

DonorsChoose

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

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

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

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

About the DonorsChoose Data Set

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

Feature	Description
<code>project_id</code>	A unique identifier for the proposed project. Example: p036502
<code>project_title</code>	Title of the project. Examples: <ul style="list-style-type: none">• Art Will Make You Happy!• First Grade Fun
<code>project_grade_category</code>	Grade level of students for which the project is targeted. One of the following enumerated values: <ul style="list-style-type: none">• Grades PreK-2• Grades 3-5• Grades 6-8• Grades 9-12
<code>project_subject_categories</code>	One or more (comma-separated) subject categories for the project from the following enumerated list of values: <ul style="list-style-type: none">• Applied Learning• Care & Hunger• Health & Sports• History & Civics• Literacy & Language• Math & Science• Music & The Arts• Special Needs• Warmth Examples: <ul style="list-style-type: none">• Music & The Arts• Literacy & Language, Math & Science
<code>school_state</code>	State where school is located (Two-letter U.S. postal code). Example: WY
<code>project_subject_subcategories</code>	One or more (comma-separated) subject subcategories for the project. Examples: <ul style="list-style-type: none">• Literacy

Feature	Description
<code>project_resource_summary</code>	An explanation of the resources needed for the project. Example: <ul style="list-style-type: none"> My students need hands on literacy materials to manage sensory needs!
<code>project_essay_1</code>	First application essay*
<code>project_essay_2</code>	Second application essay*
<code>project_essay_3</code>	Third application essay*
<code>project_essay_4</code>	Fourth application essay*
<code>project_submitted_datetime</code>	Datetime when project application was submitted. Example: 2016-04-28 12:43:56.245
<code>teacher_id</code>	A unique identifier for the teacher of the proposed project. Example: bdf8baa8fedef6bfeec7ae4ff1c15c56
<code>teacher_prefix</code>	Teacher's title. One of the following enumerated values: <ul style="list-style-type: none"> nan Dr. Mr. Mrs. Ms. Teacher.
<code>teacher_number_of_previously_posted_projects</code>	Number of project applications previously submitted by the same teacher. Example: 2

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

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

Feature	Description
<code>id</code>	A <code>project_id</code> value from the <code>train.csv</code> file. Example: p036502
<code>description</code>	Description of the resource. Example: Tenor Saxophone Reeds, Box of 25
<code>quantity</code>	Quantity of the resource required. Example: 3
<code>price</code>	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
<code>project_is_approved</code>	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"

your neighborhood, and your school are all helpful.

- `__project_essay_2__` "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

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

In [5]:

```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

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

from tqdm import tqdm
import os

from plotly import plotly
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
from collections import Counter
```

1.1 Reading Data

In [6]:

```
project_data = pd.read_csv('train_data.csv')
resource_data = pd.read_csv('resources.csv')
```

In [7]:

```
print("Number of data points in train data", project_data.shape)
print('-'*50)
print("The attributes of data :", project_data.columns.values)
project_data['teacher_prefix'].shape
```

Number of data points in train data (109248, 17)

```
-----
The attributes of data : ['Unnamed: 0' 'id' 'teacher_id' 'teacher_prefix' 'school_state'
 'project_submitted_datetime' 'project_grade_category'
 'project_subject_categories' 'project_subject_subcategories'
 'project_title' 'project_essay_1' 'project_essay_2' 'project_essay_3'
 'project_essay_4' 'project_resource_summary'
 'teacher_number_of_previously_posted_projects' 'project_is_approved']
```

Out[7]:
(109248,)

In [8]:

```
print("Number of data points in train data", resource_data.shape)
print(resource_data.columns.values)
resource_data.head(2)
```

Number of data points in train data (1541272, 4)
['id' 'description' 'quantity' 'price']

Out[8]:

	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

1.2 Data Analysis

In [9]:

```
# PROVIDE CITATIONS TO YOUR CODE IF YOU TAKE IT FROM ANOTHER WEBSITE.
# https://matplotlib.org/gallery/pie_and_polar_charts/pie_and_donut_labels.html#sphx-glr-gallery-pie-and-polar-charts-pie-and-donut-labels-py
```

```
y_value_counts = project_data['project_is_approved'].value_counts()
print("Number of projects thar are approved for funding ", y_value_counts[1], ", (",
      (y_value_counts[1]/(y_value_counts[1]+y_value_counts[0]))*100,"%")
print("Number of projects thar are not approved for funding ", y_value_counts[0], ", (",
      (y_value_counts[0]/(y_value_counts[1]+y_value_counts[0]))*100,"%")
```

```
fig, ax = plt.subplots(figsize=(6, 6), subplot_kw=dict(aspect="equal"))
recipe = ["Accepted", "Not Accepted"]
```

```
data = [y_value_counts[1], y_value_counts[0]]
```

```
wedges, texts = ax.pie(data, wedgeprops=dict(width=0.5), startangle=-40)
```

```
bbox_props = dict(boxstyle="square,pad=0.3", fc="w", ec="k", lw=0.72)
kw = dict(xycoords='data', textcoords='data', arrowprops=dict(arrowstyle="-"),
          bbox=bbox_props, zorder=0, va="center")
```

```
for i, p in enumerate(wedges):
    ang = (p.theta2 - p.theta1)/2. + p.theta1
    y = np.sin(np.deg2rad(ang))
    x = np.cos(np.deg2rad(ang))
    horizontalalignment = {-1: "right", 1: "left"}[int(np.sign(x))]
    connectionstyle = "angle,angleA=0,angleB={}".format(ang)
    kw["arrowprops"].update({"connectionstyle": connectionstyle})
    ax.annotate(recipe[i], xy=(x, y), xytext=(1.35*np.sign(x), 1.4*y),
                horizontalalignment=horizontalalignment, **kw)
```

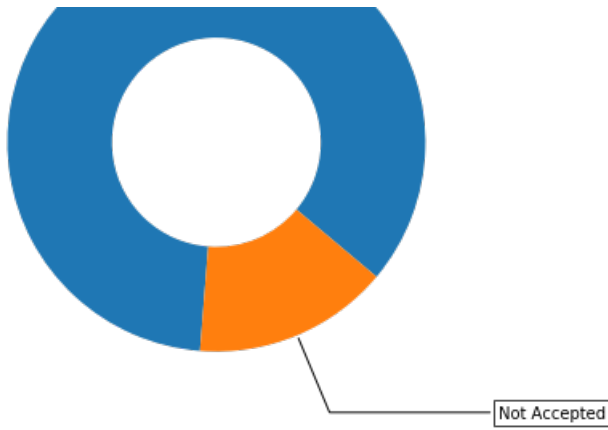
```
ax.set_title("Nmber of projects that are Accepted and not accepted")
```

```
plt.show()
```

Number of projects thar are approved for funding 92706 , (84.85830404217927 %)
Number of projects thar are not approved for funding 16542 , (15.141695957820739 %)

Accepted Number of projects that are Accepted and not accepted





1.2.1 Univariate Analysis: School State

In [10]:

```
# Pandas dataframe groupby count, mean: https://stackoverflow.com/a/19385591/4084039

temp = pd.DataFrame(project_data.groupby("school_state")
["project_is_approved"].apply(np.mean)).reset_index()
# if you have data which contain only 0 and 1, then the mean = percentage (think about it)
temp.columns = ['state_code', 'num_proposals']

# How to plot US state heatmap: https://datascience.stackexchange.com/a/9620

scl = [[0.0, 'rgb(242,240,247)'], [0.2, 'rgb(218,218,235)'], [0.4, 'rgb(188,189,220)'], \
      [0.6, 'rgb(158,154,200)'], [0.8, 'rgb(117,107,177)'], [1.0, 'rgb(84,39,143)']]

data = [ dict(
    type='choropleth',
    colorscale = scl,
    autocolorscale = False,
    locations = temp['state_code'],
    z = temp['num_proposals'].astype(float),
    locationmode = 'USA-states',
    text = temp['state_code'],
    marker = dict(line = dict (color = 'rgb(255,255,255)',width = 2)),
    colorbar = dict(title = "% of pro")
) ]

layout = dict(
    title = 'Project Proposals % of Acceptance Rate by US States',
    geo = dict(
        scope='usa',
        projection=dict( type='albers usa' ),
        showlakes = True,
        lakecolor = 'rgb(255, 255, 255)',
    ),
)

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='us-map-heat-map')
```

In [11]:

```
# https://www.csi.cuny.edu/sites/default/files/pdf/administration/ops/2letterstabbrev.pdf
temp.sort_values(by=['num_proposals'], inplace=True)
print("States with lowest % approvals")
print(temp.head(5))
print('='*50)
print("States with highest % approvals")
print(temp.tail(5))
```

States with lowest % approvals

	state_code	num_proposals
46	VT	0.800000
7	DC	0.802326
43	TX	0.813142
26	MT	0.816327
18	LA	0.831245

States with highest % approvals

	state_code	num_proposals
30	NH	0.873563
35	OH	0.875152
47	WA	0.876178
28	ND	0.888112
8	DE	0.897959

In [12]:

```
#stacked bar plots matplotlib:
https://matplotlib.org/gallery/lines\_bars\_and\_markers/bar\_stacked.html
def stack_plot(data, xtick, col2='project_is_approved', col3='total'):
    ind = np.arange(data.shape[0])

    plt.figure(figsize=(20,5))
    p1 = plt.bar(ind, data[col3].values)
    p2 = plt.bar(ind, data[col2].values)

    plt.ylabel('Projects')
    plt.title('Number of projects approved vs rejected')
    plt.xticks(ind, list(data[xtick].values))
    plt.legend((p1[0], p2[0]), ('total', 'accepted'))
    plt.show()
```

In [13]:

```
def univariate_barplots(data, col1, col2='project_is_approved', top=False):
    # Count number of zeros in dataframe python: https://stackoverflow.com/a/51540521/4084039
    temp = pd.DataFrame(project_data.groupby(col1)[col2].agg(lambda x: x.eq(1).sum())).reset_index()

    # Pandas dataframe grouby count: https://stackoverflow.com/a/19385591/4084039
    temp['total'] = pd.DataFrame(project_data.groupby(col1)
    [col2].agg({'total': 'count'})).reset_index()['total']
    temp['Avg'] = pd.DataFrame(project_data.groupby(col1)[col2].agg({'Avg': 'mean'})).reset_index()['Avg']

    temp.sort_values(by=['total'], inplace=True, ascending=False)

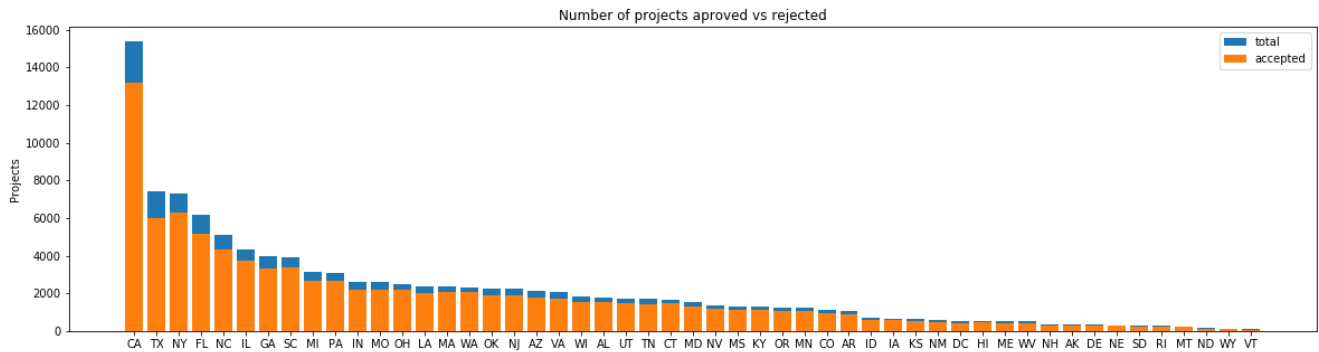
    if top:
```

```
temp = temp[0:top]

stack_plot(temp, xtick=coll, col2=col2, col3='total')
print(temp.head(5))
print("="*50)
print(temp.tail(5))
```

In [14]:

```
univariate_barplots(project_data, 'school_state', 'project_is_approved', False)
```



	school_state	project_is_approved	total	Avg
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

