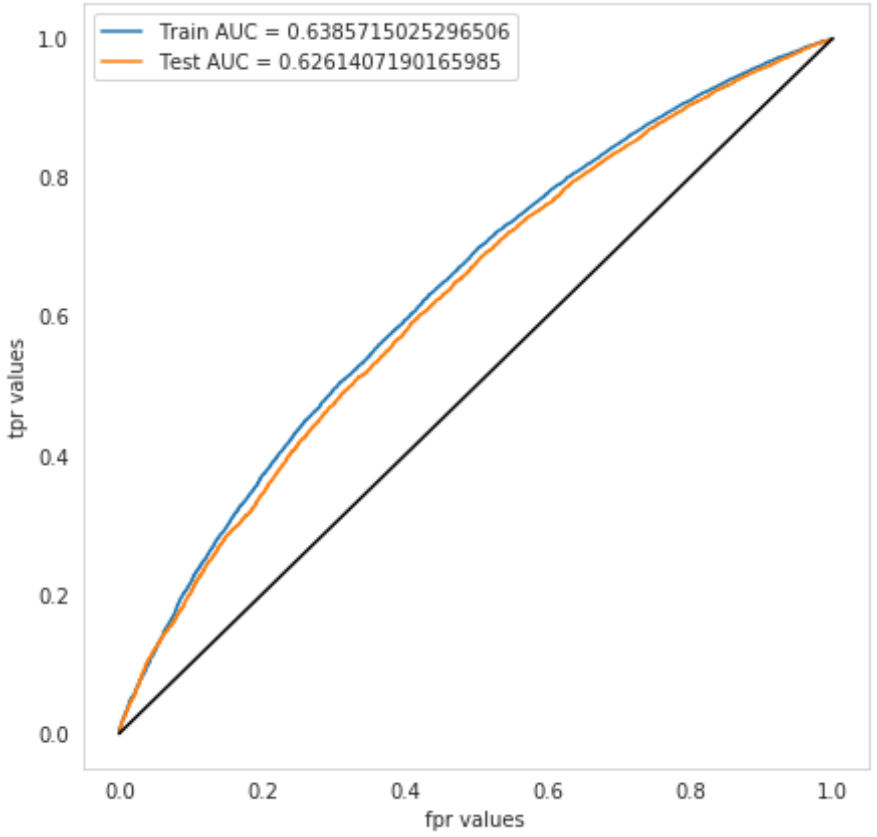
```
teData,\
                                        projectGradeTransformedCrossValid
ateData,\
                                        priceTransformedCrossValidateData
,\
                                        previouslyPostedTransformedCrossV
alidateData,\
                                        numberOfWordsInTitleTransformedCr
ossValidateData,\
                                        numberOfWordsInEssayTransformedCr
ossValidateData,\
                                        positiveSentimentScoreCrossValida
teData,\
                                        negativeSentimentScoreCrossValida
teData,\
                                        neutralSentimentScoreCrossValidat
eData,\
                                        compoundSentimentScoreCrossValida
teData));
    testMergedData = hstack((categoriesTransformedTestData,\
                             subCategoriesTransformedTestData,\
                             teacherPrefixTransformedTestData,\
                             schoolStateTransformedTestData,\
                             projectGradeTransformedTestData,\
                             priceTransformedTestData,\
                             previouslyPostedTransformedTestData,\
                             numberOfWordsInTitleTransformedTestData,\
                             numberOfWordsInEssayTransformedTestData,\
                             positiveSentimentScoreTestData,\
                             negativeSentimentScoreTestData,\
                             neutralSentimentScoreTestData,\
                             compoundSentimentScoreTestData));

    lrClassifier = LogisticRegression(penalty = 'l2');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);
```

```
    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l2', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);
```

```
    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using data {}  using logistic regression
 classifier: ".format(technique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
```

