DonorsChoose

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

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

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

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

About the DonorsChoose Data Set

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

Feature	Description
`project_id`	A unique identifier for the proposed project. **Example:** `p036502`
`project_title`	Title of the project. **Examples:** • `Art Will Make You Happy!` • `First Grade Fun`
`project_grade_category`	Grade level of students for which the project is targeted. One of the following enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`
`project_subject_categories`	One or more (comma-separated) subject categories for the project from the following enumerated list of values: • `Applied Learning` • `Care & Hunger` • `Health & Sports` • `History & Civics` • `Literacy & Language` • `Math & Science` • `Music & The Arts` • `Special Needs` • `Warmth` **Examples:** • `Music & The Arts` • `Literacy & Language, Math & Science`
`school_state`	State where school is located (Two-letter U.S. postal code). **Example:** `WY`
`project_subject_subcategories`	One or more (comma-separated) subject subcategories for the project. **Examples:** • `Literacy` • `Literature & Writing, Social Sciences`
`project_resource_summary`	An explanation of the resources needed for the project. **Example:** • `My students need hands on literacy materials to manage sensory needs!`
`project_essay_1`	First application essay*
`project_essay_2`	Second application essay*
`project_essay_3`	Third application essay*

Feature	Description
`project_essay_4`	Fourth application essay
`project_submitted_datetime`	Datetime when project application was submitted.**Example:** `2016-04-28 12:43:56.245`
`teacher_id`	A unique identifier for the teacher of the proposed project.**Example:** `bdf8baa8fedef6bfeec7ae4ff1c15c56`
`teacher_prefix`	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
`teacher_number_of_previously_posted_projects`	Number of project applications previously submitted by the same teacher.**Example:** `2`

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

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

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

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

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

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

Notes on the Essay Data

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

- __project_essay_1:__ "Introduce us to your classroom"
- __project_essay_2:__ "Tell us more about your students"
- __project_essay_3:__ "Describe how your students will use the materials you're requesting"
- __project_essay_3:__ "Close by sharing why your project will make a difference"

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

- __project_essay_1:__ "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."
- __project_essay_2:__ "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

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

In [1]:

```
%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
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

import time
from tqdm import tqdm
import os
import pickle

from chart_studio 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 [2]:

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

In [3]:

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

```
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']
```

In [4]:

```
# how to replace elements in list python: https://stackoverflow.com/a/2582163/4084039
cols = ['Date' if x=='project_submitted_datetime' else x for x in list(project_data.columns)]


#sort dataframe based on time pandas python: https://stackoverflow.com/a/49702492/4084039
project_data['Date'] = pd.to_datetime(project_data['project_submitted_datetime'])
project_data.drop('project_submitted_datetime', axis=1, inplace=True)
project_data.sort_values(by=['Date'], inplace=True)


# how to reorder columns pandas python: https://stackoverflow.com/a/13148611/4084039
project_data = project_data[cols]


project_data.head(2)
```

Out[4]:

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

In [5]:

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

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

# Join two data frames
project_data = pd.merge(project_data, price_data, on='id', how='left')
project_data.head(5)
```

```
Number of data points in train data (1541272, 4)
['id' 'description' 'quantity' 'price']
        id                                        description  quantity  \
0  p233245  LC652 - Lakeshore Double-Space Mobile Drying Rack         1
1  p069063        Bouncy Bands for Desks (Blue support pipes)         3

    price
0  149.00
1   14.95
```

Out[5]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_subje
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2	M
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5	
2	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Grades PreK-2	Litera
3	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Grades PreK-2	Ap
4	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Grades 3-5	Litera

1.2 preprocessing of `project_subject_categories`

In [6]:

```
catogories = 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 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 & H
unger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Scienc
e"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i
.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math &
Science"=>"Math&Science"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())

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

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

cat_dict = dict(my_counter)
sorted_cat_dict = dict(sorted(cat_dict.items(), key=lambda kv: kv[1]))
```

1.3 preprocessing of `project_subject_subcategories`

In [7]:

```
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 & H
unger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Scienc
e"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i
.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math &
Science"=>"Math&Science"
        temp +=j.strip()+" "#" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_')
    sub_cat_list.append(temp.strip())

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

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

sub_cat_dict = dict(my_counter)
sorted_sub_cat_dict = dict(sorted(sub_cat_dict.items(), key=lambda kv: kv[1]))
```

1.3 Text preprocessing

1.3 Text preprocessing

In [8]:

```
# 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 [9]:

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

Out[9]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_title
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2	Engineering STEAM into the Primary Classroom
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5	Sensory Tools for Focus

In [10]:

```
#### 1.4.2.3 Using Pretrained Models: TFIDF weighted W2V
```

In [11]:

```
# printing some random reviews
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)
```

```
I have been fortunate enough to use the Fairy Tale STEM kits in my classroom as well as the STEM j
ournals, which my students really enjoyed.  I would love to implement more of the Lakeshore STEM k
its in my classroom for the next school year as they provide excellent and engaging STEM
lessons.My students come from a variety of backgrounds, including language and socioeconomic statu
s.  Many of them don't have a lot of experience in science and engineering and these kits give me
the materials to provide these exciting opportunities for my students.Each month I try to do
several science or STEM/STEAM projects.  I would use the kits and robot to help guide my science i
nstruction in engaging and meaningful ways.  I can adapt the kits to my current language arts paci
ng guide where we already teach some of the material in the kits like tall tales (Paul Bunyan) or
Johnny Appleseed.  The following units will be taught in the next school year where I will
implement these kits: magnets, motion, sink vs. float, robots.  I often get to these units and don
't know If I am teaching the right way or using the right materials.    The kits will give me
additional ideas, strategies, and lessons to prepare my students in science.It is challenging to d
evelop high quality science activities.  These kits give me the materials I need to provide my
students with science activities that will go along with the curriculum in my classroom.  Although
I have some things (like magnets) in my classroom, I don't know how to use them effectively.  The
kits will provide me with the right amount of materials and show me how to use them in an
appropriate way.
==================================================
I teach high school English to students with learning and behavioral disabilities. My students all
vary in their ability level. However, the ultimate goal is to increase all students literacy level
s. This includes their reading, writing, and communication levels.I teach a really dynamic group o
f students. However, my students face a lot of challenges. My students all live in poverty and in
a dangerous neighborhood. Despite these challenges, I have students who have the the desire to def
```

a dangerous neighborhood. Despite these challenges, I have students who have the the desire to defeat these challenges. My students all have learning disabilities and currently all are performing below grade level. My students are visual learners and will benefit from a classroom that fulfills their preferred learning style.The materials I am requesting will allow my students to be prepared for the classroom with the necessary supplies. Too often I am challenged with students who come to school unprepared for class due to economic challenges. I want my students to be able to focus on learning and not how they will be able to get school supplies. The supplies will last all year. Students will be able to complete written assignments and maintain a classroom journal. The chart paper will be used to make learning more visual in class and to create posters to aid students in their learning. The students have access to a classroom printer. The toner will be used to print student work that is completed on the classroom Chromebooks.I want to try and remove all barriers for the students learning and create opportunities for learning. One of the biggest barriers is the students not having the resources to get pens, paper, and folders. My students will be able to increase their literacy skills because of this project.

==

\"Life moves pretty fast. If you don't stop and look around once in awhile, you could miss it.\" from the movie, Ferris Bueller's Day Off. Think back...what do you remember about your grandparents? How amazing would it be to be able to flip through a book to see a day in their lives?My second graders are voracious readers! They love to read both fiction and nonfiction books. Their favorite characters include Pete the Cat, Fly Guy, Piggie and Elephant, and Mercy Watson. They also love to read about insects, space and plants. My students are hungry bookworms! My students are eager to learn and read about the world around them. My kids love to be at school and are like little sponges absorbing everything around them. Their parents work long hours and usually do not see their children. My students are usually cared for by their grandparents or a family friend. Most of my students do not have someone who speaks English at home. Thus it is difficult for my students to acquire language.Now think forward... wouldn't it mean a lot to your kids, nieces or nephews or grandchildren, to be able to see a day in your life today 30 years from now? Memories are so precious to us and being able to share these memories with future generations will be a rewarding experience. As part of our social studies curriculum, students will be learning about changes over time. Students will be studying photos to learn about how their community has changed over time. In particular, we will look at photos to study how the land, buildings, clothing, and schools have changed over time. As a culminating activity, my students will capture a slice of their history and preserve it through scrap booking. Key important events in their young lives will be documented with the date, location, and names.   Students will be using photos from home and from school to create their second grade memories.   Their scrap books will preserve their unique stories for future generations to enjoy.Your donation to this project will provide my second graders with an opportunity to learn about social studies in a fun and creative manner. Through their scrapbooks, children will share their story with others and have a historical document for the rest of their lives.

==

\"A person's a person, no matter how small.\" (Dr.Seuss) I teach the smallest students with the biggest enthusiasm for learning. My students learn in many different ways using all of our senses and multiple intelligences. I use a wide range of techniques to help all my students succeed. \r\nStudents in my class come from a variety of different backgrounds which makes for wonderful sharing of experiences and cultures, including Native Americans.\r\nOur school is a caring community of successful learners which can be seen through collaborative student project based learning in and out of the classroom. Kindergarteners in my class love to work with hands-on materials and have many different opportunities to practice a skill before it is mastered. Having the social skills to work cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is the perfect place to learn about agriculture and nutrition. My students love to role play in our pretend kitchen in the early childhood classroom. I have had several kids ask me, \"Can we try cooking with REAL food?\" I will take their idea and create \"Common Core Cooking Lessons\" where we learn important math and writing concepts while cooking delicious healthy food for snack time. My students will have a grounded appreciation for the work that went into making the food and knowledge of where the ingredients came from as well as how it's healthy for their bodies. This project would expand our learning of nutrition and agricultural cooking recipes by having us peel our own apples to make homemade applesauce, make our own bread, and mix up healthy plants from our classroom garden in the spring. We will also create our own cookbooks to be printed and shared with families. \r\nStudents will gain math and literature skills as well as a life long enjoyment for healthy cooking.nannan

==

My classroom consists of twenty-two amazing sixth graders from different cultures and backgrounds. They are a social bunch who enjoy working in partners and working with groups. They are hard-working and eager to head to middle school next year. My job is to get them ready to make this transition and make it as smooth as possible. In order to do this, my students need to come to school every day and feel safe and ready to learn. Because they are getting ready to head to middle school, I give them lots of choice- choice on where to sit and work, the order to complete assignments, choice of projects, etc. Part of the students feeling safe is the ability for them to come into a welcoming, encouraging environment. My room is colorful and the atmosphere is casual. I want them to take ownership of the classroom because we ALL share it together. Because my time with them is limited, I want to ensure they get the most of this time and enjoy it to the best of their abilities.Currently, we have twenty-two desks of differing sizes, yet the desks are similar to the ones the students will use in middle school. We also have a kidney table with crates for seating. I allow my students to choose their own spots while they are working independently or in groups. More often than not, most of them move out of their desks and onto the crates. Believe it or not, this has proven to be more successful than making them stay at their desks! It is because of this that I am looking toward the “Flexible Seating” option for my classroom.\r\n The students look forward to their work time so they can move around the room. I would like to get rid of the s