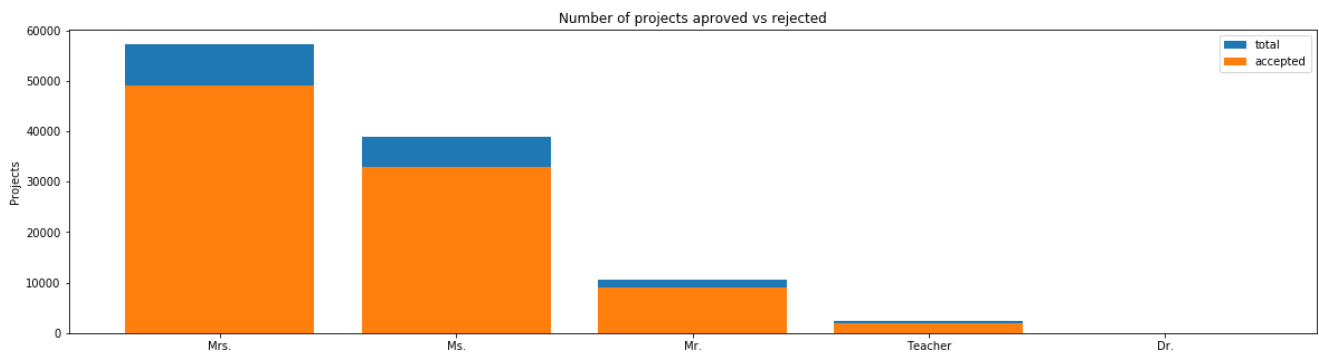
	school_state	project_is_approved	total	Avg
39	RI	243	285	0.852632
26	MT	200	245	0.816327
28	ND	127	143	0.888112
50	WY	82	98	0.836735
46	VT	64	80	0.800000

SUMMARY: Every state has greater than 80% success rate in approval 1) California has highest number of project proposals with 85% of them getting approved

1.2.2 Univariate Analysis: teacher_prefix

In [15]:

```
univariate_barplots(project_data, 'teacher_prefix', 'project_is_approved' , top=False)
```



	teacher_prefix	project_is_approved	total	Avg
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

	teacher_prefix	project_is_approved	total	Avg
2	Mrs.	48997	57269	0.855559

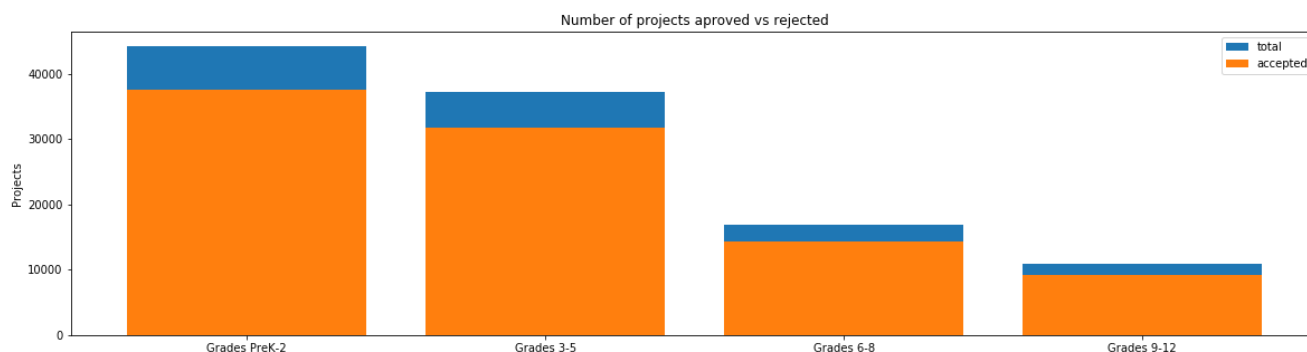
2	Mrs.	40991	31209	0.800000
3	Ms.	32860	38955	0.843537
1	Mr.	8960	10648	0.841473
4	Teacher	1877	2360	0.795339
0	Dr.	9	13	0.692308

Summary: 1) Teachers with prefix 'Mrs' have high number of project proposals with 85% of them being approved. 2) Teachers with doctors degree apply for very few projects compared to others and percentage of acceptance is also low compared to others

1.2.3 Univariate Analysis: project_grade_category

In [16]:

```
univariate_barplots(project_data, 'project_grade_category', 'project_is_approved', top=False)
```



project_grade_category	project_is_approved	total	Avg
Grades PreK-2	37536	44225	0.848751
Grades 3-5	31729	37137	0.854377
Grades 6-8	14258	16923	0.842522
Grades 9-12	9183	10963	0.837636

Summary: 1) Percentage of acceptance for all grades are very closer. 2) Differences in number of projects in grade6-8 and grades3-5 is high compared to other successive differences.

1.2.4 Univariate Analysis: project_subject_categories

In [17]:

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


In [18]:

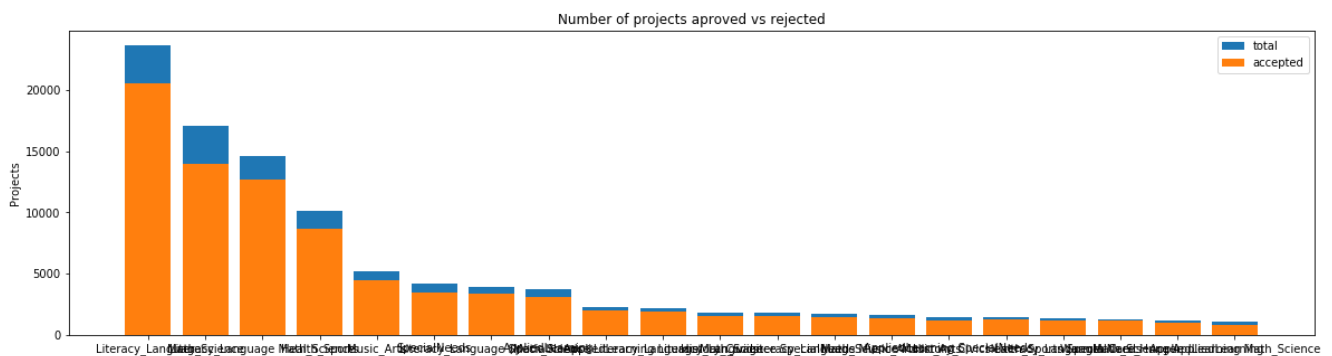
```
project_data['clean_categories'] = cat_list
project_data.drop(['project_subject_categories'], axis=1, inplace=True)
project_data.head(2)
```

Out[18]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	pro
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN	2016-12-05 13:43:57	Gra
1	140945	p258326	897464ce9ddc600bcd1151f324dd63a	Mr.	FL	2016-10-25 09:22:10	Gra

In [19]:

```
univariate_barplots(project_data, 'clean_categories', 'project_is_approved', top=20)
```



```
clean_categories project_is_approved total Avg
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
=====
clean_categories project_is_approved total Avg
19 History_Civics Literacy_Language 1271 1421 0.894441
14 Health_Sports SpecialNeeds 1215 1391 0.873472
50 Warmth Care_Hunger 1212 1309 0.925898
33 Math_Science AppliedLearning 1019 1220 0.835246
4 AppliedLearning Math_Science 855 1052 0.812738
```

Summary: 1)Literacy_Language category has highest number of projects with approx 87% of acceptance. 2)Very critical and sensitive category like warmth_hunger has higher acceptance percentage compared to others.

In [20]:

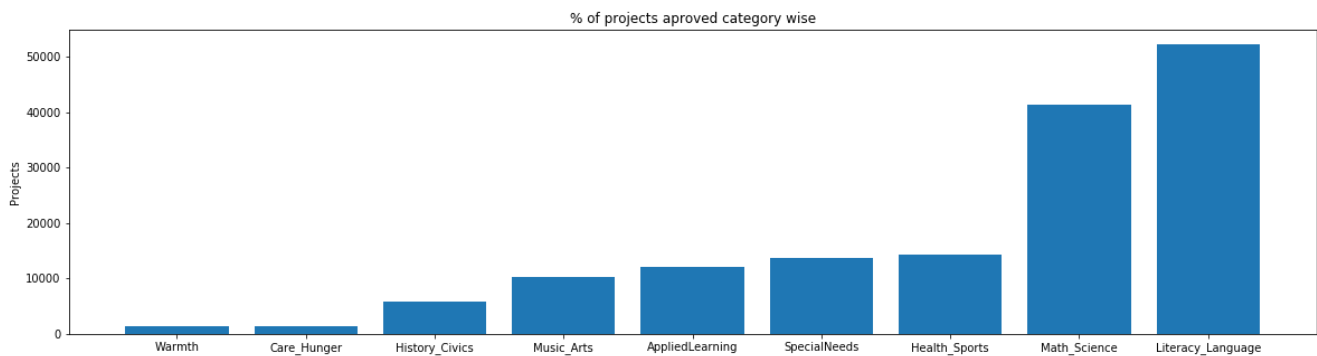
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
from collections import Counter
my_counter = Counter()
for word in project_data['clean_categories'].values:
    my_counter.update(word.split())
```

In [21]:

```
# dict sort by value python: https://stackoverflow.com/a/613218/4084039
cat_dict = dict(my_counter)
sorted_cat_dict = dict(sorted(cat_dict.items(), key=lambda kv: kv[1]))

ind = np.arange(len(sorted_cat_dict))
plt.figure(figsize=(20,5))
p1 = plt.bar(ind, list(sorted_cat_dict.values()))

plt.ylabel('Projects')
plt.title('% of projects aproved category wise')
plt.xticks(ind, list(sorted_cat_dict.keys()))
plt.show()
```



Summary: 1) From the above plot it is clear that plot with two category combined gives us varied data than expected as same category is counted many number of time. But from above plot we get unique subject categories and their acceptance percentage.

In [22]:

```
for i, j in sorted_cat_dict.items():
    print("{:20} :{:10}".format(i,j))
```

```
Warmth           :      1388
Care_Hunger      :      1388
History_Civics   :      5914
Music_Arts       :     10293
AppliedLearning  :     12135
SpecialNeeds     :     13642
Health_Sports    :     14223
Math_Science     :     41421
Literacy_Language :     52239
```

1.2.5 Univariate Analysis: project_subject_subcategories

In [23]:

```
sub_catogories = list(project_data['project_subject_subcategories'].values)
# 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

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

```
sub_cat_list.append(temp.strip())
```

In [24]:

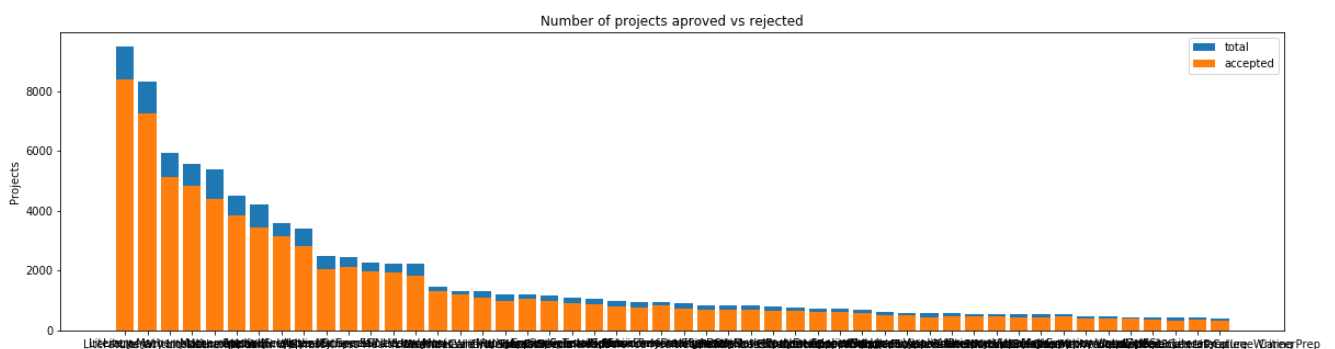
```
project_data['clean_subcategories'] = sub_cat_list
project_data.drop(['project_subject_subcategories'], axis=1, inplace=True)
project_data.head(2)
```

Out[24]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	pro
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN	2016-12-05 13:43:57	Gra
1	140945	p258326	897464ce9ddc600bcd1151f324dd63a	Mr.	FL	2016-10-25 09:22:10	Gra

In [25]:

```
univariate_barplots(project_data, 'clean_subcategories', 'project_is_approved', top=50)
```



```
clean_subcategories  project_is_approved  total  Avg
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
=====
clean_subcategories  project_is_approved  total  Avg
196      EnvironmentalScience Literacy      389      444  0.876126
127      ESL      349      421  0.828979
79      College_CareerPrep      343      421  0.814727
17      AppliedSciences Literature_Writing      361      420  0.859524
3      AppliedSciences College_CareerPrep      330      405  0.814815
```

Summary: 1)Sub Category Literacy has high number of projects 2)AppliedSciences and college related sub categoried have very few total projects.