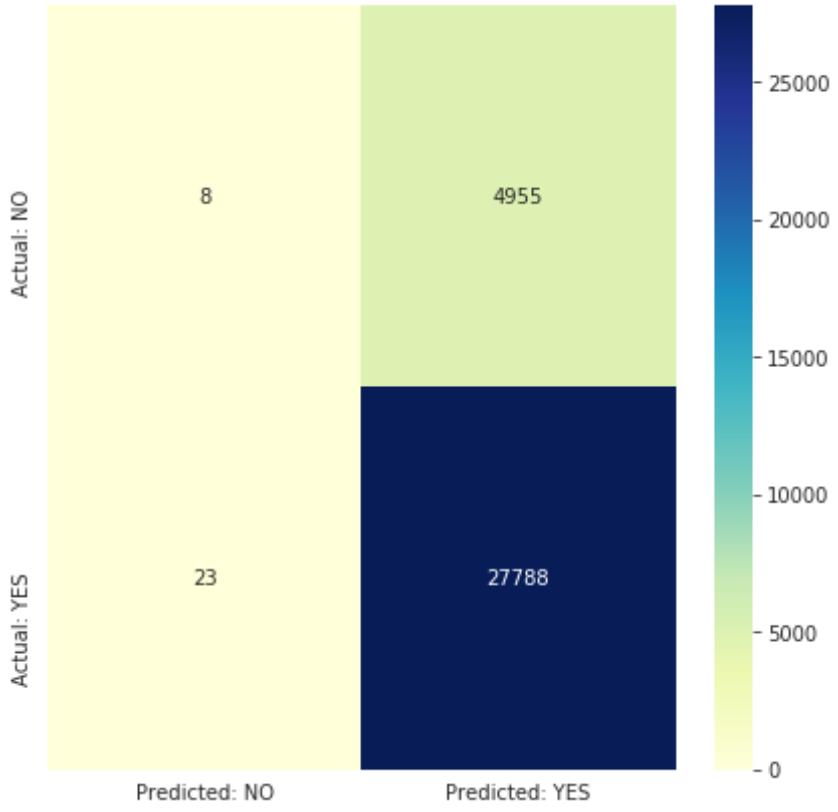
```
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l1)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```