```
look forward to their work time so they can move around the room. I would like to get rid of the c
onstricting desks and move toward more “fun” seating options. I am requesting various seating so m
y students have more options to sit. Currently, I have a stool and a papasan chair I inherited fro
m the previous sixth-grade teacher as well as five milk crate seats I made, but I would like to gi
ve them more options and reduce the competition for the “good seats”. I am also requesting two rug
s as not only more seating options but to make the classroom more welcoming and appealing. In orde
r for my students to be able to write and complete work without desks, I am requesting a class set
of clipboards. Finally, due to curriculum that requires groups to work together, I am requesting t
ables that we can fold up when we are not using them to leave more room for our flexible seating o
ptions.\r\nI know that with more seating options, they will be that much more excited about coming
to school! Thank you for your support in making my classroom one students will remember
forever!nannan
==================================================
```

In [12]:

```
# 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 [13]:

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

```
\"A person is a person, no matter how small.\" (Dr.Seuss) I teach the smallest students with the b
iggest enthusiasm for learning. My students learn in many different ways using all of our senses a
nd multiple intelligences. I use a wide range of techniques to help all my students succeed. \r\nS
tudents in my class come from a variety of different backgrounds which makes for wonderful sharing
of experiences and cultures, including Native Americans.\r\nOur school is a caring community of su
ccessful learners which can be seen through collaborative student project based learning in and ou
t of the classroom. Kindergarteners in my class love to work with hands-on materials and have many
different opportunities to practice a skill before it is mastered. Having the social skills to wor
k cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is the
perfect place to learn about agriculture and nutrition. My students love to role play in our
pretend kitchen in the early childhood classroom. I have had several kids ask me, \"Can we try coo
king with REAL food?\" I will take their idea and create \"Common Core Cooking Lessons\" where we
learn important math and writing concepts while cooking delicious healthy food for snack time. My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it is healthy for their bodies. This project
would expand our learning of nutrition and agricultural cooking recipes by having us peel our own
apples to make homemade applesauce, make our own bread, and mix up healthy plants from our classro
om garden in the spring. We will also create our own cookbooks to be printed and shared with famil
ies. \r\nStudents will gain math and literature skills as well as a life long enjoyment for health
y cooking.nannan
==================================================
```

In [14]:

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

```
 A person is a person, no matter how small.  (Dr.Seuss) I teach the smallest students with the big
gest enthusiasm for learning. My students learn in many different ways using all of our senses and
```

```
multiple intelligences. I use a wide range of techniques to help all my students succeed.
Students in my class come from a variety of different backgrounds which makes for wonderful
sharing of experiences and cultures, including Native Americans.  Our school is a caring community
of successful learners which can be seen through collaborative student project based learning in a
nd out of the classroom. Kindergarteners in my class love to work with hands-on materials and have
many different opportunities to practice a skill before it is mastered. Having the social skills t
o work cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is
the perfect place to learn about agriculture and nutrition. My students love to role play in our p
retend kitchen in the early childhood classroom. I have had several kids ask me,  Can we try cooki
ng with REAL food?  I will take their idea and create  Common Core Cooking Lessons  where we learn
important math and writing concepts while cooking delicious healthy food for snack time. My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it is healthy for their bodies. This project
would expand our learning of nutrition and agricultural cooking recipes by having us peel our own
apples to make homemade applesauce, make our own bread, and mix up healthy plants from our classro
om garden in the spring. We will also create our own cookbooks to be printed and shared with famil
ies.   Students will gain math and literature skills as well as a life long enjoyment for healthy
cooking.nannan
```

In [15]:

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

```
 A person is a person no matter how small Dr Seuss I teach the smallest students with the biggest
enthusiasm for learning My students learn in many different ways using all of our senses and multi
ple intelligences I use a wide range of techniques to help all my students succeed Students in my
class come from a variety of different backgrounds which makes for wonderful sharing of
experiences and cultures including Native Americans Our school is a caring community of successful
learners which can be seen through collaborative student project based learning in and out of the
classroom Kindergarteners in my class love to work with hands on materials and have many different
opportunities to practice a skill before it is mastered Having the social skills to work
cooperatively with friends is a crucial aspect of the kindergarten curriculum Montana is the
perfect place to learn about agriculture and nutrition My students love to role play in our
pretend kitchen in the early childhood classroom I have had several kids ask me Can we try cooking
with REAL food I will take their idea and create Common Core Cooking Lessons where we learn
important math and writing concepts while cooking delicious healthy food for snack time My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it is healthy for their bodies This project w
ould expand our learning of nutrition and agricultural cooking recipes by having us peel our own a
pples to make homemade applesauce make our own bread and mix up healthy plants from our classroom
garden in the spring We will also create our own cookbooks to be printed and shared with families
Students will gain math and literature skills as well as a life long enjoyment for healthy cooking
nannan
```

In [16]:

```python
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords= ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', '
while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'
, 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll'
, 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "do
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn', \
"mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"]
```

In [17]:

```
# Create function that will filter sentance
def filterSentance(sentance):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    sent = sent.lower()
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    return sent.strip()
```

In [18]:

```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_essays = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['essay'].values):
    preprocessed_essays.append(filterSentance(sentance))
```

```
100%|██████████| 109248/109248 [00:56<00:00, 1947.70it/s]
```

In [19]:

```
# after preprocessing
preprocessed_essays[20000]
```

Out[19]:

```
'person person no matter small dr seuss teach smallest students biggest enthusiasm learning
students learn many different ways using senses multiple intelligences use wide range techniques h
elp students succeed students class come variety different backgrounds makes wonderful sharing exp
eriences cultures including native americans school caring community successful learners seen coll
aborative student project based learning classroom kindergarteners class love work hands materials
many different opportunities practice skill mastered social skills work cooperatively friends cruc
ial aspect kindergarten curriculum montana perfect place learn agriculture nutrition students love
role play pretend kitchen early childhood classroom several kids ask try cooking real food take id
ea create common core cooking lessons learn important math writing concepts cooking delicious heal
thy food snack time students grounded appreciation work went making food knowledge ingredients cam
e well healthy bodies project would expand learning nutrition agricultural cooking recipes us peel
apples make homemade applesauce make bread mix healthy plants classroom garden spring also create
cookbooks printed shared families students gain math literature skills well life long enjoyment he
althy cooking nannan'
```

1.4 Preprocessing of `project_title`

In [20]:

```
# similarly you can preprocess the titles also
# Combining all the above stundents
from tqdm import tqdm
preprocessed_titles = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['project_title'].values):
    preprocessed_titles.append(filterSentance(sentance))
```

```
100%|██████████| 109248/109248 [00:02<00:00, 37461.92it/s]
```

In [21]:

```
# after preprocessing
```

```
print(preprocessed_titles[20000])
```

```
health nutritional cooking kindergarten
```

In [22]:

```
# similarly you can preprocess the project_resource_summary also
# Combining all the above stundents
from tqdm import tqdm
preprocessed_resource_summary = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['project_resource_summary'].values):
    preprocessed_resource_summary.append(filterSentance(sentance))
```

```
100%|██████████| 109248/109248 [00:05<00:00, 20868.80it/s]
```

In [23]:

```
# after preprocessing
print(preprocessed_resource_summary[20000])
```

```
students need cooking supplies help us healthy learn nutrition mixer apple spiralizer kitchen
tools nutrition kit kid friendly healthy literature ink make cookbooks
```

In [24]:

```
# Preprocess teacher_prefix
from tqdm import tqdm
preprocessed_teacher_prefix = []
# tqdm is for printing the status bar
for teacher_prefix in tqdm(project_data['teacher_prefix'].values):
    teacher_prefix = str(teacher_prefix)
    clean_teacher_prefix = decontracted(teacher_prefix)
    clean_teacher_prefix = clean_teacher_prefix.replace('\\r', ' ')
    clean_teacher_prefix = clean_teacher_prefix.replace('\\"', ' ')
    clean_teacher_prefix = clean_teacher_prefix.replace('\\n', ' ')
    clean_teacher_prefix = re.sub('[^A-Za-z0-9]+', ' ', clean_teacher_prefix)
    clean_teacher_prefix = clean_teacher_prefix.lower()
    if clean_teacher_prefix in stopwords:
        continue
    preprocessed_teacher_prefix.append(clean_teacher_prefix.strip())
```

```
100%|██████████| 109248/109248 [00:01<00:00, 79767.45it/s]
```

In [25]:

```
preprocessed_teacher_prefix[0:10]
```

Out[25]:

```
['mrs', 'ms', 'mrs', 'mrs', 'mrs', 'mrs', 'mrs', 'ms', 'ms', 'mrs']
```

In [26]:

```
# Preprocess project_grade_category
from tqdm import tqdm
preprocessed_project_grade_category = []
# tqdm is for printing the status bar
for project_grade_category in tqdm(project_data['project_grade_category'].values):
    project_grade_category = str(project_grade_category)
    clean_project_grade_category = decontracted(project_grade_category)
    clean_project_grade_category = clean_project_grade_category.replace('\\r', ' ')
    clean_project_grade_category = clean_project_grade_category.replace('\\"', ' ')
    clean_project_grade_category = clean_project_grade_category.replace('\\n', ' ')
    clean_project_grade_category = re.sub('[^A-Za-z0-9]+', ' ', clean_project_grade_category)
    clean_project_grade_category = clean_project_grade_category.lower()
    if clean_project_grade_category in stopwords:
        continue
    clean_project_grade_category = clean_project_grade_category.strip()
#     whitespaces are creating problems because we are treating this as categorical feature
#     whitespace are creating problems because we are treating this as categorical feature
    preprocessed_project_grade_category.append(clean_project_grade_category.replace(' ', '_'))
```

```
100%|██████████| 109248/109248 [00:01<00:00, 76807.88it/s]
```

In [27]:

```
preprocessed_project_grade_category[0:10]
```

Out[27]:

```
['grades_prek_2',
 'grades_3_5',
 'grades_prek_2',
 'grades_prek_2',
 'grades_3_5',
 'grades_3_5',
 'grades_3_5',
 'grades_3_5',
 'grades_prek_2',
 'grades_3_5']
```

In [28]:

```
# Replace original columns with preprocessed column values
project_data['clean_essays'] = preprocessed_essays
project_data['clean_titles'] = preprocessed_titles
project_data['project_resource_summary'] = preprocessed_resource_summary
project_data['teacher_prefix'] = preprocessed_teacher_prefix
project_data['project_grade_category'] = preprocessed_project_grade_category
# Drop essays column
project_data.drop(['project_essay_1'], axis=1, inplace=True)
project_data.drop(['project_essay_2'], axis=1, inplace=True)
project_data.drop(['project_essay_3'], axis=1, inplace=True)
project_data.drop(['project_essay_4'], axis=1, inplace=True)
```

In [29]:

```
project_data.head(5)
```

Out[29]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_title
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	mrs	CA	2016-04-27 00:27:36	grades_prek_2	Engineering STEAM into the Primary Classroom
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	ms	UT	2016-04-27 00:31:25	grades_3_5	Sensory Tools for Focus
2	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	mrs	CA	2016-04-27 00:46:53	grades_prek_2	Mobile Learning with a Mobile Listening Center
3	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	mrs	GA	2016-04-27 00:53:00	grades_prek_2	Flexible Seating for Flexible Learning
4	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	mrs	WA	2016-04-27 01:05:25	grades_3_5	Going Deep: The Art of Inner Thinking!