In [26]:

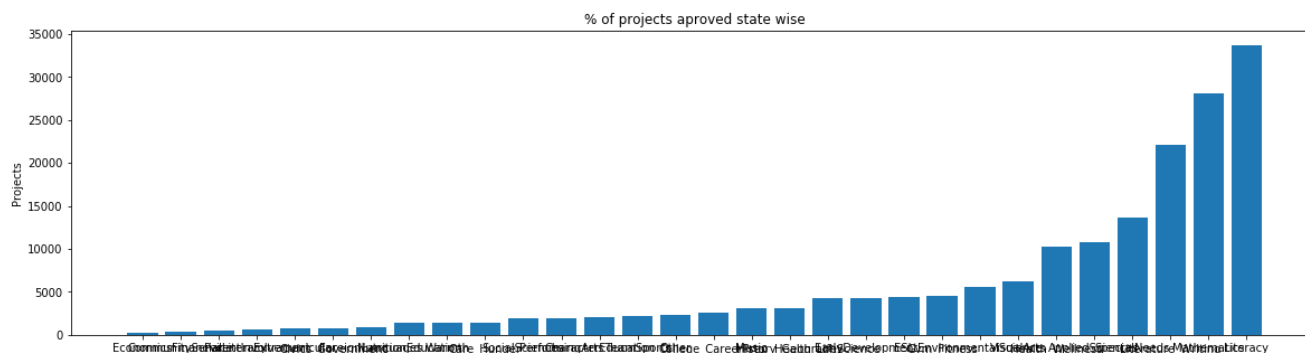
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
from collections import Counter
my_counter = Counter()
for word in project_data['clean_subcategories'].values:
    my_counter.update(word.split())
```

In [27]:

```
# dict sort by value python: https://stackoverflow.com/a/613218/4084039
sub_cat_dict = dict(my_counter)
sorted_sub_cat_dict = dict(sorted(sub_cat_dict.items(), key=lambda kv: kv[1]))

ind = np.arange(len(sorted_sub_cat_dict))
plt.figure(figsize=(20,5))
p1 = plt.bar(ind, list(sorted_sub_cat_dict.values()))

plt.ylabel('Projects')
plt.title('% of projects aproved state wise')
plt.xticks(ind, list(sorted_sub_cat_dict.keys()))
plt.show()
```



In [28]:

```
for i, j in sorted_sub_cat_dict.items():
    print("{:20} {:10}".format(i,j))
```

Economics	:	269
CommunityService	:	441
FinancialLiteracy	:	568
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
AppliedSciences	:	10816
SpecialNeeds	:	13642
Literature_Writing	:	22179
Mathematics	:	28074
Literacy	:	33700

Summary: 1) Still subcategory Literacy has high number of projects but now lowest number of subcategories mentioned is Economic and not Applied Science or College. thus this plot gives accurate results than previous plot as in previous plot two or more subcategories was considered.

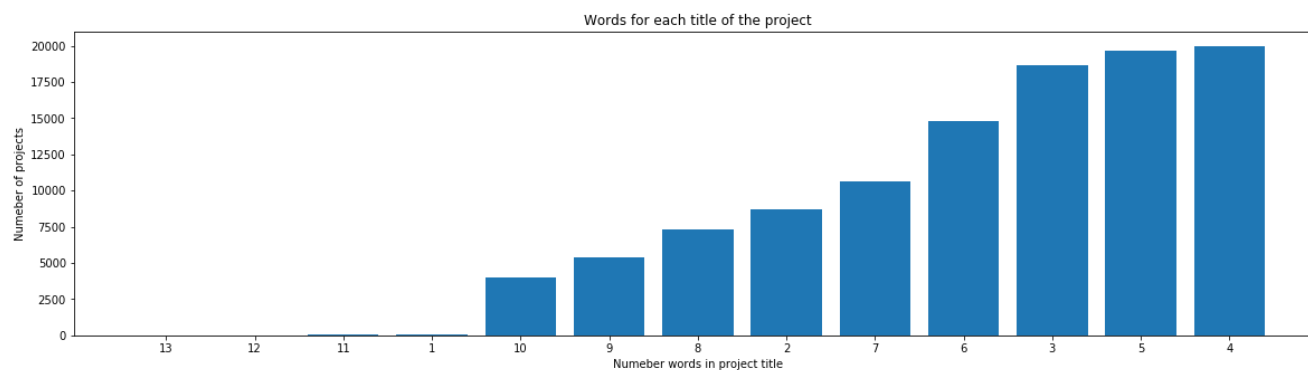
1.2.6 Univariate Analysis: Text features (Title)

In [29]:

```
#How to calculate number of words in a string in DataFrame:
https://stackoverflow.com/a/37483537/4084039
word_count = project_data['project_title'].str.split().apply(len).value_counts()
word_dict = dict(word_count)
word_dict = dict(sorted(word_dict.items(), key=lambda kv: kv[1]))

ind = np.arange(len(word_dict))
plt.figure(figsize=(20,5))
p1 = plt.bar(ind, list(word_dict.values()))

plt.ylabel('Numeber of projects')
plt.xlabel('Numeber words in project title')
plt.title('Words for each title of the project')
plt.xticks(ind, list(word_dict.keys()))
plt.show()
```



Summary: Projects with Project title having 5 and 4 words are high

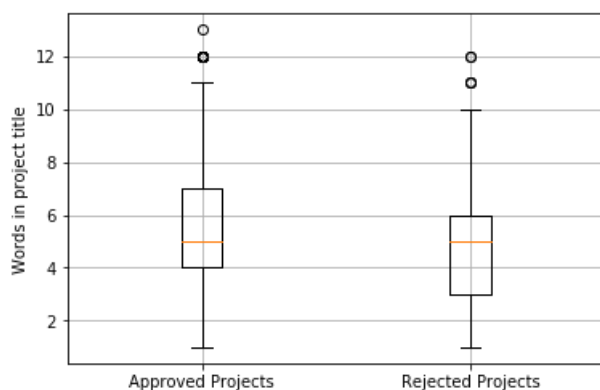
In [30]:

```
approved_title_word_count = project_data[project_data['project_is_approved']==1]['project_title'].str.split().apply(len)
approved_title_word_count = approved_title_word_count.values

rejected_title_word_count = project_data[project_data['project_is_approved']==0]['project_title'].str.split().apply(len)
rejected_title_word_count = rejected_title_word_count.values
```

In [31]:

```
# https://glowingpython.blogspot.com/2012/09/boxplot-with-matplotlib.html
plt.boxplot([approved_title_word_count, rejected_title_word_count])
plt.xticks([1,2], ('Approved Projects', 'Rejected Projects'))
plt.ylabel('Words in project title')
plt.grid()
plt.show()
```

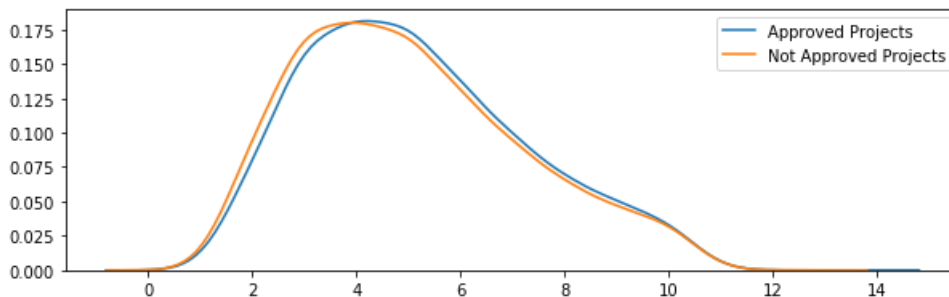


Summary: 1) Mean of both approved and rejected projects are almost same. 2) But 25th and 75th percentile differs and gives clear picture i.e. rejected projects have minimum of 3 word in project title and maximum of 6 the difference between approved and

picture no rejected projects have minimum of 0 word in project title and maximum of same difference between approved and rejected is just 1 word.

In [32]:

```
plt.figure(figsize=(10,3))
sns.kdeplot(approved_title_word_count,label="Approved Projects", bw=0.6)
sns.kdeplot(rejected_title_word_count,label="Not Approved Projects", bw=0.6)
plt.legend()
plt.show()
```



Summary: 1)Both approved and not approved projects are almost same with slightly having higher pdf.

1.2.7 Univariate Analysis: Text features (Project Essay's)

In [33]:

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

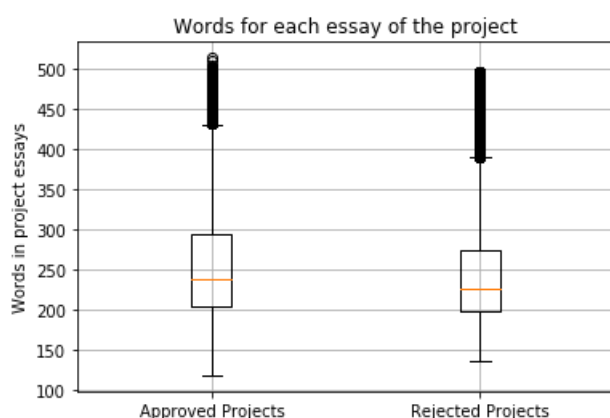
In [34]:

```
approved_word_count = project_data[project_data['project_is_approved']==1]['essay'].str.split().app
ly(len)
approved_word_count = approved_word_count.values

rejected_word_count = project_data[project_data['project_is_approved']==0]['essay'].str.split().app
ly(len)
rejected_word_count = rejected_word_count.values
```

In [35]:

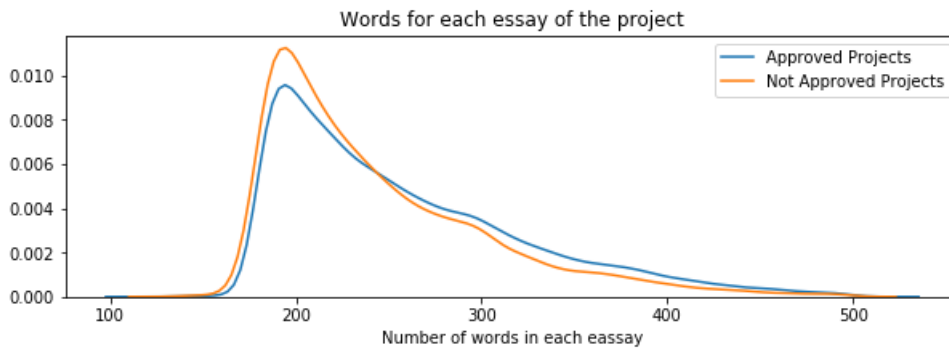
```
# https://glowingpython.blogspot.com/2012/09/boxplot-with-matplotlib.html
plt.boxplot([approved_word_count, rejected_word_count])
plt.title('Words for each essay of the project')
plt.xticks([1,2],('Approved Projects','Rejected Projects'))
plt.ylabel('Words in project essays')
plt.grid()
plt.show()
```



Summary: 1) No major difference in mean between two box plot hence no concrete conclusion can be made from above plot.

In [36]:

```
plt.figure(figsize=(10,3))
sns.distplot(approved_word_count, hist=False, label="Approved Projects")
sns.distplot(rejected_word_count, hist=False, label="Not Approved Projects")
plt.title('Words for each essay of the project')
plt.xlabel('Number of words in each eassay')
plt.legend()
plt.show()
```



Summary: 1) Interestingly, both approved and not approved projects have their highest probability for essays with 200 words.
2) Approved projects with words in the range of 300-400 have a higher probability than not approved projects.

1.2.8 Univariate Analysis: Cost per project

In [37]:

```
# we get the cost of the project using resource.csv file
resource_data.head(2)
```

Out[37]:

	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

In [38]:

```
# https://stackoverflow.com/questions/22407798/how-to-reset-a-dataframes-indexes-for-all-groups-in-one-step
price_data = resource_data.groupby('id').agg({'price': 'sum', 'quantity': 'sum'}).reset_index()
price_data.head(2)
```

Out[38]:

	id	price	quantity
0	p000001	459.56	7
1	p000002	515.89	21

In [39]:

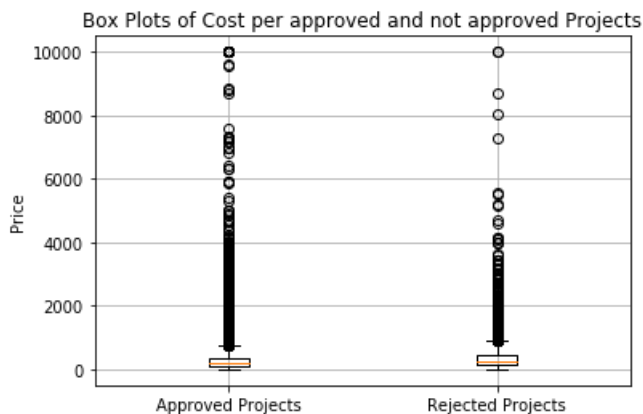
```
# join two dataframes in python:
project_data = pd.merge(project_data, price_data, on='id', how='left')
```

In [40]:

```
approved_price = project_data[project_data['project_is_approved']==1]['price'].values
rejected_price = project_data[project_data['project_is_approved']==0]['price'].values
```

In [41]:

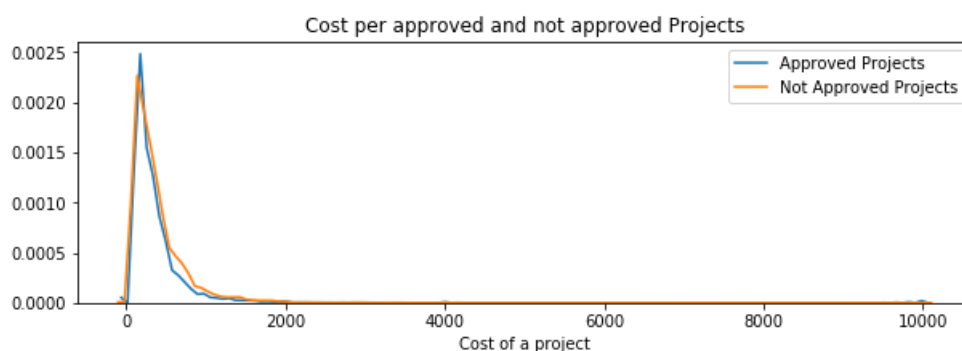
```
# https://glowingpython.blogspot.com/2012/09/boxplot-with-matplotlib.html
plt.boxplot([approved_price, rejected_price])
plt.title('Box Plots of Cost per approved and not approved Projects')
plt.xticks([1,2], ('Approved Projects', 'Rejected Projects'))
plt.ylabel('Price')
plt.grid()
plt.show()
```



Summary: Nothing much can be inferred from above plot other than most of rejected projects have price less than 3000

In [42]:

```
plt.figure(figsize=(10,3))
sns.distplot(approved_price, hist=False, label="Approved Projects")
sns.distplot(rejected_price, hist=False, label="Not Approved Projects")
plt.title('Cost per approved and not approved Projects')
plt.xlabel('Cost of a project')
plt.legend()
plt.show()
```



Summary: 1) Again both approved and not approved projects have their high probability for a value in cost of a project approx. around 250. 2) Probability of approved projects decreases steeply from approx. 750 to 1000 but for not approved it slowly decreases.

In [43]:

```
# http://zetcode.com/python/prettytable/
from prettytable import PrettyTable

#If you get a ModuleNotFoundError error , install prettytable using: pip3 install prettytable

x = PrettyTable()
x.field_names = ["Percentile", "Approved Projects", "Not Approved Projects"]

for i in range(0,101.5):
```



```

        x.add_row([i,np.round(np.percentile(approved_price,i), 3), np.round(np.percentile(rejected_price,i), 3)])
    print(x)

```

Percentile	Approved Projects	Not Approved Projects
0	0.66	1.97
5	13.59	41.9
10	33.88	73.67
15	58.0	99.109
20	77.38	118.56
25	99.95	140.892
30	116.68	162.23
35	137.232	184.014
40	157.0	208.632
45	178.265	235.106
50	198.99	263.145
55	223.99	292.61
60	255.63	325.144
65	285.412	362.39
70	321.225	399.99
75	366.075	449.945
80	411.67	519.282
85	479.0	618.276
90	593.11	739.356
95	801.598	992.486
100	9999.0	9999.0

1.2.9 Univariate Analysis: teacher_number_of_previously_posted_projects

Please do this on your own based on the data analysis that was done in the above cells

In [44]:

```
univariate_barplots(project_data, 'teacher_number_of_previously_posted_projects',
'project is approved', top=50)
```



	teacher_number_of_previously_posted_projects	project_is_approved	total	\
0	0	24652	30014	
1	1	13329	16058	
2	2	8705	10350	
3	3	5997	7110	
4	4	4452	5266	

	Avg
0	0.821350
1	0.830054
2	0.841063
3	0.843460
4	0.845423

	teacher_number_of_previously_posted_projects	project_is_approved	total	\
46	46	149	164	
45	45	141	153	
47	47	129	144	
49	49	128	143	

```

Avg
46 0.908537
45 0.921569
47 0.895833
49 0.895105
48 0.964286

```

Summary: 1) Teachers who are posting for first time i.e with 0 past project proposals have higher projects near to 30000. 2) The percentage of those getting accepted is 82%

1.2.10 Univariate Analysis: project_resource_summary

Please do this on your own based on the data analysis that was done in the above cells

Check if the presence of the numerical digits in the `project_resource_summary` effects the acceptance of the project or not. If you observe that presence of the numerical digits is helpful in the classification, please include it for further process or you can ignore it.

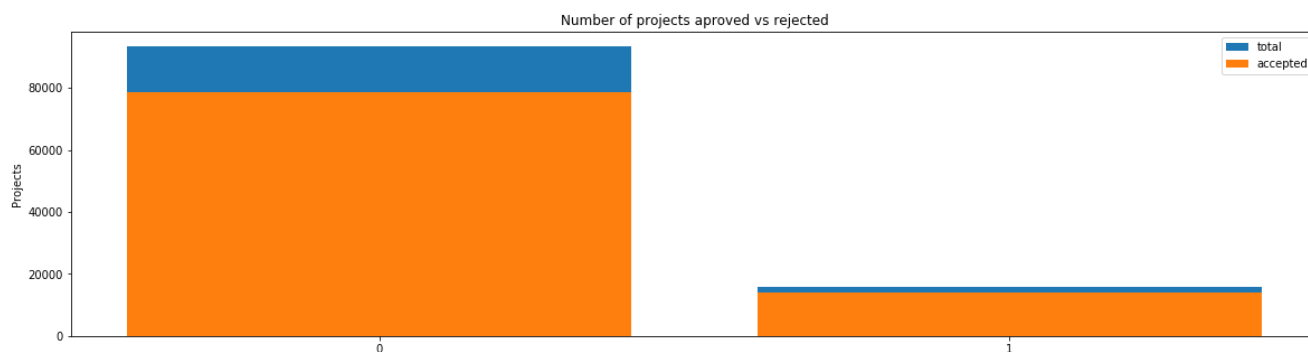
In [45]:

```

presencenumeric=[]
for i in project_data['project_resource_summary']:
    if bool(re.search(r'\d', i))==True:
        presencenumeric.append(1)
    else:
        presencenumeric.append(0)
project_data['presence_numeric']=presencenumeric

univariate_barplots(project_data, 'presence_numeric', 'project_is_approved', top=50)

```



presence_numeric	project_is_approved	total	Avg	
0	0	78616	93492	0.840885
1	1	14090	15756	0.894263

presence_numeric	project_is_approved	total	Avg	
0	0	78616	93492	0.840885
1	1	14090	15756	0.894263

Summary: 1) Resource Summary without integer has high projects and accepted projects percentage is 84% 2) very few projects containing numbers in resource summary with acceptance percentage of 89%

1.3 Text preprocessing

1.3.1 Essay Text

In [46]:

```
project_data.head(2)
```

Out[46]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	proj
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN	2016-12-05 13:43:57	Gra
1	140945	p258326	897464ce9ddc600bcd1151f324dd63a	Mr.	FL	2016-10-25 09:22:10	Gra

2 rows \times 21 columns

In [47]:

```
# printing some random essays.
print(project_data['essay'].values[0])
print("="*50)
print(project_data['essay'].values[150])
print("="*50)
print(project_data['essay'].values[1000])
print("="*50)
print(project_data['essay'].values[20000])
print("="*50)
print(project_data['essay'].values[99999])
print("="*50)
```

My students are English learners that are working on English as their second or third languages. We are a melting pot of refugees, immigrants, and native-born Americans bringing the gift of language to our school. \r\n\r\n We have over 24 languages represented in our English Learner program with students at every level of mastery. We also have over 40 countries represented with the families within our school. Each student brings a wealth of knowledge and experiences to us that open our eyes to new cultures, beliefs, and respect.\"The limits of your language are the limits of your world.\"-Ludwig Wittgenstein Our English learner's have a strong support system at home that begs for more resources. Many times our parents are learning to read and speak English alongside of their children. Sometimes this creates barriers for parents to be able to help their child learn phonetics, letter recognition, and other reading skills.\r\n\r\nBy providing these dvd's and players, students are able to continue their mastery of the English language even if no one at home is able to assist. All families with students within the Level 1 proficiency status, will be offered to be a part of this program. These educational videos will be specially chosen by the English Learner Teacher and will be sent home regularly to watch. The 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 opportunity to check out a dvd player to use for the year. The plan is to use these videos and educational dvd's for the years to come for other EL students.\r\n\r\nannan

The 51 fifth grade students that will cycle through my classroom this year all love learning, at least most of the time. At our school, 97.3% of the students receive free or reduced price lunch. Of the 560 students, 97.3% are minority students. The school has a vibrant community that loves to get together and celebrate. Around Halloween there is a whole school parade to show off the beautiful costumes that students wear. On Cinco de Mayo we put on a big festival with crafts made by the students, dances, and games. At the end of the year the school hosts a carnival to celebrate the hard work put in during the school year, with a dunk tank being the most popular activity. My students will use these five brightly colored Hokki stools in place of regular, stationary, 4-legged chairs. As I will only have a total of ten in the classroom and not enough for each student to have an individual one, they will be used in a variety of ways. During independent reading time they will be used as special chairs students will each use on occasion. I will utilize them in place of chairs at my small group tables during math and reading times. The rest of the day they will be used by the students who need the highest amount of movement in their life in order to stay focused on school. Whenever asked what the classroom is missing, my students always say more Hokki Stools. They can't get their fill of the 5 stools we already have. When the students are sitting in a group with me on the Hokki Stools, they are always moving, but at the same time doing their work. Anytime the students get to pick where they can sit, the Hokki Stools are the first to be taken.

ken. There are always students who head over to the kidney table to get one of the stools who are disappointed as there are not enough of them. \r\n\r\nWe ask a lot of students to sit for 7 hours a day. The Hokki stools will be a compromise that allow my students to do desk work and move at the same time. These stools will help students to meet their 60 minutes a day of movement by allowing them to activate their core muscles for balance while they sit. For many of my students, these chairs will take away the barrier that exists in schools for a child who can't sit still.nannan

=====

How do you remember your days of school? Was it in a sterile environment with plain walls, rows of desks, and a teacher in front of the room? A typical day in our room is nothing like that. I work hard to create a warm inviting themed room for my students look forward to coming to each day.\r\n\r\nMy class is made up of 28 wonderfully unique boys and girls of mixed races in Arkansas.\r\n\r\nThey attend a Title I school, which means there is a high enough percentage of free and reduced-price lunch to qualify. Our school is an "open classroom" concept, which is very unique as there are no walls separating the classrooms. These 9 and 10 year-old students are very eager learners; they are like sponges, absorbing all the information and experiences and keep on wanting more. With these resources such as the comfy red throw pillows and the whimsical nautical hanging decor and the blue fish nets, I will be able to help create the mood in our classroom setting to be one of a themed nautical environment. Creating a classroom environment is very important in the success in each and every child's education. The nautical photo props will be used with each child as they step foot into our classroom for the first time on Meet the Teacher evening. I'll take pictures of each child with them, have them developed, and then hung in our classroom ready for their first day of 4th grade. This kind gesture will set the tone before even the first day of school! The nautical thank you cards will be used throughout the year by the students as they create thank you cards to their team groups.\r\n\r\n\r\nYour generous donations will help me to help make our classroom a fun, inviting, learning environment from day one.\r\n\r\n\r\nIt costs a lot of money out of my own pocket on resources to get our classroom ready. Please consider helping with this project to make our new school year a very successful one. Thank you!nannan

=====

My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work their hardest working past their limitations. \r\n\r\n\r\nThe materials we have are the ones I seek out for my students. I teach in a Title I school where most of the students receive free or reduced price lunch. Despite their disabilities and limitations, my students love coming to school and come eager to learn and explore. Have you ever felt like you had ants in your pants and you needed to groove and move as you were in a meeting? This is how my kids feel all the time. They want to be able to move as they learn or so they say. Wobble chairs are the answer and I love them because they develop their core, which enhances gross motor and fine motor skills. \r\n\r\nThey also want to learn through games, my kids don't want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is the key to our success. The number toss and color and shape mats can make that happen. My students will forget they are doing work and just have the fun a 6 year old deserves.nannan

=====

The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires. -William A. Ward\r\n\r\n\r\nMy school has 803 students which is made up of 97.6% African-American, making up the largest segment of the student body. A typical school in Dallas is made up of 23.2% African-American students. Most of the students are on free or reduced lunch. We aren't receiving doctors, lawyers, or engineers children from rich backgrounds or neighborhoods. As an educator I am inspiring minds of young children and we focus not only on academics but on smart, effective, efficient, and disciplined students with good character. In our classroom we can utilize the Bluetooth for swift transitions during class. I use a speaker which doesn't amplify the sound enough to receive the message. Due to the volume of my speaker my students can't hear videos or books clearly and it isn't making the lessons as meaningful. But with the bluetooth speaker my students will be able to hear and I can stop, pause and replay it at any time.\r\n\r\nThe cart will allow me to have more room for storage of things that are needed for the day and has an extra part to it I can use. The table top chart has all of the letter, words and pictures for students to learn about different letters and it is more accessible.nannan

=====

In [48]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

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

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

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

In [49]:

```
sent = decontracted(project_data['essay'].values[20000])
print(sent)
print(" "*50)
```

My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work their hardest working past their limitations. \r\n\r\nThe materials we have are the ones I seek out for my students. I teach in a Title I school where most of the students receive free or reduced price lunch. Despite their disabilities and limitations, my students love coming to school and come eager to learn and explore. Have you ever felt like you had ants in your pants and you needed to groove and move as you were in a meeting? This is how my kids feel all the time. They want to be able to move as they learn or so they say. Wobble chairs are the answer and I love them because they develop their core, which enhances gross motor and in turn fine motor skills. \r\n\r\nThey also want to learn through games, my kids do not want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is the key to our success. The number toss and color and shape mats can make that happen. My students will forget they are doing work and just have the fun a 6 year old deserves. nannan

In [50]:

```
# \r \n \t remove from string python: http://texthandler.com/info/remove-line-breaks-python/
sent = sent.replace('\r', ' ')
sent = sent.replace('\n', ' ')
sent = sent.replace('\t', ' ')
print(sent)
```

My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work their hardest working past their limitations. The materials we have are the ones I seek out for my students. I teach in a Title I school where most of the students receive free or reduced price lunch. Despite their disabilities and limitations, my students love coming to school and come eager to learn and explore. Have you ever felt like you had ants in your pants and you needed to groove and move as you were in a meeting? This is how my kids feel all the time. They want to be able to move as they learn or so they say. Wobble chairs are the answer and I love them because they develop their core, which enhances gross motor and in turn fine motor skills. They also want to learn through games, my kids do not want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is the key to our success. The number toss and color and shape mats can make that happen. My students will forget they are doing work and just have the fun a 6 year old deserves. nannan

In [51]:

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

My kindergarten students have varied disabilities ranging from speech and language delays cognitive delays gross fine motor delays to autism They are eager beavers and always strive to work their hardest working past their limitations The materials we have are the ones I seek out for my students I teach in a Title I school where most of the students receive free or reduced price lunch Despite their disabilities and limitations my students love coming to school and come eager to learn and explore Have you ever felt like you had ants in your pants and you needed to groove and move as you were in a meeting This is how my kids feel all the time They want to be able to move as they learn or so they say Wobble chairs are the answer and I love them because they develop their core which enhances gross motor and in turn fine motor skills They also want to learn through games my kids do not want to sit and do worksheets They want to learn to count by jumping and playing Physical engagement is the key to our success The number toss and color and shape mats can make that happen My students will forget they are doing work and just have the fun a 6 year old deserves nannan

In [52]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords= ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
```

◀ ▶

```
100% |██████████████████████████████████████████████████████████████████████████| 109248/109248  
[01:21<00:00, 1333.24it/s]
```

1.3.2 Project title Text

```
100%|██████████████████████████████████████████████████████████████████████████| 109248/109248  
[00:04<00:00, 26395.84it/s]
```

```
preprocessed_title[20000]
```

```
'we need to move it while we input it'
```

```
project data.columns
```

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
      'project_submitted_datetime', 'project_grade_category', 'project_title',
      'project_essay_1', 'project_essay_2', 'project_essay_3',
      'project_essay_4', 'project_resource_summary',
      'teacher_number_of_previously_posted_projects', 'project_is_approved',
      'clean_categories', 'clean_subcategories', 'essay', 'price', 'quantity',
      'presence_numeric'],
      dtype='object')
```

- ```
- school_state : categorical data
- clean_categories : categorical data
- clean_subcategories : categorical data
- project_grade_category : categorical data
- teacher_prefix : categorical data

- project_title : text data
- text : text data
- project_resource_summary: text data

- quantity : numerical
- teacher_number_of_previously_posted_projects : numerical
- price : numerical
```

```
we use count vectorizer to convert the values into one hot encoded features
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(project_data['clean_categories'].values)
print(vectorizer.get_feature_names())
```

```
print(vectorizer.get_feature_names())
```

```
categories_one_hot = vectorizer.transform(project_data['clean_categories'].values)
print("Shape of matrix after one hot encodig ",categories_one_hot.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig (109248, 9)
```

In [59]:

```
we use count vectorizer to convert the values into one hot encoded features
vectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(project_data['clean_subcategories'].values)
print(vectorizer.get_feature_names())
```

```
sub_categories_one_hot = vectorizer.transform(project_data['clean_subcategories'].values)
print("Shape of matrix after one hot encodig ",sub_categories_one_hot.shape)
```

```
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular',
'Civics_Government', 'ForeignLanguages', 'NutritionEducation', 'Warmth', 'Care_Hunger',
'SocialSciences', 'PerformingArts', 'CharacterEducation', 'TeamSports', 'Other',
'College_CareerPrep', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopment', 'ESL',
'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedSciences',
'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Literacy']
Shape of matrix after one hot encodig (109248, 30)
```

In [60]:

```
Please do the similar feature encoding with state, teacher_prefix and project_grade_category also
vectorizer = CountVectorizer()
vectorizer.fit(project_data['school_state'].values)
print(vectorizer.get_feature_names())
```

```
school_state_one_hot = vectorizer.transform(project_data['school_state'].values)
print("Shape of matrix after one hot encodig ",school_state_one_hot.shape)
```

```
['ak', 'al', 'ar', 'az', 'ca', 'co', 'ct', 'dc', 'de', 'fl', 'ga', 'hi', 'ia', 'id', 'il', 'in', 'ks',
's', 'ky', 'la', 'ma', 'md', 'me', 'mi', 'mn', 'mo', 'ms', 'mt', 'nc', 'nd', 'ne', 'nh', 'nj', 'nm',
'nv', 'ny', 'oh', 'ok', 'or', 'pa', 'ri', 'sc', 'sd', 'tn', 'tx', 'ut', 'va', 'vt', 'wa', 'wi', 'wv',
'wy']
Shape of matrix after one hot encodig (109248, 51)
```

In [61]:

```
#project_data['teacher_prefix'].fillna("NotDefined",inplace=False)
project_data['teacher_prefix'].shape
```

Out[61]:

```
(109248,)
```

Points Observed: In doing one hot encoding for teacher prefix the following error was observed. ValueError: np.nan is an invalid document, expected byte or unicode string So first I tried to remove missing values but number of rows decreased by 3 and in further when trying to combine all one hot encoding I was getting another error because number of rows were unequal. Then through this link found how to fix <https://stackoverflow.com/questions/39303912/tfidfvectorizer-in-scikit-learn-valueerror-np-nan-is-an-invalid-document/39308809#39308809>

In [62]:

```
vectorizer = CountVectorizer()
vectorizer.fit(project_data['teacher_prefix'].values.astype("U"))
print(vectorizer.get_feature_names())
teach_state_one_hot = vectorizer.transform(project_data['teacher_prefix'].values.astype("U"))
print("Shape of matrix after one hot encodig ",teach_state_one_hot.shape)
```



```
['dr', 'mr', 'mrs', 'ms', 'nan', 'teacher']
Shape of matrix after one hot encoding (109248, 6)
```

In [63]:

```
vectorizer = CountVectorizer()
vectorizer.fit(project_data['project_grade_category'].values)
print(vectorizer.get_feature_names())
project_grade_one_hot = vectorizer.transform(project_data['project_grade_category'].values)
print("Shape of matrix after one hot encoding ", project_grade_one_hot.shape)
```

```
['12', 'grades', 'prek']
Shape of matrix after one hot encoding (109248, 3)
```

## 1.4.2 Vectorizing Text data

### 1.4.2.1 Bag of words

In [64]:

```
We are considering only the words which appeared in at least 10 documents (rows or projects).
vectorizer = CountVectorizer(min_df=10)
text_bow = vectorizer.fit_transform(preprocessed_essays)
print("Shape of matrix after one hot encoding ", text_bow.shape)
```

Shape of matrix after one hot encoding (109248, 16623)

#### 1.4.2.2 Bag of Words on 'project\_title'

In [65]:

```
you can vectorize the title also
before you vectorize the title make sure you preprocess it
vectorizer = CountVectorizer(min_df=10)
title_bow = vectorizer.fit_transform(preprocessed_title)
print("Shape of matrix after one hot encoding ", title_bow.shape)
```

Shape of matrix after one hot encoding (109248, 3329)

### 1.4.2.3 TFIDF vectorizer

In [66]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)
text_tfidf = vectorizer.fit_transform(preprocessed_essays)
print("Shape of matrix after one hot encoding ", text_tfidf.shape)
```

Shape of matrix after one hot encoding (109248, 16623)

#### 1.4.2.4 TFIDF Vectorizer on 'project\_title'

In [67]:

```
Similarly you can vectorize for title also
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)
title_tfidf = vectorizer.fit_transform(preprocessed_title)
print("Shape of matrix after one hot encoding ", title_tfidf.shape)
```

Shape of matrix after one hot encoding (109248, 3329)

### 1.4.2.5 Using Pretrained Models: Avg W2V

In [68]:

```
'''
Reading glove vectors in python: https://stackoverflow.com/a/38230349/4084039
def loadGloveModel(gloveFile):
 print ("Loading Glove Model")
 f = open(gloveFile,'r', encoding="utf8")
 model = {}
 for line in tqdm(f):
 splitLine = line.split()
 word = splitLine[0]
 embedding = np.array([float(val) for val in splitLine[1:]])
 model[word] = embedding
 print ("Done.",len(model)," words loaded!")
 return model
model = loadGloveModel('glove.42B.300d.txt')

=====
Output:

Loading Glove Model
1917495it [06:32, 4879.69it/s]
Done. 1917495 words loaded!

=====

words = []
for i in preproced_texts:
 words.extend(i.split(' '))

for i in preproced_titles:
 words.extend(i.split(' '))
print("all the words in the coupus", len(words))
words = set(words)
print("the unique words in the coupus", len(words))

inter_words = set(model.keys()).intersection(words)
print("The number of words that are present in both glove vectors and our coupus", \
 len(inter_words), "("",np.round(len(inter_words)/len(words)*100,3),"%")

words_courpus = {}
words_glove = set(model.keys())
for i in words:
 if i in words_glove:
 words_courpus[i] = model[i]
print("word 2 vec length", len(words_courpus))

stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-sa
ve-and-load-variables-in-python/

import pickle
with open('glove_vectors', 'wb') as f:
 pickle.dump(words_courpus, f)

'''
```