```
Results of analysis using data without text features  using logistic re
gression classifier:
========================================================================
AUC values of train data:
========================================
[0.54713727 0.61111339 0.63071284 0.63834498 0.63940997 0.63951588
 0.63952582]
========================================
Optimal Hyper parameter Value:  10000
========================================
AUC value of test data:  0.6261407190165985
========================================
Confusion Matrix :
```

==========




## Classification using balanced data with no text features by logistic regression(l1 regularizer)

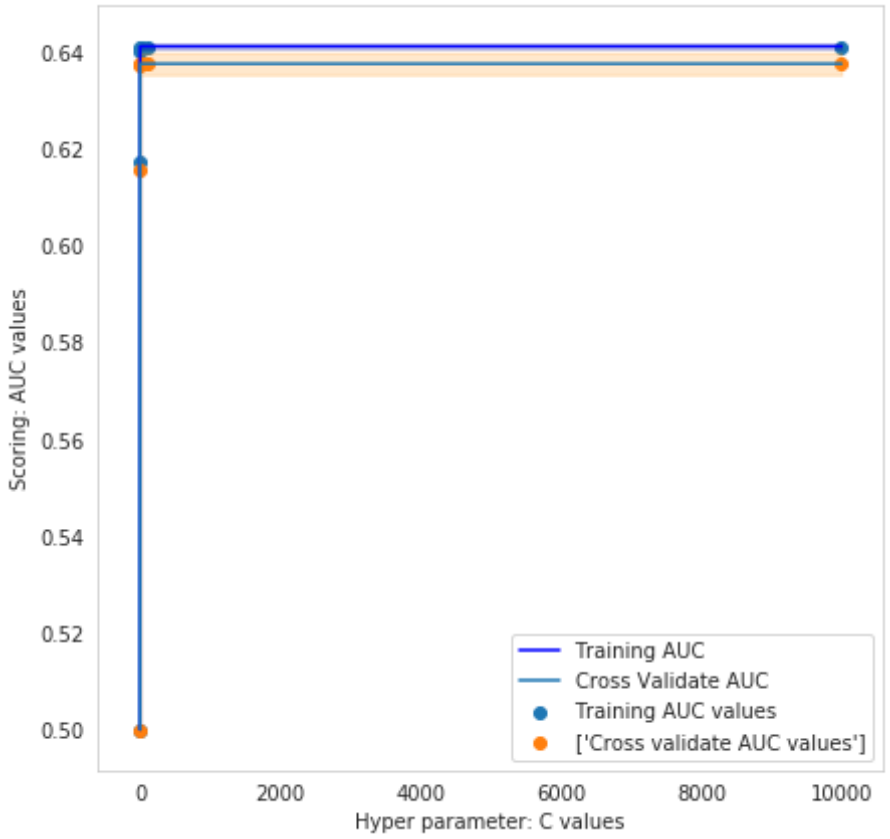
In [249]:
```
techniques = ['without text features'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
                                 teacherPrefixVectorsSub,\
```

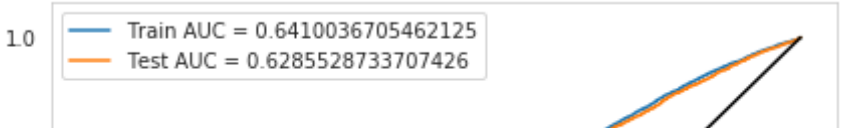
```
                                    schoolStateVectorsSub,\
                                    projectGradeVectorsSub,\
                                    priceStandardizedSub,\
                                    previouslyPostedStandardizedSub,\
                                    numberOfWordsInTitleStandardizedSub,\
                                    numberOfWordsInEssayStandardizedSub,\
                                    positiveSentimentScoreSub,\
                                    negativeSentimentScoreSub,\
                                    neutralSentimentScoreSub,\
                                    compoundSentimentScoreSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                      subCategoriesTransformedCrossVali
dateData,\
                                      teacherPrefixTransformedCrossVali
dateData,\
                                      schoolStateTransformedCrossValida
teData,\
                                      projectGradeTransformedCrossValid
ateData,\
                                      priceTransformedCrossValidateData
,\
                                      previouslyPostedTransformedCrossV
alidateData,\
                                      numberOfWordsInTitleTransformedCr
ossValidateData,\
                                      numberOfWordsInEssayTransformedCr
ossValidateData,\
                                      positiveSentimentScoreCrossValida
teData,\
                                      negativeSentimentScoreCrossValida
teData,\
                                      neutralSentimentScoreCrossValidat
eData,\
                                      compoundSentimentScoreCrossValida
teData));
    testMergedData = hstack((categoriesTransformedTestData,\
                             subCategoriesTransformedTestData,\
                             teacherPrefixTransformedTestData,\
```

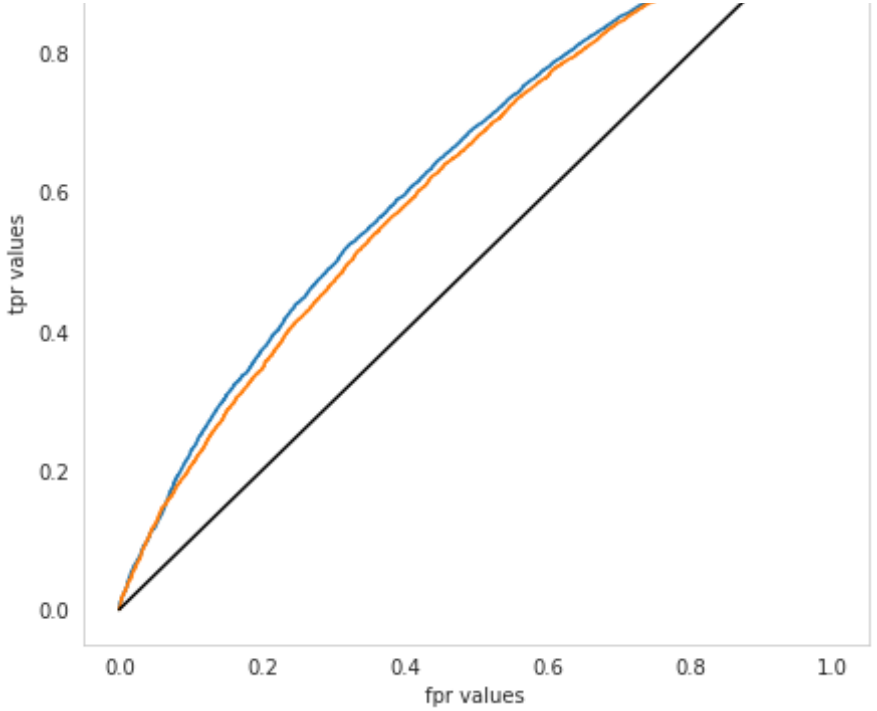
```
                                schoolStateTransformedTestData,\
                                projectGradeTransformedTestData,\
                                priceTransformedTestData,\
                                previouslyPostedTransformedTestData,\
                                numberOfWordsInTitleTransformedTestData,\
                                numberOfWordsInEssayTransformedTestData,\
                                positiveSentimentScoreTestData,\
                                negativeSentimentScoreTestData,\
                                neutralSentimentScoreTestData,\
                                compoundSentimentScoreTestData));