In [30]:

```
project_data.tail(5)
```

Out[30]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	projec
109243	45036	p194916	29cf137e5a40b0f141d9fd7898303a5c	mrs	HI	2017-04-30 23:11:45	grades_9_12	Na F Pro S
109244	12610	p162971	22fee80f2078c694c2d244d3ecb1c390	ms	NM	2017-04-30 23:23:24	grades_prek_2	Op Organ
109245	179833	p096829	c8c81a73e29ae3bdd4140be8ad0bea00	mrs	IL	2017-04-30 23:25:42	grades_3_5	B Agri Sustai t
109246	13791	p184393	65545a295267ad9df99f26f25c978fd0	mrs	HI	2017-04-30 23:27:07	grades_9_12	((M N
109247	124250	p028318	1fff5a88945be8b2c728c6a85c31930f	mrs	CA	2017-04-30 23:45:08	grades_prek_2	Ne

In [31]:

```
print(set(preprocessed_project_grade_category))
```

```
{'grades_9_12', 'grades_6_8', 'grades_3_5', 'grades_prek_2'}
```

In [32]:

```
project_data['teacher_prefix'] = project_data['teacher_prefix'].fillna('null')
```

In [33]:

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

Out[33]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_title
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	mrs	CA	2016-04-27 00:27:36	grades_prek_2	Engineering STEAM into the Primary Classroom
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	ms	UT	2016-04-27 00:31:25	grades_3_5	Sensory Tools for Focus

Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category	project_title

1.5 Preparing data for models

In [34]:

```
project_data.columns
```

Out[34]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'Date', 'project_grade_category', 'project_title',
       'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'project_is_approved',
       'quantity', 'price', 'clean_categories', 'clean_subcategories', 'essay',
       'clean_essays', 'clean_titles'],
      dtype='object')
```

we are going to consider

```
      - 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 (optinal)

      - quantity : numerical (optinal)
      - teacher_number_of_previously_posted_projects : numerical
      - price : numerical
```

In [35]:

```
print(project_data.shape)

# I am taking 30% of data points for my analysis
project_data = project_data.sample(frac=0.3)

print(project_data.shape)
```

```
(109248, 18)
(32774, 18)
```

In [36]:

```
# Assigning data
y = project_data['project_is_approved'].values
project_data.drop(['project_is_approved'], axis=1, inplace=True)
X = project_data
project_data.shape
```

Out[36]:

```
(32774, 17)
```

In [37]:

```
# Split Train, CV and Test data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, stratify=y)
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.33, stratify=y_train)

print('Train Data Set', X_train.shape, y_train.shape)
print('Cross Validate Data Set', X_cv.shape, y_cv.shape)
print('Cross Validate Data Set', X_cv.shape, y_cv.shape)
print('Test Data Set', X_test.shape, y_test.shape)
```

```
Train Data Set (14711, 17) (14711,)
Cross Validate Data Set (7247, 17) (7247,)
Test Data Set (10816, 17) (10816,)
```

In [38]:

```
print('Train Data Set', X_train.shape, y_train.shape)
print('Cross Validate Data Set', X_cv.shape, y_cv.shape)
print('Test Data Set', X_test.shape, y_test.shape)
print('*'*100)
```

```
Train Data Set (14711, 17) (14711,)
Cross Validate Data Set (7247, 17) (7247,)
Test Data Set (10816, 17) (10816,)
*************************************************************************************************
```

1.5.1 Vectorizing Categorical data

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

In [39]:

```
# One hot encoding of Categorical Feature
# - school_state : categorical data
vectorizer = CountVectorizer()
vectorizer.fit(X_train['school_state'].values)# Fit has to happen only on train data

X_train_school_state_ohe = vectorizer.transform(X_train['school_state'].values)
X_cv_school_state_ohe = vectorizer.transform(X_cv['school_state'].values)
X_test_school_state_ohe = vectorizer.transform(X_test['school_state'].values)

school_state_features = vectorizer.get_feature_names()

print(X_train_school_state_ohe.shape, y_train.shape)
print(X_cv_school_state_ohe.shape, y_cv.shape)
print(X_test_school_state_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print('*'*100)
```

```
(14711, 51) (14711,)
(7247, 51) (7247,)
(10816, 51) (10816,)
['ak', 'al', 'ar', 'az', 'ca', 'co', 'ct', 'dc', 'de', 'fl', 'ga', 'hi', 'ia', 'id', 'il', 'in', 'k
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']
*************************************************************************************************
```

In [40]:

```
# One hot encoding of Categorical Feature
# - clean_categories : categorical data
vectorizer = CountVectorizer()
vectorizer.fit(X_train['clean_categories'].values)# Fit has to happen only on train data

X_train_clean_categories_ohe = vectorizer.transform(X_train['clean_categories'].values)
X_cv_clean_categories_ohe = vectorizer.transform(X_cv['clean_categories'].values)
X_test_clean_categories_ohe = vectorizer.transform(X_test['clean_categories'].values)

clean_categories_features = vectorizer.get_feature_names()

print(X_train_clean_categories_ohe.shape, y_train.shape)
print(X_cv_clean_categories_ohe.shape, y_cv.shape)
print(X_test_clean_categories_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print('*'*100)
```

```
(14711, 9) (14711,)
(7247, 9) (7247,)
(10816, 9) (10816,)
['appliedlearning', 'care_hunger', 'health_sports', 'history_civics', 'literacy_language',
'math_science', 'music_arts', 'specialneeds', 'warmth']
***************************************************************************************************
```

In [41]:

```
# One hot encoding of Categorical Feature
# - clean_subcategories : categorical data
vectorizer = CountVectorizer()
vectorizer.fit(X_train['clean_subcategories'].values)# Fit has to happen only on train data

X_train_clean_subcategories_ohe = vectorizer.transform(X_train['clean_subcategories'].values)
X_cv_clean_subcategories_ohe = vectorizer.transform(X_cv['clean_subcategories'].values)
X_test_clean_subcategories_ohe = vectorizer.transform(X_test['clean_subcategories'].values)

clean_subcategories_features = vectorizer.get_feature_names()

print(X_train_clean_subcategories_ohe.shape, y_train.shape)
print(X_cv_clean_subcategories_ohe.shape, y_cv.shape)
print(X_test_clean_subcategories_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print('*'*100)
```

```
(14711, 30) (14711,)
(7247, 30) (7247,)
(10816, 30) (10816,)
['appliedsciences', 'care_hunger', 'charactereducation', 'civics_government',
'college_careerprep', 'communityservice', 'earlydevelopment', 'economics', 'environmentalscience',
'esl', 'extracurricular', 'financialliteracy', 'foreignlanguages', 'gym_fitness',
'health_lifescience', 'health_wellness', 'history_geography', 'literacy', 'literature_writing', 'm
athematics', 'music', 'nutritioneducation', 'other', 'parentinvolvement', 'performingarts', 'socia
lsciences', 'specialneeds', 'teamsports', 'visualarts', 'warmth']
***************************************************************************************************
```

In [42]:

```
print(X_train['project_grade_category'])
# One hot encoding of Categorical Feature
# - project_grade_category : categorical data
# Convert one hot encoding for project grade category
vectorizer = CountVectorizer()
vectorizer.fit(X_train['project_grade_category'].values)# Fit has to happen only on train data

X_train_project_grade_category_ohe = vectorizer.transform(X_train['project_grade_category'].values
)
X_cv_project_grade_category_ohe = vectorizer.transform(X_cv['project_grade_category'].values)
X_test_project_grade_category_ohe = vectorizer.transform(X_test['project_grade_category'].values)

project_grade_category_features = vectorizer.get_feature_names()

print(X_train_project_grade_category_ohe.shape, y_train.shape)
print(X_cv_project_grade_category_ohe.shape, y_cv.shape)
print(X_test_project_grade_category_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print('*'*100)
```

```
22433        grades_3_5
68758        grades_3_5
21494       grades_9_12
107392       grades_6_8
67845        grades_3_5
66687        grades_3_5
92835     grades_prek_2
82331     grades_prek_2
5990         grades_3_5
24638        grades_3_5
```

```
2658         grades_3_5
100370    grades_prek_2
64830       grades_9_12
69086     grades_prek_2
73004        grades_6_8
71055       grades_9_12
9424         grades_3_5
72558     grades_prek_2
11628        grades_6_8
49539        grades_3_5
22283     grades_prek_2
93441        grades_6_8
101545    grades_prek_2
96358     grades_prek_2
93463     grades_prek_2
96272        grades_3_5
462       grades_prek_2
85774        grades_3_5
107576       grades_3_5
82834        grades_3_5
              ...
28549        grades_6_8
79720       grades_9_12
66616     grades_prek_2
31730     grades_prek_2
77084     grades_prek_2
89813     grades_prek_2
8563         grades_6_8
29237        grades_3_5
33713        grades_3_5
22389        grades_6_8
91200        grades_6_8
101          grades_3_5
86689     grades_prek_2
42705     grades_prek_2
57290     grades_prek_2
72438        grades_3_5
70294     grades_prek_2
27090     grades_prek_2
61816        grades_6_8
73657     grades_prek_2
66584        grades_6_8
49435     grades_prek_2
104368       grades_3_5
84479     grades_prek_2
5203      grades_prek_2
109058      grades_9_12
37913     grades_prek_2
72075     grades_prek_2
94103        grades_3_5
31528        grades_6_8
Name: project_grade_category, Length: 14711, dtype: object
(14711, 4) (14711,)
(7247, 4) (7247,)
(10816, 4) (10816,)
['grades_3_5', 'grades_6_8', 'grades_9_12', 'grades_prek_2']
****************************************************************************************************
```

In [43]:

```
print(X_train_project_grade_category_ohe.toarray())
```

```
[[1 0 0 0]
 [1 0 0 0]
 [0 0 1 0]
 ...
 [0 0 0 1]
 [1 0 0 0]
 [0 1 0 0]]
```

In [44]:

```
# One hot encoding of Categorical Feature
# - teacher_prefix : categorical data
print(X_train['teacher_prefix'])
vectorizer = CountVectorizer()
vectorizer.fit(X_train['teacher_prefix'].values)# Fit has to happen only on train data

X_train_teacher_prefix_ohe = vectorizer.transform(X_train['teacher_prefix'].values)
X_cv_clean_teacher_prefix_ohe = vectorizer.transform(X_cv['teacher_prefix'].values)
X_test_clean_teacher_prefix_ohe = vectorizer.transform(X_test['teacher_prefix'].values)

teacher_prefix_features = vectorizer.get_feature_names()

print(X_train_teacher_prefix_ohe.shape, y_train.shape)
print(X_cv_clean_teacher_prefix_ohe.shape, y_cv.shape)
print(X_test_clean_teacher_prefix_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print('*'*100)
```

```
22433      mrs
68758       ms
21494       mr
107392      mr
67845       ms
66687       mr
92835      mrs
82331       ms
5990       mrs
24638      mrs
2658        ms
100370     mrs
64830       mr
69086      mrs
73004       ms
71055       ms
9424        ms
72558       ms
11628       ms
49539       ms
22283       ms
93441       mr
101545      ms
96358      mrs
93463       ms
96272       mr
462         mr
85774      mrs
107576     mrs
82834      mrs
         ...
28549       mr
79720      mrs
66616      mrs
31730      mrs
77084      mrs
89813       ms
8563       mrs
29237       ms
33713      mrs
22389       ms
91200       ms
101        mrs
86689      mrs
42705       ms
57290      mrs
72438      mrs
70294      mrs
27090      mrs
61816       ms
73657       ms
66584       mr
49435       ms
104368      ms
84479      mrs
5203        ms
109058      ms
37913       mr
72075      mrs
94103       ms
```

```
31528      ms
Name: teacher_prefix, Length: 14711, dtype: object
(14711, 5) (14711,)
(7247, 5) (7247,)
(10816, 5) (10816,)
['dr', 'mr', 'mrs', 'ms', 'teacher']
****************************************************************************************************
```

In [45]:

```
print(X_train_teacher_prefix_ohe.toarray())
```

```
[[0 0 1 0 0]
 [0 0 0 1 0]
 [0 1 0 0 0]
 ...
 [0 0 1 0 0]
 [0 0 0 1 0]
 [0 0 0 1 0]]
```

1.5.2 Vectorizing Text data

1.5.2.1 Bag of words

In [46]:

```
# - project_title : text data
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = CountVectorizer(min_df=10,ngram_range=(1,4), max_features=5000)
vectorizer.fit(X_train['clean_titles'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_title_bow = vectorizer.transform(X_train['clean_titles'].values)
X_cv_title_bow = vectorizer.transform(X_cv['clean_titles'].values)
X_test_title_bow = vectorizer.transform(X_test['clean_titles'].values)

clean_titles_bow_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_title_bow.shape, y_train.shape)
print(X_cv_title_bow.shape, y_cv.shape)
print(X_test_title_bow.shape, y_test.shape)
# print(vectorizer.get_feature_names())
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************

After vectorizations
(14711, 1112) (14711,)
(7247, 1112) (7247,)
(10816, 1112) (10816,)
****************************************************************************************************
```

In [47]:

```
# - text : text data
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = CountVectorizer(min_df=10,ngram_range=(1,4), max_features=5000)
vectorizer.fit(X_train['clean_essays'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_essay_bow = vectorizer.transform(X_train['clean_essays'].values)
X_cv_essay_bow = vectorizer.transform(X_cv['clean_essays'].values)
X_test_essay_bow = vectorizer.transform(X_test['clean_essays'].values)

easy_bow_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_essay_bow.shape, y_train.shape)
print(X_cv_essay_bow.shape, y_cv.shape)
print(X_test_essay_bow.shape, y_test.shape)
# print(vectorizer.get_feature_names())
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************

After vectorizations
(14711, 5000) (14711,)
(7247, 5000) (7247,)
(10816, 5000) (10816,)
****************************************************************************************************
```

In [48]:

```
# - project_resource_summary: text data (optinal)
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = CountVectorizer(min_df=10,ngram_range=(1,4), max_features=5000)
vectorizer.fit(X_train['project_resource_summary'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_project_resource_summary_bow = vectorizer.transform(X_train['project_resource_summary'].va
lues)
X_cv_project_resource_summary_bow = vectorizer.transform(X_cv['project_resource_summary'].values)
X_test_project_resource_summary_bow =
vectorizer.transform(X_test['project_resource_summary'].values)

project_resource_summary_bow_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_project_resource_summary_bow.shape, y_train.shape)
print(X_cv_project_resource_summary_bow.shape, y_cv.shape)
print(X_test_project_resource_summary_bow.shape, y_test.shape)
# print(vectorizer.get_feature_names())
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************

After vectorizations
(14711, 4121) (14711,)
(7247, 4121) (7247,)
(10816, 4121) (10816,)
****************************************************************************************************
```

1.5.2.2 TFIDF vectorizer

In [49]:

```
# - project_title : text data
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

from sklearn.feature_extraction.text import TfidfVectorizer
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = TfidfVectorizer(min_df=10)
vectorizer.fit(X_train['clean_titles'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_title_tfidf = vectorizer.transform(X_train['clean_titles'].values)
X_cv_title_tfidf = vectorizer.transform(X_cv['clean_titles'].values)
X_test_title_tfidf = vectorizer.transform(X_test['clean_titles'].values)

clean_titles_tfidf_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_title_tfidf.shape, y_train.shape)
print(X_cv_title_tfidf.shape, y_cv.shape)
print(X_cv_title_tfidf.shape, y_test.shape)
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************
After vectorizations
(14711, 826) (14711,)
(7247, 826) (7247,)
(7247, 826) (10816,)
****************************************************************************************************
```

In [50]:

```
# - text : text data
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

from sklearn.feature_extraction.text import TfidfVectorizer
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = TfidfVectorizer(min_df=10)#, ngram_range=(2,2), max_features=5000
vectorizer.fit(X_train['clean_essays'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_essay_tfidf = vectorizer.transform(X_train['clean_essays'].values)
X_cv_essay_tfidf = vectorizer.transform(X_cv['clean_essays'].values)
X_test_essay_tfidf = vectorizer.transform(X_test['clean_essays'].values)

easy_tfidf_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_essay_tfidf.shape, y_train.shape)
print(X_cv_essay_tfidf.shape, y_cv.shape)
print(X_test_essay_tfidf.shape, y_test.shape)
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************
```

```
After vectorizations
(14711, 7310) (14711,)
(7247, 7310) (7247,)
(10816, 7310) (10816,)
****************************************************************************************************
```

In [51]:

```
# - project_resource_summary: text data (optinal)
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("*"*100)

from sklearn.feature_extraction.text import TfidfVectorizer
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = TfidfVectorizer(min_df=10)
vectorizer.fit(X_train['project_resource_summary'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_project_resource_summary_tfidf = vectorizer.transform(X_train['project_resource_summary'].
values)
X_cv_project_resource_summary_tfidf = vectorizer.transform(X_cv['project_resource_summary'].values
)
X_test_project_resource_summary_tfidf =
vectorizer.transform(X_test['project_resource_summary'].values)

project_resource_summary_tfidf_features = vectorizer.get_feature_names()

print("After vectorizations")
print(X_train_project_resource_summary_tfidf.shape, y_train.shape)
print(X_cv_project_resource_summary_tfidf.shape, y_cv.shape)
print(X_test_project_resource_summary_tfidf.shape, y_test.shape)
print("*"*100)
```

```
(14711, 17) (14711,)
(7247, 17) (7247,)
(10816, 17) (10816,)
****************************************************************************************************

After vectorizations
(14711, 1886) (14711,)
(7247, 1886) (7247,)
(10816, 1886) (10816,)
****************************************************************************************************
```

1.5.2.3 Using Pretrained Models: Avg W2V

In [52]:

```
'''
# 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[52]:

```
'\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        m
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 =
set(model.keys())\nfor i in words:\n    if i in words_glove:\n        words_courpus[i] = model[i]\n
print("word 2 vec length", len(words_courpus))\n\n\n# stronging variables into pickle files python
: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-python/\n\nimport pic
kle\nwith open(\'glove_vectors\', \'wb\') as f:\n    pickle.dump(words_courpus, f)\n\n\n'
```

In [53]:

```
# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-sa
ve-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 [54]:

```
# average Word2Vec for train text
# compute average word2vec for each review.
avg_w2v_vectors_text_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['clean_essays'].values): # 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:
```

```
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors_text_train.append(vector)

print(len(avg_w2v_vectors_text_train))
print(len(avg_w2v_vectors_text_train[0]))
```

```
100%|██████████| 14711/14711 [00:03<00:00, 4280.88it/s]
```

```
14711
300
```

In [55]:

```
# average Word2Vec for CV text
# compute average word2vec for each review.
avg_w2v_vectors_text_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv['clean_essays'].values): # 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_text_cv.append(vector)

print(len(avg_w2v_vectors_text_cv))
print(len(avg_w2v_vectors_text_cv[0]))
```

```
100%|██████████| 7247/7247 [00:01<00:00, 4142.44it/s]
```

```
7247
300
```

In [56]:

```
# average Word2Vec for test text
# compute average word2vec for each review.
avg_w2v_vectors_text_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['clean_essays'].values): # 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_text_test.append(vector)

print(len(avg_w2v_vectors_text_test))
print(len(avg_w2v_vectors_text_test[0]))
```

```
100%|██████████| 10816/10816 [00:02<00:00, 4116.85it/s]
```

```
10816
300
```

In [57]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_title_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['clean_titles']): # 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_train.append(vector)

print(len(avg_w2v_vectors_title_train))
print(len(avg_w2v_vectors_title_train[0]))
```

```
100%|██████████| 14711/14711 [00:00<00:00, 69759.33it/s]
```

```
14711
300
```

In [58]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_title_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv['clean_titles']): # 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_cv.append(vector)

print(len(avg_w2v_vectors_title_cv))
print(len(avg_w2v_vectors_title_cv[0]))
```

```
100%|██████████| 7247/7247 [00:00<00:00, 69502.90it/s]
```

```
7247
300
```

In [59]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_title_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['clean_titles']): # 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_test.append(vector)

print(len(avg_w2v_vectors_title_test))
print(len(avg_w2v_vectors_title_test[0]))
```

```
100%|██████████| 10816/10816 [00:00<00:00, 67690.54it/s]
```

```
10816
300
```