Out[68]:

```
'\n# Reading glove vectors in python: https://stackoverflow.com/a/38230349/4084039\ndef
loadGloveModel(gloveFile):\n print ("Loading Glove Model")\n f = open(gloveFile,\r',
encoding="utf8")\n model = {}\n for line in tqdm(f):\n splitLine = line.split()\n
word = splitLine[0]\n embedding = np.array([float(val) for val in splitLine[1:]])\n
odel[word] = embedding\n print ("Done.",len(model)," words loaded!")\n return model\nmodel =
loadGloveModel(\glove.42B.300d.txt)\n\n# =====\nOutput:\n \nLoading G
love Model\n1917495it [06:32, 4879.69it/s]\nDone. 1917495 words loaded!\n\n#
=====\n\nwords = []\nfor i in preproced_texts:\n words.extend(i.split(\
'))\n\nfor i in preproced_titles:\n words.extend(i.split(\
'))\nprint("all the words in the
coupus", len(words))\nwords = set(words)\nprint("the unique words in the coupus",
len(words))\n\ninter_words = set(model.keys()).intersection(words)\nprint("The number of words tha
t are present in both glove vectors and our coupus",
len(inter_words),
("np.round(len(inter words)/len(words)*100.3),"%")\n\nwords_courpus = {}\nwords_glove =
```

◀ ▶

In [69]:

```
stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-python/
make sure you have the glove_vectors file
with open('glove_vectors', 'rb') as f:
 model = pickle.load(f)
 glove_words = set(model.keys())
```

In [70]:

```
average Word2Vec
compute average word2vec for each review.
avg_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(preprocessed_essays): # for each review/sentence
 vector = np.zeros(300) # as word vectors are of zero length
 cnt_words = 0; # num of words with a valid vector in the sentence/review
 for word in sentence.split(): # for each word in a review/sentence
 if word in glove_words:
 vector += model[word]
 cnt_words += 1
 if cnt_words != 0:
 vector /= cnt_words
 avg_w2v_vectors.append(vector)

print(len(avg_w2v_vectors))
print(len(avg_w2v_vectors[5]))
```

```
100%|██| 109248/109248
[00:47<00:00, 2296.56it/s]
```

109248  
300

#### 1.4.2.6 Using Pretrained Models: AVG W2V on `project\_title`

In [71]:

```
Similarly you can vectorize for title also
avg_w2v_vectors_title= []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(preprocessed_title): # for each review/sentence
 vector = np.zeros(300) # as word vectors are of zero length
 cnt_words = 0; # num of words with a valid vector in the sentence/review
 for word in sentence.split(): # for each word in a review/sentence
 if word in glove_words:
 vector += model[word]
 cnt_words += 1
 if cnt_words != 0:
 vector /= cnt_words
 avg_w2v_vectors_title.append(vector)

print(len(avg_w2v_vectors_title))
print(len(avg_w2v_vectors_title[5]))
```

```
100%|██| 109248/109248
[00:02<00:00, 52148.52it/s]
```

109248  
300

#### 1.4.2.7 Using Pretrained Models: TFIDF weighted W2V

In [72]:



## 14.3 Vectorizing Numerical Features

In [75]:

```
check this one: https://www.youtube.com/watch?v=0HOqOcIn3Z4&t=530s
standardization sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
from sklearn.preprocessing import StandardScaler

price_standardized = standardScaler.fit(project_data['price'].values)
this will rise the error
ValueError: Expected 2D array, got 1D array instead: array=[725.05 213.03 329. ... 399. 287.
73 5.5].
Reshape your data either using array.reshape(-1, 1)

price_scalar = StandardScaler()
price_scalar.fit(project_data['price'].values.reshape(-1,1)) # finding the mean and standard
deviation of this data
print(f"Mean : {price_scalar.mean_[0]}, Standard deviation : {np.sqrt(price_scalar.var_[0])}")

Now standardize the data with above maen and variance.
price_standardized = price_scalar.transform(project_data['price'].values.reshape(-1, 1))
```

Mean : 298.1193425966608, Standard deviation : 367.49634838483496

In [76]:

```
price_standardized
```

Out[76]:

```
array([[-0.3905327],
 [0.00239637],
 [0.59519138],
 ...,
 [-0.15825829],
 [-0.61243967],
 [-0.51216657]])
```

In [77]:

```
from sklearn.preprocessing import StandardScaler

price_standardized =
standardScaler.fit(project_data['teacher_number_of_previously_posted_projects'].values)
this will rise the error
ValueError: Expected 2D array, got 1D array instead: array=[725.05 213.03 329. ... 399. 287.
73 5.5].
Reshape your data either using array.reshape(-1, 1)

previous_scalar = StandardScaler()
previous_scalar.fit(project_data['teacher_number_of_previously_posted_projects'].values.reshape(-1,1)) # finding the mean and standard deviation of this data
print(f"Mean : {previous_scalar.mean_[0]}, Standard deviation :
{np.sqrt(previous_scalar.var_[0])}")

Now standardize the data with above maen and variance.
previous_standardized =
previous_scalar.transform(project_data['teacher_number_of_previously_posted_projects'].values.reshape(-1, 1))
```

Mean : 11.153165275336848, Standard deviation : 27.77702641477403

In [78]:

```
previous_standardized
```

Out[78]:

```
array([[-0.40152481],
 [-0.14951799],
 [-0.36552384],
 ...,
 ...])
```

```
[-0.29352189],
[-0.40152481],
[-0.40152481]])
```

### 1.4.4 Merging all the above features

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

In [79]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(text_bow.shape)
print(price_standardized.shape)
print(teach_state_one_hot.shape)
#print(.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 16623)
(109248, 1)
(109248, 6)
```

In [80]:

```
merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
with the same hstack function we are concatenating a sparse matrix and a dense matrix :)
X = hstack((categories_one_hot, sub_categories_one_hot, teach_state_one_hot, school_state_one_hot, price_standardized, previous_standardized))
print(X.shape)
```

```
(109248, 98)
```

## Assignment 2: Apply TSNE

If you are using any code snippet from the internet, you have to provide the reference/citations, as we did in the above cells. Otherwise, it will be treated as plagiarism without citations.

1. In the above cells we have plotted and analyzed many features. Please observe the plots and write the observations in markdown cells below every plot.
2. EDA: Please complete the analysis of the feature: teacher\_number\_of\_previously\_posted\_projects
3. Build the data matrix using these features
  - school\_state : categorical data (one hot encoding)
  - clean\_categories : categorical data (one hot encoding)
  - clean\_subcategories : categorical data (one hot encoding)
  - teacher\_prefix : categorical data (one hot encoding)
  - project\_grade\_category : categorical data (one hot encoding)
  - project\_title : text data (BOW, TFIDF, AVG W2V, TFIDF W2V)
  - price : numerical
  - teacher\_number\_of\_previously\_posted\_projects : numerical
4. Now, plot FOUR t-SNE plots with each of these feature sets.
  - A. categorical, numerical features + project\_title(BOW)
  - B. categorical, numerical features + project\_title(TFIDF)
  - C. categorical, numerical features + project\_title(AVG W2V)
  - D. categorical, numerical features + project\_title(TFIDF W2V)
5. Concatenate all the features and Apply TSNE on the final data matrix
6. **Note 1: The TSNE accepts only dense matrices**
7. **Note 2: Consider only 5k to 6k data points to avoid memory issues. If you run into memory error issues, reduce the number of data points but clearly state the number of data-points you are using**

In [81]:

```
from scipy.sparse import hstack
```

```
with the same hstack function we are concatenating a sparse matrix and a dense matrix :)
X = hstack((categories_one_hot,
sub_categories_one_hot,teach_state_one_hot,school_state_one_hot,project_grade_one_hot,previous_star
dardized , price_standardized,title_bow))
print(X.shape)
```

(109248, 3430)

## 2.1 TSNE with `BOW` encoding of `project\_title` feature

In [82]:

```
from sklearn.manifold import TSNE
X = X.tocsr()
X_new = X[0:5000,:]
```

In [83]:

```
X_new.shape
```

Out[83]:

```
(5000, 3430)
```

In [84]:

```
type(X_new)
```

Out[84]:

```
scipy.sparse.csr.csr_matrix
```

In [85]:

```
X_new=X_new.toarray()
model=TSNE(n_components=2,perplexity=100.0,random_state=0)
tsne_data_b=model.fit_transform(X_new)
```

In [86]:

```
labels=project_data["project_is_approved"]
labels_new = labels[0: 5000]
len(labels_new)
```

Out[86]:

```
5000
```

In [87]:

```
labels_new
```

Out[87]:

```
0 0
1 1
2 0
3 1
4 1
5 1
6 1
7 1
8 1
9 1
10 1
11 1
12 0
13 1
...
```

```
14 0
15 1
16 1
17 1
18 1
19 1
20 1
21 1
22 0
23 1
24 1
25 0
26 1
27 1
28 0
29 1
..
4970 0
4971 0
4972 1
4973 1
4974 1
4975 1
4976 0
4977 1
4978 0
4979 1
4980 1
4981 1
4982 1
4983 1
4984 1
4985 1
4986 1
4987 1
4988 1
4989 1
4990 1
4991 1
4992 1
4993 1
4994 1
4995 1
4996 1
4997 0
4998 1
4999 1
Name: project_is_approved, Length: 5000, dtype: int64
```

In [88]:

```
tsne_data_b = np.vstack((tsne_data_b.T, labels_new)).T
tsne_df_b = pd.DataFrame(tsne_data_b, columns = ("1st_Dim", "2nd_Dim", "Labels"))
```

In [89]:

```
tsne_df_b.shape
```

Out[89]:

```
(5000, 3)
```

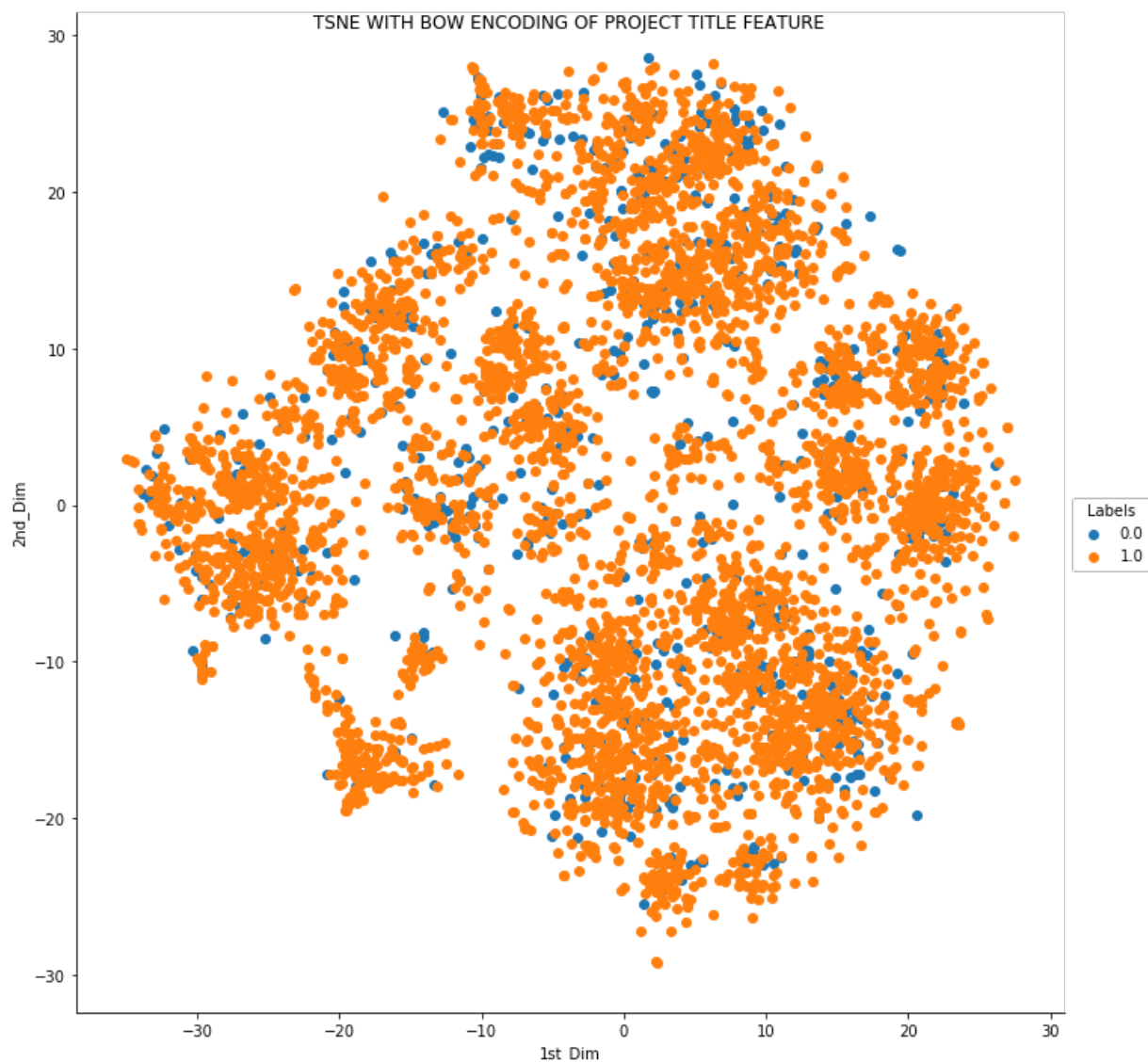
In [90]:

```
please write all of the code with proper documentation and proper titles for each subsection
when you plot any graph make sure you use
 # a. Title, that describes your plot, this will be very helpful to the reader
 # b. Legends if needed
 # c. X-axis label
 # d. Y-axis label
import numpy as np
from sklearn.manifold import TSNE
from sklearn import datasets
import pandas as pd
```



```
import matplotlib.pyplot as plt
```

```
sns.FacetGrid(tsne_df_b, hue = "Labels", size = 10).map(plt.scatter, "1st_Dim", "2nd_Dim").add_legend().fig.suptitle("TSNE WITH BOW ENCODING OF PROJECT TITLE FEATURE ")
plt.show()
```



Summary: 1.From the above plot all points are clustered and hence we cannot use tsne with BOW encoding to categorize projects approved

## 2.2 TSNE with `TFIDF` encoding of `project\_title` feature

In [95]:

```
X = hstack((categories_one_hot,
sub_categories_one_hot,teach_state_one_hot,school_state_one_hot,project_grade_one_hot,previous_star
dardized , price_standardized,title_tfidf))
print(X.shape)
```

(109248, 3430)

In [100]:

```
from sklearn.manifold import TSNE
X = X.tocsr()
X_new = X[0:5000,:]
```

```
In [101]:
```

```
X_new = X_new.toarray()
model = TSNE(n_components = 2, perplexity = 100.0, random_state = 0)
tsne_data_tfidf = model.fit_transform(X_new)
tsne_data_tfidf = np.vstack((tsne_data_tfidf.T, labels_new)).T
tsne_df_tfidf = pd.DataFrame(tsne_data_tfidf, columns = ("1st_Dim", "2nd_Dim", "Labels"))
```

```
In [103]:
```

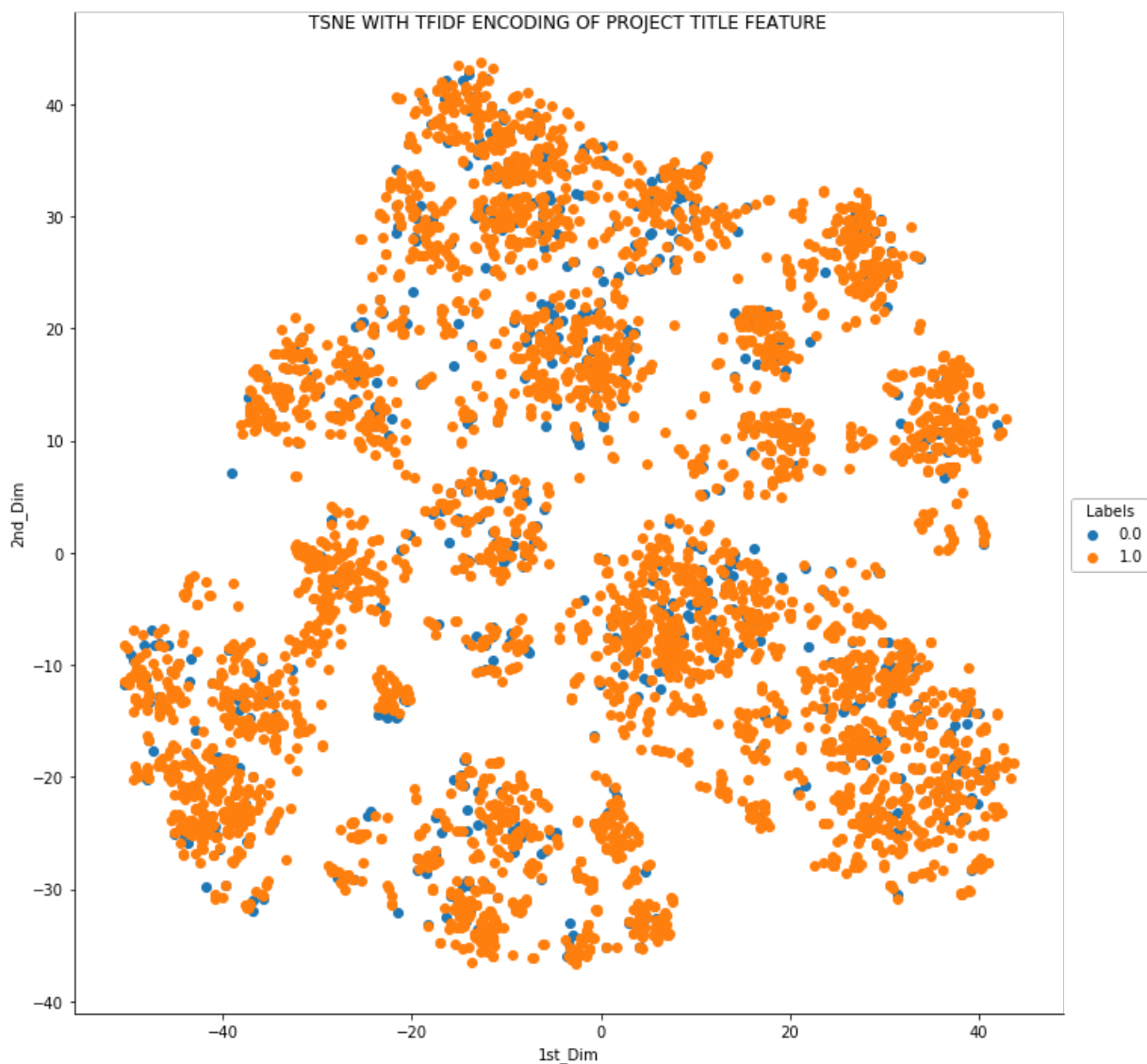
```
tsne_df_tfidf.shape
```

```
Out[103]:
```

```
(5000, 3)
```

```
In [121]:
```

```
please write all the code with proper documentation, and proper titles for each subsection
when you plot any graph make sure you use
a. Title, that describes your plot, this will be very helpful to the reader
b. Legends if needed
c. X-axis label
d. Y-axis label
sns.FacetGrid(tsne_df_tfidf, hue = "Labels", size = 10).map(plt.scatter, "1st_Dim", "2nd_Dim").add_
legend().fig.suptitle("TSNE WITH TFIDF ENCODING OF PROJECT TITLE FEATURE ")
plt.show()
```



Summary: 1 From the above plot all points are clustered and hence we cannot use tsne with TFIDF encoding to categorize projects

Summary: If from the above plot all points are clustered and hence we cannot use t-SNE with TF-IDF encoding to categorize projects approved

## 2.3 TSNE with `AVG W2V` encoding of `project\_title` feature

In [105]:

```
from scipy.sparse import hstack
with the same hstack function we are concatenating a sparse matrix and a dense matrix :)
X = hstack((categories_one_hot,
sub_categories_one_hot,teach_state_one_hot,school_state_one_hot,project_grade_one_hot,previous_star
dardized , price_standardized,avg_w2v_vectors_title))
print(X.shape)
```

(109248, 401)

In [106]:

```
X = X.tocsr()
X_new = X[0:5000,:]
```

In [107]:

```
X_new = X_new.toarray()
model = TSNE(n_components = 2, perplexity = 100.0, random_state = 0)
tsne_data_avgw2v = model.fit_transform(X_new)
tsne_data_avgw2v = np.vstack((tsne_data_avgw2v.T, labels_new)).T
tsne_df_avgw2v = pd.DataFrame(tsne_data_avgw2v, columns = ("1st_Dim","2nd_Dim","Labels"))
```

In [108]:

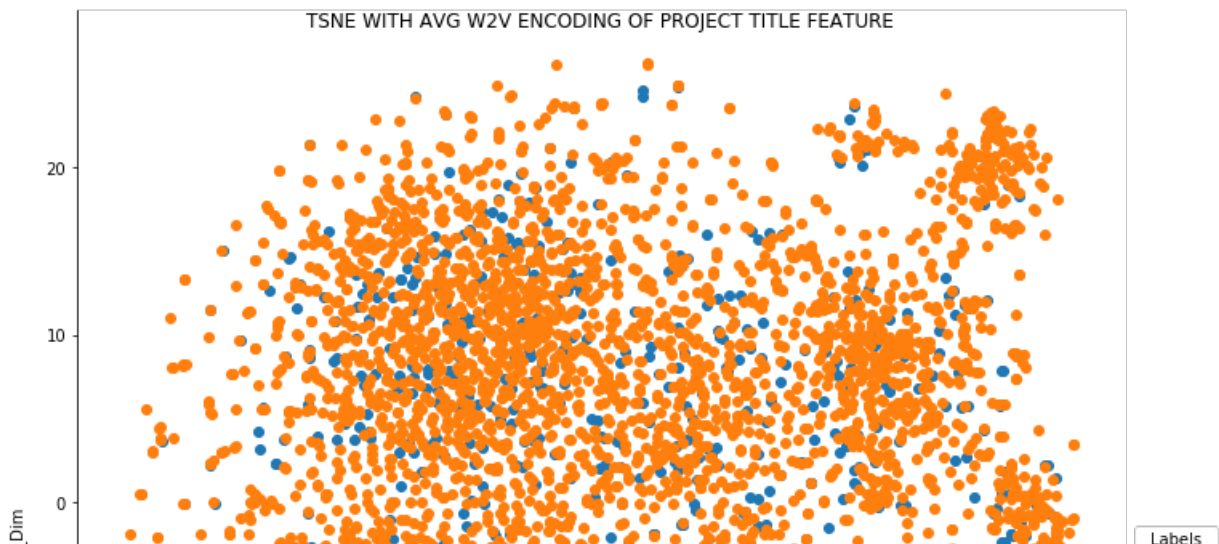
```
tsne_df_avgw2v.shape
```

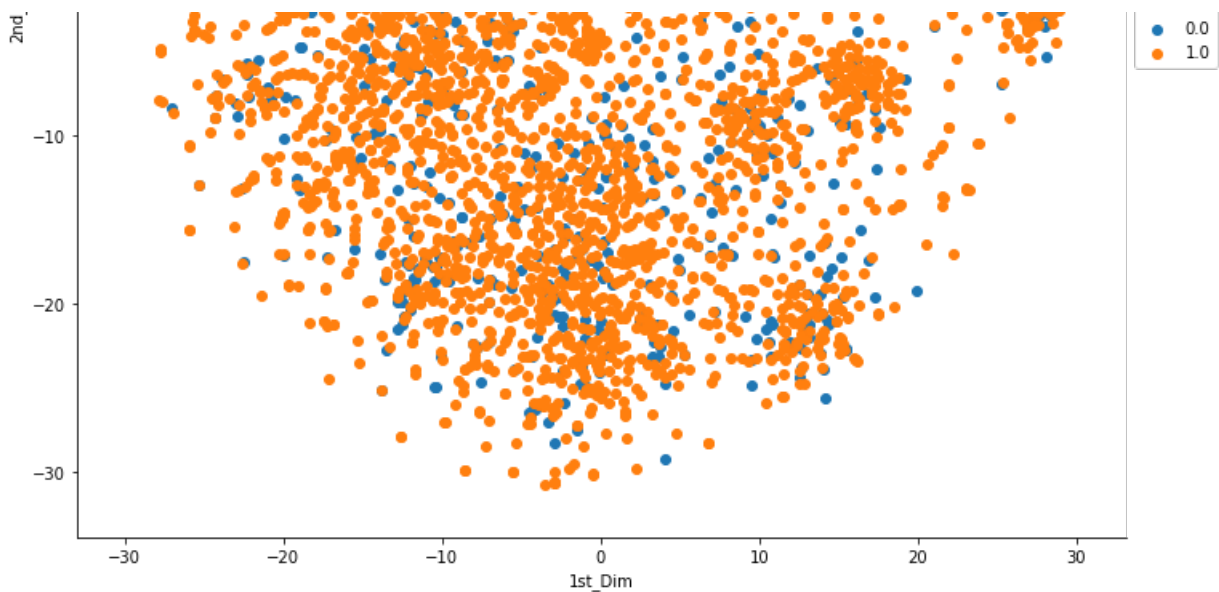
Out[108]:

(5000, 3)

In [109]:

```
please write all the code with proper documentation, and proper titles for each subsection
when you plot any graph make sure you use
a. Title, that describes your plot, this will be very helpful to the reader
b. Legends if needed
c. X-axis label
d. Y-axis label
sns.FacetGrid(tsne_df_avgw2v, hue = "Labels", size = 10).map(plt.scatter, "1st_Dim", "2nd_Dim").add
_legend().fig.suptitle("TSNE WITH AVG W2V ENCODING OF PROJECT TITLE FEATURE ")
plt.show()
```





Summary: 1. From the above plot all points are clustered and hence we cannot use tsne with avgw2v encoding to categorize projects approved

## 2.4 TSNE with `TFIDF Weighted W2V` encoding of `project\_title` feature

In [110]:

```
from scipy.sparse import hstack
with the same hstack function we are concatenating a sparse matrix and a dense matrix :)
X = hstack((categories_one_hot,
sub_categories_one_hot, teach_state_one_hot, school_state_one_hot, project_grade_one_hot, previous_star
dardized , price_standardized, tfidf_w2v_vectors))
print(X.shape)
```

(109248, 401)

In [111]:

```
X = X.tocsr()
X_new = X[0:5000,:]
```

In [112]:

```
X_new = X_new.toarray()
model = TSNE(n_components = 2, perplexity = 100.0, random_state = 0)
tsne_data_tfidfw2v = model.fit_transform(X_new)
tsne_data_tfidfw2v = np.vstack((tsne_data_tfidfw2v.T, labels_new)).T
tsne_df_tfidfw2v = pd.DataFrame(tsne_data_tfidfw2v, columns = ("1st_Dim", "2nd_Dim", "Labels"))
```

In [113]:

```
tsne_df_tfidfw2v.shape
```

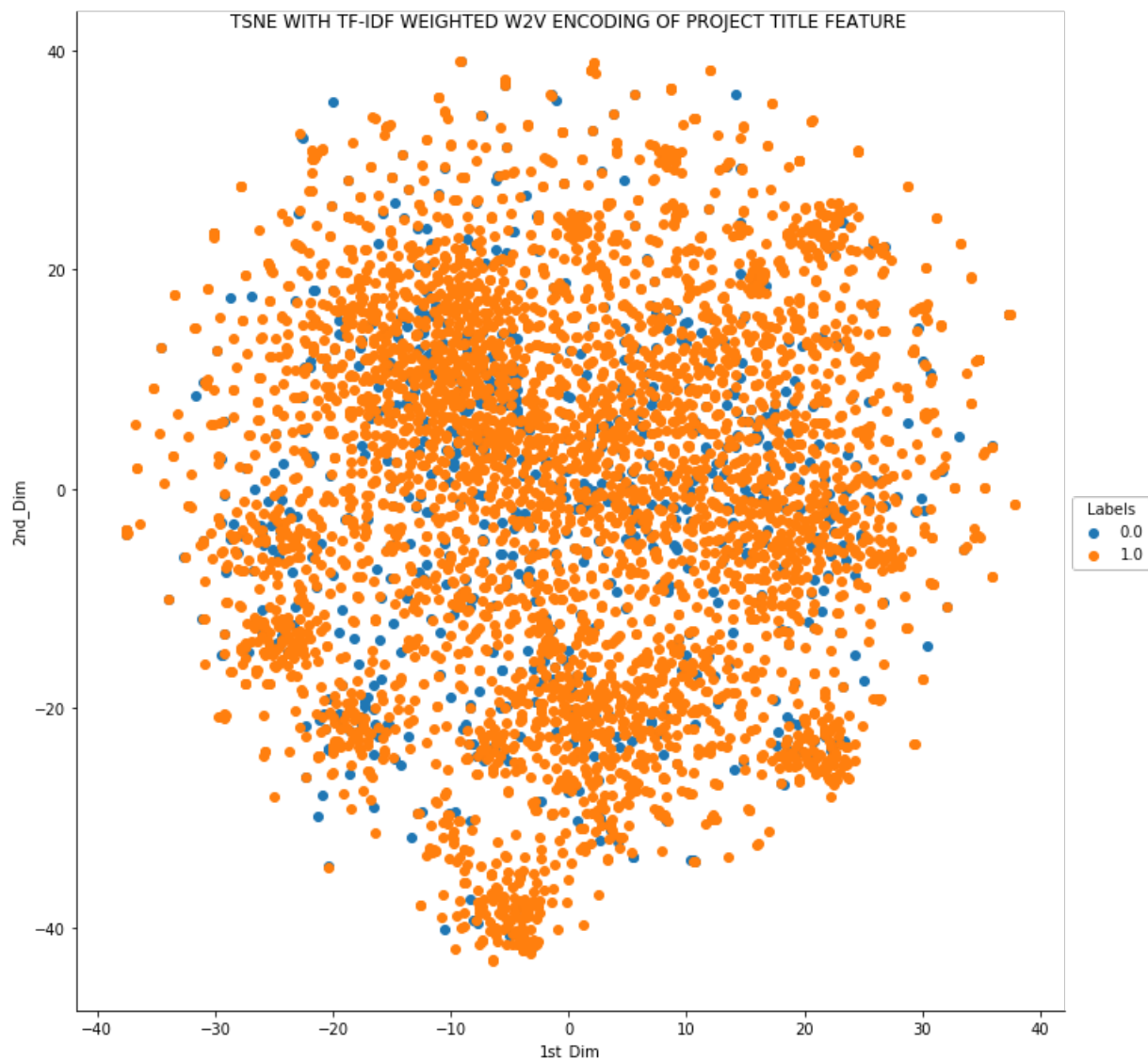
Out[113]:

(5000, 3)

In [114]:

```
please write all the code with proper documentation, and proper titles for each subsection
when you plot any graph make sure you use
a. Title, that describes your plot, this will be very helpful to the reader
b. Legends if needed
c. X-axis label
d. Y-axis label
sns.FacetGrid(tsne_df_tfidfw2v, hue = "Labels", size = 10).map(plt.scatter, "1st_Dim", "2nd_Dim").
```

```
add_legend().fig.suptitle("TSNE WITH TF-IDF WEIGHTED W2V ENCODING OF PROJECT TITLE FEATURE ")
plt.show()
```



Summary:1.From the above plot all points are clustered and hence we cannot use tsne with TFIDF weighted W2V encoding to categorize projects approved

## 2.5 TSNE with BOW, TFIDF, AVG W2V, TFIDF Weighted W2V encoding of project\_title feature

In [115]:

```
X = hstack((categories_one_hot,
sub_categories_one_hot,teach_state_one_hot,school_state_one_hot,project_grade_one_hot,previous_star
dardized , price_standardized,title_bow,title_tfidf,tfidf_w2v_vectors,avg_w2v_vectors_title))
print(X.shape)
```

(109248, 7359)

In [116]:

```
X = X.tocsr()
X_new = X[0:5000,:]
```

In [117]:

```
X_new = X_new.toarray()
model = TSNE(n_components = 2, perplexity = 100.0, random_state = 0)
tsne_data_complete = model.fit_transform(X_new)
tsne_data_complete = np.vstack((tsne_data_complete.T, labels_new)).T
tsne_df_complete = pd.DataFrame(tsne_data_complete, columns = ("1st_Dim", "2nd_Dim", "Labels"))
```

In [118]:

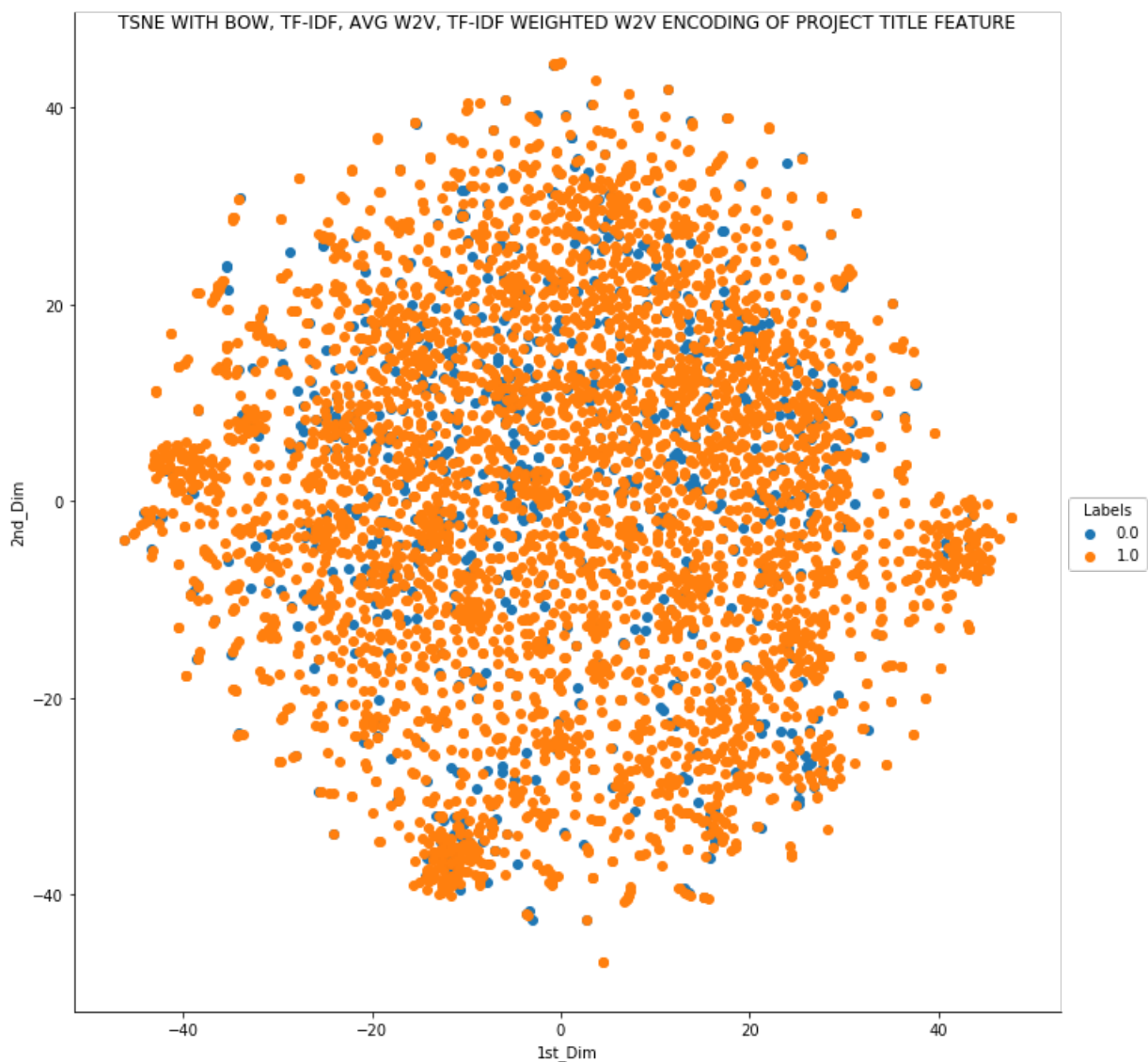
```
tsne_df_complete.shape
```

Out[118]:

```
(5000, 3)
```

In [120]:

```
sns.FacetGrid(tsne_df_complete, hue = "Labels", size = 10).map(plt.scatter, "1st_Dim", "2nd_Dim").add_legend().fig.suptitle("TSNE WITH BOW, TF-IDF, AVG W2V, TF-IDF WEIGHTED W2V ENCODING OF PROJECT TITLE FEATURE ")
plt.show()
```



Summary:1.From the above plot all points are clustered and hence we cannot use tsne with combination of BOW,Avg W2V,TFIDF,TFIDF weighted W2V encoding to categorize projects approved

## 2.6 CONCLUSION

We observe TSNE doesnot yield expected results as it is unclear to determine results from Bag of words,Avg W2V,TFIDF,TFIDF weighted W2V.So we have to think of any other methods to categorize projects approved