    lrClassifier = LogisticRegression(penalty = 'l1');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
    plt.xlabel('Hyper parameter: C values');
```

```
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l1', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using data {}  using logistic regression
 classifier: ".format(technique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
    equalsBorder(40);
```

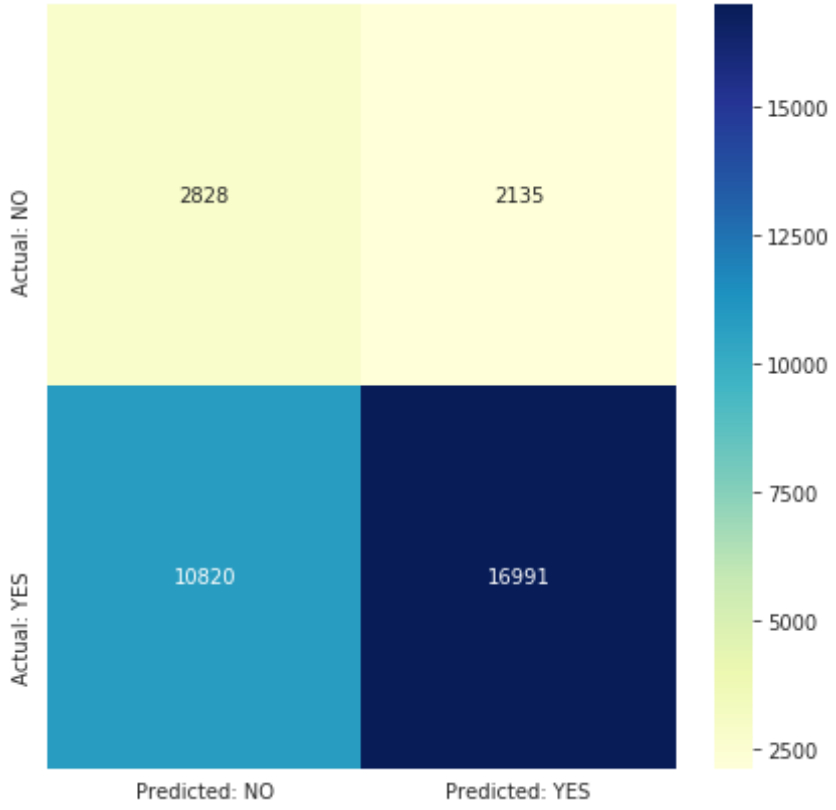
```
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l1)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```






```
Results of analysis using data without text features  using logistic re
gression classifier:
========================================================================
AUC values of train data:
========================================
[0.5        0.6172983  0.64034065 0.64127787 0.64130945 0.6413104
 0.64131048]
========================================
Optimal Hyper parameter Value:  1
========================================
AUC value of test data:  0.6285528733707426
========================================
Confusion Matrix :
============================================================
```