In [60]:

```
# Similarly you can vectorize for project_resource_summary also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_resource_summary_train = []; # the avg-w2v for each sentence/review is sto
red in this list
for sentence in tqdm(X_train['project_resource_summary']): # 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_project_resource_summary_train.append(vector)

print(len(avg_w2v_vectors_project_resource_summary_train))
print(len(avg_w2v_vectors_project_resource_summary_train[0]))
```

```
100%|██████████| 14711/14711 [00:00<00:00, 33055.24it/s]
```

```
14711
300
```

In [61]:

```
# Similarly you can vectorize for project_resource_summary also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_resource_summary_cv = []; # the avg-w2v for each sentence/review is stored
in this list
for sentence in tqdm(X_cv['project_resource_summary']): # 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_project_resource_summary_cv.append(vector)

print(len(avg_w2v_vectors_project_resource_summary_cv))
print(len(avg_w2v_vectors_project_resource_summary_cv[0]))
```

```
100%|██████████| 7247/7247 [00:00<00:00, 31379.49it/s]
```

```
7247
300
```

In [62]:

```
# Similarly you can vectorize for project_resource_summary also
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_resource_summary_test = []; # the avg-w2v for each sentence/review is stor
ed in this list
for sentence in tqdm(X_test['project_resource_summary']): # 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_project_resource_summary_test.append(vector)

print(len(avg_w2v_vectors_project_resource_summary_test))
print(len(avg_w2v_vectors_project_resource_summary_test[0]))
```

```
100%|██████████| 10816/10816 [00:00<00:00, 33901.47it/s]
```

```
10816
300
```

1.5.2.3 Using Pretrained Models: TFIDF weighted W2V

In [63]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['clean_essays'])
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model.get_feature_names(), list(tfidf_model.idf_)))
tfidf_words = set(tfidf_model.get_feature_names())
```

In [64]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_text_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['clean_essays']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf 
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf 
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_text_train.append(vector)

print(len(tfidf_w2v_vectors_text_train))
print(len(tfidf_w2v_vectors_text_train[0]))
```

```
100%|██████████| 14711/14711 [00:25<00:00, 587.34it/s]
```

```
14711
300
```

In [65]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_text_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv['clean_essays']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_text_cv.append(vector)

print(len(tfidf_w2v_vectors_text_cv))
print(len(tfidf_w2v_vectors_text_cv[0]))
```

```
100%|██████████| 7247/7247 [00:12<00:00, 584.96it/s]
```

```
7247
300
```

In [66]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_text_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['clean_essays']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_text_test.append(vector)

print(len(tfidf_w2v_vectors_text_test))
print(len(tfidf_w2v_vectors_text_test[0]))
```

```
100%|██████████| 10816/10816 [00:18<00:00, 584.74it/s]
```

```
10816
300
```

In [67]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['clean_titles'])
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model.get_feature_names(), list(tfidf_model.idf_)))
tfidf_words = set(tfidf_model.get_feature_names())
```

In [68]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_title_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['clean_titles']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_title_train.append(vector)

print(len(tfidf_w2v_vectors_title_train))
print(len(tfidf_w2v_vectors_title_train[0]))
```

```
100%|██████████| 14711/14711 [00:00<00:00, 34328.59it/s]
```

```
14711
300
```

In [69]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_title_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv['clean_titles']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_title_cv.append(vector)

print(len(tfidf_w2v_vectors_title_cv))
print(len(tfidf_w2v_vectors_title_cv[0]))
```

```
100%|██████████| 7247/7247 [00:00<00:00, 37227.06it/s]
```

```
7247
300
```

In [70]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_title_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['clean_titles']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_title_test.append(vector)
```

```
print(len(tfidf_w2v_vectors_title_test))
print(len(tfidf_w2v_vectors_title_test[0]))
```

```
100%|██████████| 10816/10816 [00:00<00:00, 37749.01it/s]
```

```
10816
300
```

In [71]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['project_resource_summary'])
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model.get_feature_names(), list(tfidf_model.idf_)))
tfidf_words = set(tfidf_model.get_feature_names())
```

In [72]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_resource_summary_train = []; # the avg-w2v for each sentence/review is s
tored in this list
for sentence in tqdm(X_train['project_resource_summary']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_project_resource_summary_train.append(vector)

print(len(tfidf_w2v_vectors_project_resource_summary_train))
print(len(tfidf_w2v_vectors_project_resource_summary_train[0]))
```

```
100%|██████████| 14711/14711 [00:01<00:00, 13051.28it/s]
```

```
14711
300
```

In [73]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_resource_summary_cv = []; # the avg-w2v for each sentence/review is stor
ed in this list
for sentence in tqdm(X_cv['project_resource_summary']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_project_resource_summary_cv.append(vector)

print(len(tfidf_w2v_vectors_project_resource_summary_cv))
print(len(tfidf_w2v_vectors_project_resource_summary_cv[0]))
```

```
100%|██████████| 7247/7247 [00:00<00:00, 12814.14it/s]
```

```
7247
300
```

In [74]:

```
# Similarly you can vectorize for title also
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_resource_summary_test = []; # the avg-w2v for each sentence/review is st
ored in this list
for sentence in tqdm(X_test['project_resource_summary']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =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) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors_project_resource_summary_test.append(vector)

print(len(tfidf_w2v_vectors_project_resource_summary_test))
print(len(tfidf_w2v_vectors_project_resource_summary_test[0]))
```

```
100%|██████████| 10816/10816 [00:00<00:00, 13064.33it/s]
```

```
10816
300
```

1.5.3 Vectorizing Numerical features

In [75]:

```
# You no need to perform standardization/normalization on numerical data,
# because you will classify data by using gini impurity in decision tree classifier.
# - quantity : numerical (optional)

X_train_quantity_norm = X_train['quantity'].values.reshape(-1,1)
X_cv_quantity_norm = X_cv['quantity'].values.reshape(-1,1)
X_test_quantity_norm = X_test['quantity'].values.reshape(-1,1)

print("After vectorizations")
print(X_train_quantity_norm.shape, y_train.shape)
print(X_cv_quantity_norm.shape, y_cv.shape)
print(X_test_quantity_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
====================================================================================================
```

In [76]:

```
# You no need to perform standardization/normalization on numerical data,
# because you will classify data by using gini impurity in decision tree classifier.
# One hot encoding of numerical feature
# - teacher_number_of_previously_posted_projects : numerical
X_train_teacher_number_of_previously_posted_projects_norm =
X_train['teacher_number_of_previously_posted_projects'].values.reshape(-1,1)
X_cv_teacher_number_of_previously_posted_projects_norm =
X_cv['teacher_number_of_previously_posted_projects'].values.reshape(-1,1)
X_test_teacher_number_of_previously_posted_projects_norm =
X_test['teacher_number_of_previously_posted_projects'].values.reshape(-1,1)

print("After vectorizations")
print(X_train_teacher_number_of_previously_posted_projects_norm.shape, y_train.shape)
print(X_cv_teacher_number_of_previously_posted_projects_norm.shape, y_cv.shape)
print(X_test_teacher_number_of_previously_posted_projects_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
====================================================================================================
```

In [77]:

```
# You no need to perform standardization/normalization on numerical data,
# because you will classify data by using gini impurity in decision tree classifier.
# - price : numerical

X_train_price_norm = X_train['price'].values.reshape(-1,1)
X_cv_price_norm = X_cv['price'].values.reshape(-1,1)
X_test_price_norm = X_test['price'].values.reshape(-1,1)

print("After vectorizations")
print(X_train_price_norm.shape, y_train.shape)
print(X_cv_price_norm.shape, y_cv.shape)
print(X_test_price_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
====================================================================================================
```

1.5.4 Merging all the above features

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

In [78]:

```
# print(categories_one_hot.shape)
# print(sub_categories_one_hot.shape)
# print(text_bow.shape)
# print(price_standardized.shape)
print('Categorical Features')
print('*'*100)
print(X_train_school_state_ohe.shape, y_train.shape)
print(X_cv_school_state_ohe.shape, y_cv.shape)
print(X_test_school_state_ohe.shape, y_test.shape)
print('*'*100)
print(X_train_clean_categories_ohe.shape, y_train.shape)
print(X_cv_clean_categories_ohe.shape, y_cv.shape)
print(X_test_clean_categories_ohe.shape, y_test.shape)
print('*'*100)
print(X_train_clean_subcategories_ohe.shape, y_train.shape)
print(X_cv_clean_subcategories_ohe.shape, y_cv.shape)
print(X_test_clean_subcategories_ohe.shape, y_test.shape)
print('*'*100)
print(X_train_project_grade_category_ohe.shape, y_train.shape)
print(X_cv_project_grade_category_ohe.shape, y_cv.shape)
print(X_test_project_grade_category_ohe.shape, y_test.shape)
print('*'*100)
print(X_train_teacher_prefix_ohe.shape, y_train.shape)
print(X_cv_clean_teacher_prefix_ohe.shape, y_cv.shape)
print(X_test_clean_teacher_prefix_ohe.shape, y_test.shape)
print('*'*100)
print('Text Encoding Features')
print('*'*100)
print(X_train_title_bow.shape, y_train.shape)
print(X_cv_title_bow.shape, y_cv.shape)
print(X_test_title_bow.shape, y_test.shape)
print('*'*100)
print(X_train_essay_bow.shape, y_train.shape)
print(X_cv_essay_bow.shape, y_cv.shape)
print(X_test_essay_bow.shape, y_test.shape)
print('*'*100)
print(X_train_project_resource_summary_bow.shape, y_train.shape)
print(X_cv_project_resource_summary_bow.shape, y_cv.shape)
print(X_test_project_resource_summary_bow.shape, y_test.shape)
print('*'*100)
print(X_train_title_tfidf.shape, y_train.shape)
print(X_cv_title_tfidf.shape, y_cv.shape)
print(X_cv_title_tfidf.shape, y_test.shape)
print('*'*100)
print(X_train_essay_tfidf.shape, y_train.shape)
print(X_cv_essay_tfidf.shape, y_cv.shape)
print(X_test_essay_tfidf.shape, y_test.shape)
print('*'*100)
print(X_train_project_resource_summary_tfidf.shape, y_train.shape)
print(X_cv_project_resource_summary_tfidf.shape, y_cv.shape)
print(X_test_project_resource_summary_tfidf.shape, y_test.shape)
print('*'*100)
print(len(avg_w2v_vectors_text_train))
print(len(avg_w2v_vectors_text_train[0]))
print('*'*100)
print(len(avg_w2v_vectors_text_cv))
print(len(avg_w2v_vectors_text_cv[0]))
print('*'*100)
print(len(avg_w2v_vectors_text_test))
print(len(avg_w2v_vectors_text_test[0]))
print('*'*100)
print(len(avg_w2v_vectors_title_train))
print(len(avg_w2v_vectors_title_train[0]))
print('*'*100)
print(len(avg_w2v_vectors_title_cv))
print(len(avg_w2v_vectors_title_cv[0]))
print('*'*100)
print(len(avg_w2v_vectors_title_test))
print(len(avg_w2v_vectors_title_test[0]))
print('*'*100)
print(len(avg_w2v_vectors_project_resource_summary_train))
print(len(avg_w2v_vectors_project_resource_summary_train[0]))
print('*'*100)
print(len(avg_w2v_vectors_project_resource_summary_cv))
print(len(avg_w2v_vectors_project_resource_summary_cv[0]))
print('*'*100)
print(len(avg_w2v_vectors_project_resource_summary_test))
print(len(avg_w2v_vectors_project_resource_summary_test[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_text_train))
print(len(tfidf_w2v_vectors_text_train[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_text_cv))
print(len(tfidf_w2v_vectors_text_cv[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_text_test))
print(len(tfidf_w2v_vectors_text_test[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_title_train))
print(len(tfidf_w2v_vectors_title_train[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_title_cv))
print(len(tfidf_w2v_vectors_title_cv[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_title_test))
print(len(tfidf_w2v_vectors_title_test[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_project_resource_summary_train))
print(len(tfidf_w2v_vectors_project_resource_summary_train[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_project_resource_summary_cv))
print(len(tfidf_w2v_vectors_project_resource_summary_cv[0]))
print('*'*100)
print(len(tfidf_w2v_vectors_project_resource_summary_test))
print(len(tfidf_w2v_vectors_project_resource_summary_test[0]))
print('*'*100)
print('Numerical Features')
print('*'*100)
print(X_train_quantity_norm.shape, y_train.shape)
print(X_cv_quantity_norm.shape, y_cv.shape)
print(X_test_quantity_norm.shape, y_test.shape)
print('*'*100)
print(X_train_teacher_number_of_previously_posted_projects_norm.shape, y_train.shape)
print(X_cv_teacher_number_of_previously_posted_projects_norm.shape, y_cv.shape)
print(X_test_teacher_number_of_previously_posted_projects_norm.shape, y_test.shape)
print('*'*100)
print(X_train_price_norm.shape, y_train.shape)
print(X_cv_price_norm.shape, y_cv.shape)
print(X_test_price_norm.shape, y_test.shape)
```