## Classification using balanced data with no text features by logistic regression(l2 regularizer)

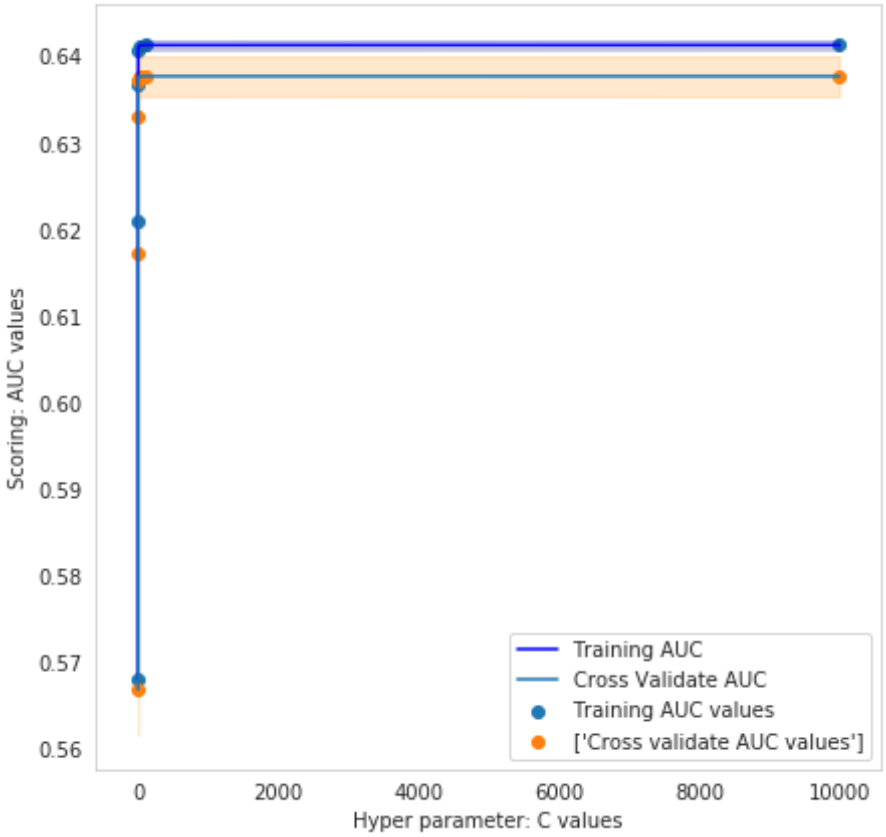
In [250]:
```
techniques = ['without text features'];
for index, technique in enumerate(techniques):
    trainingMergedData = hstack((categoriesVectorsSub,\
                                 subCategoriesVectorsSub,\
```

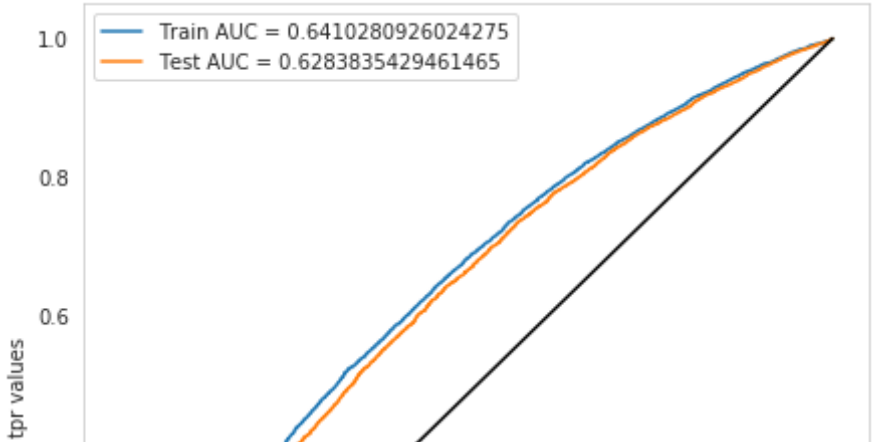
```
                                    teacherPrefixVectorsSub,\
                                    schoolStateVectorsSub,\
                                    projectGradeVectorsSub,\
                                    priceStandardizedSub,\
                                    previouslyPostedStandardizedSub,\
                                    numberOfWordsInTitleStandardizedSub,\
                                    numberOfWordsInEssayStandardizedSub,\
                                    positiveSentimentScoreSub,\
                                    negativeSentimentScoreSub,\
                                    neutralSentimentScoreSub,\
                                    compoundSentimentScoreSub));
    crossValidateMergedData = hstack((categoriesTransformedCrossValidat
eData,\
                                         subCategoriesTransformedCrossVali
dateData,\
                                         teacherPrefixTransformedCrossVali
dateData,\
                                         schoolStateTransformedCrossValida
teData,\
                                         projectGradeTransformedCrossValid
ateData,\
                                         priceTransformedCrossValidateData
,\
                                         previouslyPostedTransformedCrossV
alidateData,\
                                         numberOfWordsInTitleTransformedCr
ossValidateData,\
                                         numberOfWordsInEssayTransformedCr
ossValidateData,\
                                         positiveSentimentScoreCrossValida
teData,\
                                         negativeSentimentScoreCrossValida
teData,\
                                         neutralSentimentScoreCrossValidat
eData,\
                                         compoundSentimentScoreCrossValida
teData));
    testMergedData = hstack((categoriesTransformedTestData,\
                             subCategoriesTransformedTestData,\
```

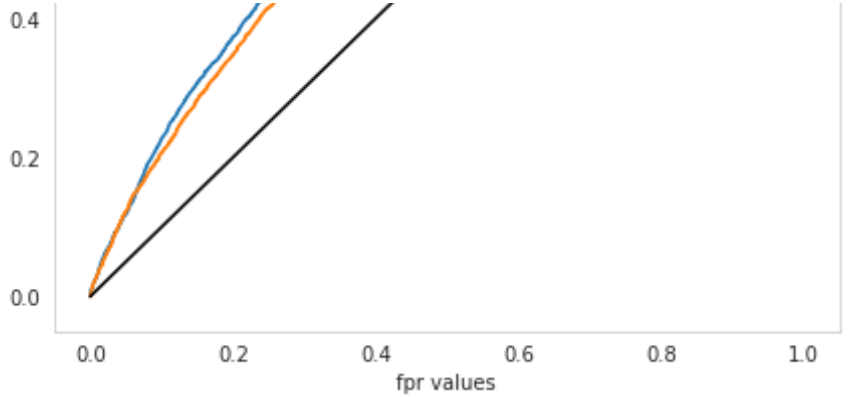
```
                                teacherPrefixTransformedTestData,\
                                schoolStateTransformedTestData,\
                                projectGradeTransformedTestData,\
                                priceTransformedTestData,\
                                previouslyPostedTransformedTestData,\
                                numberOfWordsInTitleTransformedTestData,\
                                numberOfWordsInEssayTransformedTestData,\
                                positiveSentimentScoreTestData,\
                                negativeSentimentScoreTestData,\
                                neutralSentimentScoreTestData,\
                                compoundSentimentScoreTestData));

    lrClassifier = LogisticRegression(penalty = 'l2');
    tunedParameters = {'C': [0.0001, 0.01, 0.1, 1, 10, 100, 10000]};
    classifier = GridSearchCV(lrClassifier, tunedParameters, cv = 5, sc
oring = 'roc_auc');
    classifier.fit(trainingMergedData, classesTraining);

    trainingAucMeanValues = classifier.cv_results_['mean_train_score'];
    trainingAucStdValues = classifier.cv_results_['std_train_score'];
    crossValidateAucMeanValues = classifier.cv_results_['mean_test_scor
e'];
    crossValidateAucStdValues = classifier.cv_results_['std_test_score'
];

    plt.plot(tunedParameters['C'], trainingAucMeanValues, 'b', label =
"Training AUC");
    plt.plot(tunedParameters['C'], crossValidateAucMeanValues, label =
"Cross Validate AUC");
    plt.scatter(tunedParameters['C'], trainingAucMeanValues, label = 'T
raining AUC values');
    plt.scatter(tunedParameters['C'], crossValidateAucMeanValues, label
 = ['Cross validate AUC values']);
    plt.gca().fill_between(tunedParameters['C'], trainingAucMeanValues
- trainingAucStdValues, trainingAucMeanValues + trainingAucStdValues, a
lpha = 0.2, color = 'darkblue');
    plt.gca().fill_between(tunedParameters['C'], crossValidateAucMeanVa
lues - crossValidateAucStdValues, crossValidateAucMeanValues + crossVal
idateAucStdValues, alpha = 0.2, color = 'darkorange');
```

```
    plt.xlabel('Hyper parameter: C values');
    plt.ylabel('Scoring: AUC values');
    plt.grid();
    plt.legend();
    plt.show();

    optimalHypParamValue = classifier.best_params_['C'];
    lrClassifier = LogisticRegression(penalty = 'l2', C = optimalHypPar
amValue);
    lrClassifier.fit(trainingMergedData, classesTraining);
    predProbScoresTraining = lrClassifier.predict_proba(trainingMergedD
ata);
    fprTrain, tprTrain, thresholdTrain = roc_curve(classesTraining, pre
dProbScoresTraining[:, 1]);
    predProbScoresTest = lrClassifier.predict_proba(testMergedData);
    fprTest, tprTest, thresholdTest = roc_curve(classesTest, predProbSc
oresTest[:, 1]);

    plt.plot(fprTrain, tprTrain, label = "Train AUC = " + str(auc(fprTr
ain, tprTrain)));
    plt.plot(fprTest, tprTest, label = "Test AUC = " + str(auc(fprTest,
 tprTest)));
    plt.plot([0, 1], [0, 1], 'k-');
    plt.xlabel("fpr values");
    plt.ylabel("tpr values");
    plt.grid();
    plt.legend();
    plt.show();

    areaUnderRocValueTest = auc(fprTest, tprTest);

    print("Results of analysis using data {}  using logistic regression
 classifier: ".format(technique));
    equalsBorder(70);
    print("AUC values of train data: ");
    equalsBorder(40);
    print(trainingAucMeanValues);
    equalsBorder(40);
    print("Optimal Hyper parameter Value: ", optimalHypParamValue);
```