```
Categorical Features
****************************************************************************************************

(14711, 51) (14711,)
(7247, 51) (7247,)
(10816, 51) (10816,)
****************************************************************************************************

(14711, 9) (14711,)
(7247, 9) (7247,)
(10816, 9) (10816,)
****************************************************************************************************

(14711, 30) (14711,)
(7247, 30) (7247,)
(10816, 30) (10816,)
****************************************************************************************************

(14711, 4) (14711,)
(7247, 4) (7247,)
(10816, 4) (10816,)
****************************************************************************************************

(14711, 5) (14711,)
(7247, 5) (7247,)
(10816, 5) (10816,)
****************************************************************************************************

Text Encoding Features
****************************************************************************************************

(14711, 1112) (14711,)
(7247, 1112) (7247,)
(10816, 1112) (10816,)
****************************************************************************************************

(14711, 5000) (14711,)
(7247, 5000) (7247,)
(10816, 5000) (10816,)
****************************************************************************************************

(14711, 4121) (14711,)
(7247, 4121) (7247,)
(10816, 4121) (10816,)
****************************************************************************************************

(14711, 826) (14711,)
(7247, 826) (7247,)
```

```
(7247, 826) (7247,)
(7247, 826) (10816,)
****************************************************************************************************

(14711, 7310) (14711,)
(7247, 7310) (7247,)
(10816, 7310) (10816,)
****************************************************************************************************

(14711, 1886) (14711,)
(7247, 1886) (7247,)
(10816, 1886) (10816,)
****************************************************************************************************

14711
300
****************************************************************************************************

7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

14711
300
****************************************************************************************************

7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

14711
300
****************************************************************************************************

7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

14711
300
****************************************************************************************************

7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

14711
300
****************************************************************************************************

7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

14711
300
****************************************************************************************************
```

```
7247
300
****************************************************************************************************

10816
300
****************************************************************************************************

Numerical Features
****************************************************************************************************

(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
****************************************************************************************************

(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
****************************************************************************************************

(14711, 1) (14711,)
(7247, 1) (7247,)
(10816, 1) (10816,)
```

In [79]:

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

X_train_real = X_train
X_cv_real = X_cv
X_test_real = X_test

X_train = hstack((X_train_school_state_ohe, X_train_clean_categories_ohe,
X_train_clean_subcategories_ohe, X_train_project_grade_category_ohe, X_train_teacher_prefix_ohe, X
_train_title_bow, X_train_essay_bow, X_train_project_resource_summary_bow, X_train_title_tfidf,
X_train_essay_tfidf, X_train_project_resource_summary_tfidf, avg_w2v_vectors_text_train,
avg_w2v_vectors_title_train, avg_w2v_vectors_project_resource_summary_train,
tfidf_w2v_vectors_text_train, tfidf_w2v_vectors_title_train,
tfidf_w2v_vectors_project_resource_summary_train, X_train_quantity_norm,
X_train_teacher_number_of_previously_posted_projects_norm, X_train_price_norm)).tocsr()
X_cv = hstack((X_cv_school_state_ohe, X_cv_clean_categories_ohe, X_cv_clean_subcategories_ohe,
X_cv_project_grade_category_ohe, X_cv_clean_teacher_prefix_ohe, X_cv_title_bow, X_cv_essay_bow, X_c
v_project_resource_summary_bow, X_cv_title_tfidf, X_cv_essay_tfidf,
X_cv_project_resource_summary_tfidf, avg_w2v_vectors_text_cv, avg_w2v_vectors_title_cv,
avg_w2v_vectors_project_resource_summary_cv, tfidf_w2v_vectors_text_cv, tfidf_w2v_vectors_title_cv
, tfidf_w2v_vectors_project_resource_summary_cv, X_cv_quantity_norm,
X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_price_norm)).tocsr()
X_test = hstack((X_test_school_state_ohe, X_test_clean_categories_ohe,
X_test_clean_subcategories_ohe, X_test_project_grade_category_ohe, X_test_clean_teacher_prefix_ohe
,X_test_title_bow, X_test_essay_bow, X_test_project_resource_summary_bow, X_test_title_tfidf,
X_test_essay_tfidf, X_test_project_resource_summary_tfidf, avg_w2v_vectors_text_test,
avg_w2v_vectors_title_test, avg_w2v_vectors_project_resource_summary_test,
tfidf_w2v_vectors_text_test, tfidf_w2v_vectors_title_test,
tfidf_w2v_vectors_project_resource_summary_test, X_test_quantity_norm,
X_test_teacher_number_of_previously_posted_projects_norm, X_test_price_norm)).tocsr()

print(X_train_real.shape)
print(X_cv_real.shape)
print(X_test_real.shape)
print(X_train.shape)
print(X_cv.shape)
print(X_test.shape)
```

```
(14711, 17)
(7247, 17)
(10816, 17)
(14711, 22157)
(7247, 22157)
(10816, 22157)
```

Computing Sentiment Scores

In [80]:

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# import nltk
# nltk.download('vader_lexicon')

sid = SentimentIntensityAnalyzer()

for_sentiment = 'a person is a person no matter how small dr seuss i teach the smallest students w
ith the biggest enthusiasm \
for learning my students learn in many different ways using all of our senses and multiple intelli
gences i use a wide range\
of techniques to help all my students succeed students in my class come from a variety of differen
t backgrounds which makes\
for wonderful sharing of experiences and cultures including native americans our school is a carin
g community of successful \
learners which can be seen through collaborative student project based learning in and out of the
classroom kindergarteners \
in my class love to work with hands on materials and have many different opportunities to practice
a skill before it is\
mastered having the social skills to work cooperatively with friends is a crucial aspect of the ki
ndergarten curriculum\
montana is the perfect place to learn about agriculture and nutrition my students love to role pla
y in our pretend kitchen\
in the early childhood classroom i have had several kids ask me can we try cooking with real food
i will take their idea \
and create common core cooking lessons where we learn important math and writing concepts while co
oking delicious healthy \
food for snack time my students will have a grounded appreciation for the work that went into maki
ng the food and knowledge \
of where the ingredients came from as well as how it is healthy for their bodies this project woul
d expand our learning of \
nutrition and agricultural cooking recipes by having us peel our own apples to make homemade apple
sauce make our own bread \
and mix up healthy plants from our classroom garden in the spring we will also create our own cook
books to be printed and \
shared with families students will gain math and literature skills as well as a life long enjoymen
t for healthy cooking \
nannan'
ss = sid.polarity_scores(for_sentiment)

for k in ss:
    print('{0}: {1}, '.format(k, ss[k]), end='')

# we can use these 4 things as features/attributes (neg, neu, pos, compound)
# neg: 0.0, neu: 0.753, pos: 0.247, compound: 0.93
```

```
neg: 0.01, neu: 0.745, pos: 0.245, compound: 0.9975,
```

Assignment 11: TruncatedSVD

- step 1 Select the top 2k words from essay text and project_title (concatinate essay text with project title and then find the top 2k words) based on their `idf_` values
- step 2 Compute the co-occurance matrix with these 2k words, with window size=5 (ref)
- step 3 Use TruncatedSVD on calculated co-occurance matrix and reduce its dimensions, choose the number of components (`n_components`) using elbow method

> - The shape of the matrix after TruncatedSVD will be 2000*n, i.e. each row represents a vector form of the corresponding word.
> - Vectorize the essay text and project titles using these word vectors. (while vectorizing, do ignore all the words which are not in top 2k words)

- step 4 Concatenate these truncatedSVD matrix, with the matrix with features

        - **school_state** : categorical data
        - **clean_categories** : categorical data
        - **clean_subcategories** : categorical data
        - **project_grade_category** :categorical data
        - **teacher_prefix** : categorical data
        - **quantity** : numerical data
        - **teacher_number_of_previously_posted_projects** : numerical data
        - **price** : numerical data
        - **sentiment score's of each of the essay** : numerical data
        - **number of words in the title** : numerical data
        - **number of words in the combine essays** : numerical data
        - **word vectors calculated in** step 3 : numerical data
- step 5: Apply GBDT on matrix that was formed in step 4 of this assignment, **DO REFER THIS BLOG: XGBOOST DMATRIX**
- **step 6:Hyper parameter tuning (Consider any two hyper parameters)**
    - **Find the best hyper parameter which will give the maximum AUC value**
    - **Find the best hyper paramter using k-fold cross validation or simple cross validation data**
    - **Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this task of hyperparameter tuning**

In [81]:

```
import sys
import math

import numpy as np
from sklearn.model_selection import learning_curve, GridSearchCV
from sklearn.metrics import roc_auc_score

# you might need to install this one
import xgboost as xgb

class XGBoostClassifier():
    def __init__(self, num_boost_round=10, **params):
        self.clf = None
        self.num_boost_round = num_boost_round
        self.params = params
        self.params.update({'objective': 'multi:softprob'})

    def fit(self, X, y, num_boost_round=None):
        num_boost_round = num_boost_round or self.num_boost_round
        self.label2num = {label: i for i, label in enumerate(sorted(set(y)))}
        dtrain = xgb.DMatrix(X, label=[self.label2num[label] for label in y])
        self.clf = xgb.train(params=self.params, dtrain=dtrain, num_boost_round=num_boost_round, verbose_eval=1)

    def predict(self, X):
        num2label = {i: label for label, i in self.label2num.items()}
        Y = self.predict_proba(X)
        y = np.argmax(Y, axis=1)
        return np.array([num2label[i] for i in y])

    def predict_proba(self, X):
        dtest = xgb.DMatrix(X)
        return self.clf.predict(dtest)

    def score(self, X, y):
        Y = self.predict_proba(X)[:,1]
        return roc_auc_score(y, Y)

    def get_params(self, deep=True):
        return self.params

    def set_params(self, **params):
        if 'num_boost_round' in params:
            self.num_boost_round = params.pop('num_boost_round')
        if 'objective' in params:
            del params['objective']
        self.params.update(params)
        return self


clf = XGBoostClassifier(eval_metric = 'auc', num_class = 2, nthread = 4,)
###################################################################
#               Change from here                                  #
#                Change from here                              #
####################################################################
parameters = {
    'num_boost_round': [100, 250, 500],
    'eta': [0.05, 0.1, 0.3],
    'max_depth': [6, 9, 12],
    'subsample': [0.9, 1.0],
    'colsample_bytree': [0.9, 1.0],
}

clf = GridSearchCV(clf, parameters)
X = np.array([[1,2], [3,4], [2,1], [4,3], [1,0], [4,5]])
Y = np.array([0, 1, 0, 1, 0, 1])
clf.fit(X, Y)

# print(clf.grid_scores_)
# best_parameters, score, _ = max(clf.grid_scores_, key=lambda x: x[1])
# print('score:', score)
# for param_name in sorted(best_parameters.keys()):
#     print("%s: %r" % (param_name, best_parameters[param_name]))
```

Out[81]:

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
             estimator=<__main__.XGBoostClassifier object at 0x1a2d9c2240>,
             iid='warn', n_jobs=None,
             param_grid={'colsample_bytree': [0.9, 1.0],
                         'eta': [0.05, 0.1, 0.3], 'max_depth': [6, 9, 12],
                         'num_boost_round': [100, 250, 500],
                         'subsample': [0.9, 1.0]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

2. TruncatedSVD

2.1 Selecting top 2000 words from `essay` and `project_title`

In [82]:

```
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# 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

# X_train_real.columns
# Concate essay and project title
X_train_combined_essay_and_title = X_train_real['clean_essays'] + ' ' + X_train_real['clean_titles'
]
# print(X_train_combined_essay_and_title[100])
# print("*"*100)
# print(X_train_real['clean_essays'][100])
# print("*"*100)
# print(X_train_real['clean_titles'][100])


# step 1 Select the top 2k words from essay text and project_title
# (concatinate essay text with project title and then find the top 2k words)
# based on their idf_ values

tfidf_model = TfidfVectorizer(min_df=10)
tfidf_model.fit_transform(X_train_combined_essay_and_title.values)

features = tfidf_model.get_feature_names()
idf = tfidf_model.idf_

indexes = np.argsort(idf)[::-1]
indexes = indexes[0:2000]
```

```
top_features_2k = []
top_idf_2k = []
for i in indexes:
  top_features_2k.append(features[i])
  top_idf_2k.append(idf[i])

list_top2k = list(zip(top_features_2k, top_idf_2k))
print(list_top2k[0])
```

```
('zoom', 8.198523493485787)
```

2.2 Computing Co-occurance matrix

In [83]:

```
%%time
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# make sure you featurize train and test data separatly

# 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

# step 2 Compute the co-occurance matrix with these 2k words, with window size=5 (ref)


# Evaluate the Co-occurence matrix with context window '5'
def get_co_occur_matrix(data, vocab, context_window=5):
    a = pd.DataFrame(np.zeros((len(vocab), len(vocab))), index=vocab, columns=vocab)
    for review in data:
        words = review.split()
        for idx in range(len(words)):
            if a.get(words[idx]) is None:
                continue
            for i in range(1, context_window+1):
                if idx-i >= 0:
                    if a.get(words[idx-i]) is not None:
                        a[words[idx-i]].loc[words[idx]] = a.get(words[idx-i]).loc[words[idx]] + 1
                        a[words[idx]].loc[words[idx-i]] = a.get(words[idx]).loc[words[idx-i]] + 1
                if idx+i < len(words):
                    if a.get(words[idx+i]) is not None:
                        a[words[idx+i]].loc[words[idx]] = a.get(words[idx+i]).loc[words[idx]] + 1
                        a[words[idx]].loc[words[idx+i]] = a.get(words[idx]).loc[words[idx+i]] + 1
    np.fill_diagonal(a.values, 0)
    return a

co_matrix = get_co_occur_matrix(X_train_combined_essay_and_title, top_features_2k)
```

```
CPU times: user 21.5 s, sys: 55.5 ms, total: 21.6 s
Wall time: 21.7 s
```

In [84]:

```
co_matrix
```

Out[84]:

	zoom	trajectory	fatigue	illiterate	segments	minneapolis	rob	transporting	trapped	roam	...	intimidating	partake
zoom	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
trajectory	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
fatigue	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
illiterate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
segments	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

	zoom	trajectory	fatigue	illiterate	segments	minneapolis	rob	transporting	trapped	roam	...	intimidating	partake
minneapolis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
rob	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
transporting	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
trapped	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
roam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
iii	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
sorely	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
fathom	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
glued	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
treasures	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
treating	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
dreaded	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
breakoutedu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
breakdown	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
bread	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
brass	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
dribbling	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
bottoms	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
troubleshooting	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
mixer	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
mock	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
modest	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
tuners	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
riddles	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
pit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
productively	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
ergonomic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
warmth	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
answered	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
shooting	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
interfere	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
suddenly	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
creek	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
amplify	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
mail	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
rambunctious	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
expend	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
intrinsically	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
paragraph	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
cumbersome	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
expansion	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
expands	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
egypt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
sibling	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
winners	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
intimidating	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
partake	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
existent	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
accidents	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

	zoom	trajectory	fatigue	illiterate	segments	minneapolis	rob	transporting	trapped	roam	...	intimidating	partake
enticing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
ran	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
echo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
advocates	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
colorado	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
bones	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

2000 rows × 2000 columns

In [85]:

```
%%time
# Display 20 samples in Co-occurence matrix
counter = 0
for i in co_matrix.index:
    if counter >= 20:
        break
    for j in co_matrix.index:
        if co_matrix.loc[j][i] != 0:
            print (i,j,"===>", co_matrix.loc[j][i])
            counter += 1
```

```
zoom font ===> 2.0
zoom closest ===> 2.0
trajectory acceleration ===> 2.0
trajectory predict ===> 4.0
fatigue prolonged ===> 2.0
fatigue decreased ===> 2.0
fatigue argue ===> 2.0
fatigue grip ===> 2.0
illiterate poem ===> 2.0
segments astronomy ===> 2.0
segments extending ===> 2.0
minneapolis metro ===> 2.0
minneapolis excitable ===> 2.0
minneapolis plagued ===> 2.0
rob dewey ===> 20.0
transporting layers ===> 2.0
transporting sidewalk ===> 2.0
transporting accident ===> 2.0
transporting mass ===> 2.0
trapped arrange ===> 2.0
trapped understandable ===> 2.0
trapped groupings ===> 2.0
CPU times: user 1.64 s, sys: 14.3 ms, total: 1.66 s
Wall time: 1.71 s
```

2.3 Applying TruncatedSVD and Calculating Vectors for `essay` and `project_title`

In [86]:

```
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# make sure you featurize train and test data separatly

# 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


# step 3 Use TruncatedSVD on calculated co-occurance matrix and reduce its dimensions,
# choose the number of components (n_components) using elbow method

# The shape of the matrix after TruncatedSVD will be 2000*n,
# i.e. each row represents a vector form of the corresponding word
# i.e. each row represents a vector form of the corresponding word.
# Vectorize the essay text and project titles using these word vectors.
# (while vectorizing, do ignore all the words which are not in top 2k words)

# Process TruncatedSVD
# Apply TruncatedSVD on TfidfVectorizer of essay text,
# choose the number of components (`n_components`) using elbow method : numerical data
# Citation https://medium.com/swlh/truncated-singular-value-decomposition-svd-using-amazon-food-re
views-891d97af5d8d for TruncatedSVD

print(co_matrix.shape)

from sklearn.decomposition import TruncatedSVD

# Program to find the optimal number of components for Truncated SVD
n_comp = [4,10,15,20,50,100,150,200,500,700,800,900,1000,1500,1999] # list containing different val
ues of components
explained = [] # explained variance ratio for each component of Truncated SVD
for x in tqdm(n_comp):
    svd = TruncatedSVD(n_components=x, random_state=42)
    svd.fit(co_matrix)
    explained.append(svd.explained_variance_ratio_.sum())
    print("Number of components = %r and explained variance = %r"%(x,svd.explained_variance_ratio_
.sum()))


# Plot the Truncated SVD spectrum
plt.figure(1, figsize=(6, 4))

plt.plot(n_comp, explained)
plt.xlabel('Number of components')
plt.ylabel("Explained Variance")
plt.title("Plot of Number of components v/s explained variance")
plt.show()
```

```
  0%|          | 0/15 [00:00<?, ?it/s]
```

```
(2000, 2000)
Number of components = 4 and explained variance = 0.1286520064553036
```

```
 13%|█▎        | 2/15 [00:00<00:01, 11.61it/s]
```

```
Number of components = 10 and explained variance = 0.2197570277992794
```

```
 27%|██▋       | 4/15 [00:00<00:00, 11.54it/s]
```

```
Number of components = 15 and explained variance = 0.2657068397854207
Number of components = 20 and explained variance = 0.3028995537974381
```