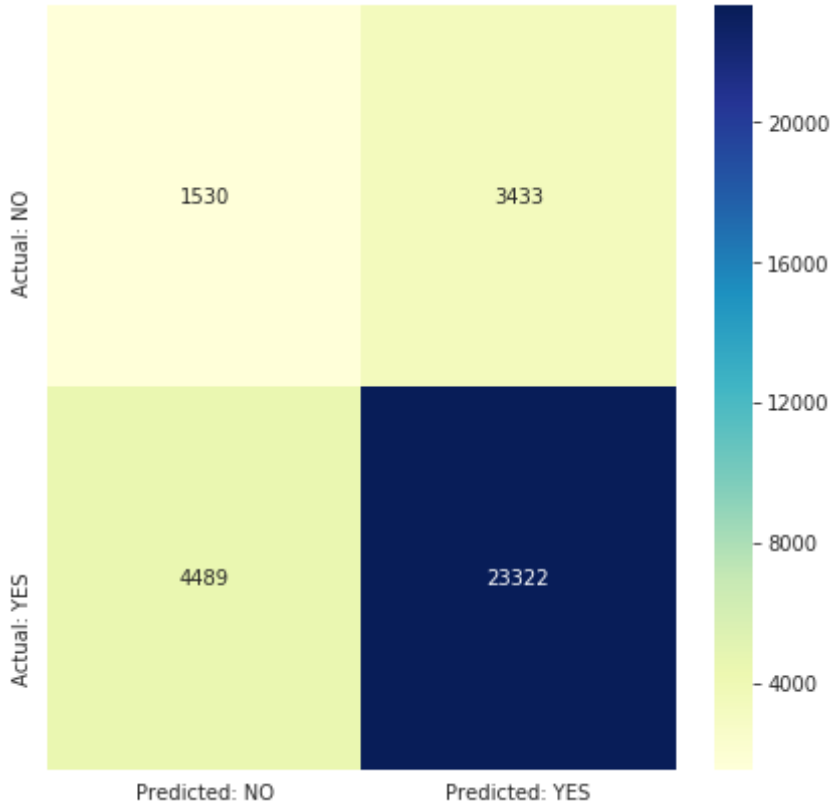
```
    equalsBorder(40);
    print("AUC value of test data: ", str(areaUnderRocValueTest));
    # Predicting classes of test data projects
    predictionClassesTest = lrClassifier.predict(testMergedData);
    equalsBorder(40);
    # Printing confusion matrix
    confusionMatrix = confusion_matrix(classesTest, predictionClassesTe
st);
    # Creating dataframe for generated confusion matrix
    confusionMatrixDataFrame = pd.DataFrame(data = confusionMatrix, ind
ex = ['Actual: NO', 'Actual: YES'], columns = ['Predicted: NO', 'Predic
ted: YES']);
    print("Confusion Matrix : ");
    equalsBorder(60);
    sbrn.heatmap(confusionMatrixDataFrame, annot = True, fmt = 'd', cma
p="YlGnBu");
    plt.show();
    # Adding results to results dataframe
    logisticResultsDataFrame = logisticResultsDataFrame.append({'Vector
izer': technique, 'Model': 'Logistic Regression(l1)', 'Hyper Parameter
 - C': optimalHypParamValue, 'AUC': areaUnderRocValueTest}, ignore_inde
x = True);
```






```
Results of analysis using data without text features  using logistic re
gression classifier:
========================================================================
AUC values of train data:
========================================
[0.56814085 0.62087497 0.63679331 0.64079708 0.64125789 0.64130574
 0.64131183]
========================================
Optimal Hyper parameter Value:  100
========================================
AUC value of test data:  0.6283835429461465
========================================
Confusion Matrix :
============================================================
```




## Summary of results of above classification using Multinomial Naive bayes

In [13]:
```
logisticResultsDataFrame
```

Out[13]:

| | Vectorizer | Model | Hyper Parameter - C | AUC | Data |
|---|---|---|---|---|---|
| **0** | Bag of words | Logistic Regression(l1) | 0.10 | 0.724 | Imbalanced |
| **1** | Tf-Idf | Logistic Regression(l1) | 1.00 | 0.722 | Imbalanced |
| **2** | Average Word2Vector | Logistic Regression(l1) | 10.00 | 0.706 | Imbalanced |
| **3** | Tf-Idf Weighted Word2Vector | Logistic Regression(l1) | 1.00 | 0.702 | Imbalanced |
| **4** | Bag of words | Logistic Regression(l2) | 0.01 | 0.719 | Imbalanced |
| **5** | Tf-Idf | Logistic Regression(l2) | 1.00 | 0.716 | Imbalanced |
| **6** | Average Word2Vector | Logistic Regression(l2) | 10.00 | 0.707 | Imbalanced |
| **7** | Tf-Idf Weighted Word2Vector | Logistic Regression(l2) | 1.00 | 0.701 | Imbalanced |
| **8** | Bag of words | Logistic Regression(l1) | 1.00 | 0.669 | Balanced |
| **9** | Tf-Idf | Logistic Regression(l1) | 10.00 | 0.661 | Balanced |
| **10** | Bag of words | Logistic Regression(l2) | 10.00 | 0.658 | Balanced |
| **11** | Tf-Idf | Logistic Regression(l2) | 10.00 | 0.670 | Balanced |

| | Vectorizer | Model | Hyper Parameter - C | AUC | Data |
|---|---|---|---|---|---|
| **12** | No technique | Logistic Regression(l1) | 1.00 | 0.626 | ImBalanced(no text features) |
| **13** | No technique | Logistic Regression(l2) | 10000.00 | 0.626 | ImBalanced(no text features) |
| **14** | No technique | Logistic Regression(l1) | 1.00 | 0.628 | Balanced(no text features) |
| **15** | No technique | Logistic Regression(l2) | 100.00 | 0.628 | Balanced(no text features) |

## Conclusions of above analysis

1. As from above analysis you can see that tf-idf with logistic regression(l1) on imbalanced data is best as the auc value obtained is larger than all. But the classification is more bias towards positive class as there are more number of positive points. The hyper parameter value would be 1.
2. The best considerable model would be tf-idf with logistic regression(l2) on balanced data as the auc value is considerably good and both negative and positive points are classified without bias. Here the best hyper parameter is 10.
3. The analysis on data without text features resulted in very low auc value and so from this we can say that text features play important role in constructing good model.