```
 33%|███▎      | 5/15 [00:00<00:01,  9.66it/s]
```

```
Number of components = 50 and explained variance = 0.4445075789921685
```

```
 40%|████      | 6/15 [00:00<00:01,  7.03it/s]
```

```
Number of components = 100 and explained variance = 0.5629133741022715
```

```
 47%|████▋     | 7/15 [00:01<00:01,  5.09it/s]
```

```
Number of components = 150 and explained variance = 0.6398121947632999
```

```
 53%|█████▎    | 8/15 [00:01<00:01,  3.76it/s]
```

```
Number of components = 200 and explained variance = 0.6931965007885699
```

```
 60%|██████    | 9/15 [00:02<00:03,  1.92it/s]
```

```
Number of components = 500 and explained variance = 0.8665554779228808
```

```
 67%|██████▋   | 10/15 [00:04<00:04,  1.23it/s]
```

```
Number of components = 700 and explained variance = 0.9259151216097632
```

```
 73%|███████▎  | 11/15 [00:05<00:04,  1.10s/it]
```

```
Number of components = 800 and explained variance = 0.9466645708640891
```

```
 80%|████████  | 12/15 [00:07<00:04,  1.39s/it]
```

```
Number of components = 900 and explained variance = 0.962496011626655
```

```
 87%|████████▋ | 13/15 [00:10<00:03,  1.68s/it]
```

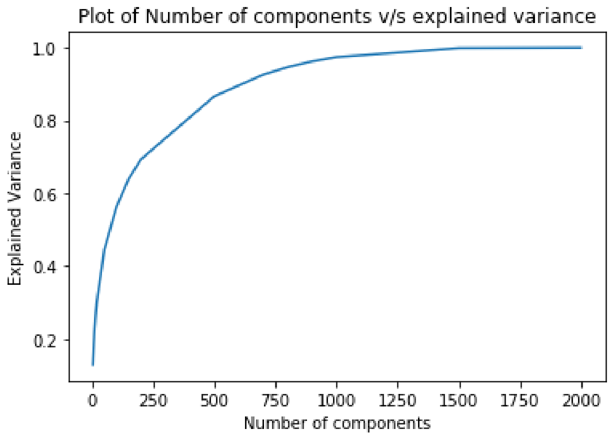
```
Number of components = 1000 and explained variance = 0.9742996619351999
```

```
 93%|█████████▎| 14/15 [00:14<00:02,  2.53s/it]
```

```
Number of components = 1500 and explained variance = 0.9990382095707607
```

```
100%|██████████| 15/15 [00:21<00:00,  3.68s/it]
```

```
Number of components = 1999 and explained variance = 0.9999999999999963
```



In [87]:

```
# Vectorize the essay text and project titles using these word vectors.
# I am taking 1000 dimension because it will cover 99.98% data
svd = TruncatedSVD(n_components=1500, random_state=42)
svd_matrix = svd.fit_transform(co_matrix)
print(svd_matrix.shape)
```

```
(2000, 1500)
```

In [88]:

```
# Vectorize the essay text and project titles using these word vectors.
# (while vectorizing, do ignore all the words which are not in top 2k words)
print(co_matrix.columns)

# collect word names
word_names = list(co_matrix.columns)

clean_essay_after_svd_train = []; # the word vectors for each sentence/review is stored in this li
st
```

```
for sentence in tqdm(X_train_real['clean_essays']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_essay_after_svd_train.append(vector)

print(len(clean_essay_after_svd_train))
print(len(clean_essay_after_svd_train[0]))
```

```
  0%|          | 25/14711 [00:00<00:59, 247.81it/s]
```

```
Index(['zoom', 'trajectory', 'fatigue', 'illiterate', 'segments',
       'minneapolis', 'rob', 'transporting', 'trapped', 'roam',
       ...
       'intimidating', 'partake', 'existent', 'accidents', 'enticing', 'ran',
       'echo', 'advocates', 'colorado', 'bones'],
      dtype='object', length=2000)
```

```
100%|██████████| 14711/14711 [00:48<00:00, 305.81it/s]
```

```
14711
1500
```

In [89]:

```
clean_essay_after_svd_cv = []; # the word vectors for each sentence/review is stored in this list
for sentence in tqdm(X_cv_real['clean_essays']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_essay_after_svd_cv.append(vector)

print(len(clean_essay_after_svd_cv))
print(len(clean_essay_after_svd_cv[0]))
```

```
100%|██████████| 7247/7247 [00:23<00:00, 303.62it/s]
```

```
7247
1500
```

In [91]:

```
clean_essay_after_svd_test = []; # the word vectors for each sentence/review is stored in this lis
t
for sentence in tqdm(X_test_real['clean_essays']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_essay_after_svd_test.append(vector)

print(len(clean_essay_after_svd_test))
print(len(clean_essay_after_svd_test[0]))
```

```
100%|██████████| 10816/10816 [00:39<00:00, 266.08it/s]
```

```
10816
1500
```

In [92]:

```
clean_title_after_svd_train = []; # the word vectors for each sentence/review is stored in this li
```

```
st
for sentence in tqdm(X_train_real['clean_titles']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_title_after_svd_train.append(vector)

print(len(clean_title_after_svd_train))
print(len(clean_title_after_svd_train[0]))
```

```
100%|██████████| 14711/14711 [00:01<00:00, 9093.40it/s]
```

```
14711
1500
```

In [93]:

```
clean_title_after_svd_cv = []; # the word vectors for each sentence/review is stored in this list
for sentence in tqdm(X_cv_real['clean_titles']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_title_after_svd_cv.append(vector)

print(len(clean_title_after_svd_cv))
print(len(clean_title_after_svd_cv[0]))
```

```
100%|██████████| 7247/7247 [00:00<00:00, 8759.93it/s]
```

```
7247
1500
```

In [94]:

```
clean_title_after_svd_test = []; # the word vectors for each sentence/review is stored in this lis
t
for sentence in tqdm(X_test_real['clean_titles']): # for each review/sentence
    vector = np.zeros(1500) # as word vectors are of zero length
    for word in sentence.split(): # for each word in a review/sentence
        if word in word_names:
            vector += svd_matrix[word_names.index(word)]
    clean_title_after_svd_test.append(vector)

print(len(clean_title_after_svd_test))
print(len(clean_title_after_svd_test[0]))
```

```
100%|██████████| 10816/10816 [00:01<00:00, 9394.83it/s]
```

```
10816
1500
```

2.4 Merge the features from step 3 and step 4

In [95]:

```
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# 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

# step 4 Concatenate these truncatedSVD matrix, with the matrix with features

#     school_state : categorical data
#     clean_categories : categorical data
#     clean_subcategories : categorical data
#     project_grade_category :categorical data
#     teacher_prefix : categorical data
#     quantity : numerical data
#     teacher_number_of_previously_posted_projects : numerical data
#     price : numerical data
#     sentiment score's of each of the essay : numerical data
#     number of words in the title : numerical data
#     number of words in the combine essays : numerical data
#     word vectors calculated in step 3 : numerical data
```

Sentiment Score

In [96]:

```
# Collect sentiment score
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sid = SentimentIntensityAnalyzer()

# Process train data
# the sentiment score for each sentence/review is stored in these lists
sentiment_score_essay_pos_train = []
sentiment_score_essay_neg_train = []
sentiment_score_essay_compound_train = [];
for sentence in tqdm(X_train_real['clean_essays']): # for each review/sentence
    ss = sid.polarity_scores(sentence)
    sentiment_score_essay_pos_train.append(ss['pos'])
    sentiment_score_essay_neg_train.append(ss['neg'])
    sentiment_score_essay_compound_train.append(ss['compound'])


sentiment_score_essay_pos_train = np.array(sentiment_score_essay_pos_train)
sentiment_score_essay_neg_train = np.array(sentiment_score_essay_neg_train)
sentiment_score_essay_compound_train = np.array(sentiment_score_essay_compound_train)

print(len(sentiment_score_essay_pos_train))
print(len(sentiment_score_essay_neg_train))
print(len(sentiment_score_essay_compound_train))

# Process cv data
# the sentiment score for each sentence/review is stored in these lists
sentiment_score_essay_pos_cv = []
sentiment_score_essay_neg_cv = []
sentiment_score_essay_compound_cv = [];
for sentence in tqdm(X_cv_real['clean_essays']): # for each review/sentence
    ss = sid.polarity_scores(sentence)
    sentiment_score_essay_pos_cv.append(ss['pos'])
    sentiment_score_essay_neg_cv.append(ss['neg'])
    sentiment_score_essay_compound_cv.append(ss['compound'])


sentiment_score_essay_pos_cv = np.array(sentiment_score_essay_pos_cv)
sentiment_score_essay_neg_cv = np.array(sentiment_score_essay_neg_cv)
sentiment_score_essay_compound_cv = np.array(sentiment_score_essay_compound_cv)

print(len(sentiment_score_essay_pos_cv))
print(len(sentiment_score_essay_neg_cv))
print(len(sentiment_score_essay_compound_cv))

# Process test data
# the sentiment score for each sentence/review is stored in these lists
sentiment_score_essay_pos_test = []
sentiment_score_essay_neg_test = []
sentiment_score_essay_compound_test = [];
sentiment_score_essay_compound_test = [];
for sentence in tqdm(X_test_real['clean_essays']): # for each review/sentence
    ss = sid.polarity_scores(sentence)
    sentiment_score_essay_pos_test.append(ss['pos'])
    sentiment_score_essay_neg_test.append(ss['neg'])
    sentiment_score_essay_compound_test.append(ss['compound'])


sentiment_score_essay_pos_test = np.array(sentiment_score_essay_pos_test)
sentiment_score_essay_neg_test = np.array(sentiment_score_essay_neg_test)
sentiment_score_essay_compound_test = np.array(sentiment_score_essay_compound_test)

print(len(sentiment_score_essay_pos_test))
print(len(sentiment_score_essay_neg_test))
print(len(sentiment_score_essay_compound_test))
```

```
100%|██████████| 14711/14711 [00:21<00:00, 688.03it/s]
  1%|          | 64/7247 [00:00<00:11, 636.27it/s]
```

```
14711
14711
14711
```

```
100%|██████████| 7247/7247 [00:10<00:00, 665.69it/s]
  1%|          | 65/10816 [00:00<00:16, 649.01it/s]
```

```
7247
7247
7247
```

```
100%|██████████| 10816/10816 [00:16<00:00, 647.80it/s]
```

```
10816
10816
10816
```

Number of words in the title for each record

In [97]:

```
title_no_of_words_train = []
for title in tqdm(X_train_real['clean_titles']):
    title_no_of_words_train.append(len(title.split()))

title_no_of_words_train = np.array(title_no_of_words_train)

title_no_of_words_cv = []
for title in tqdm(X_cv_real['clean_titles']):
    title_no_of_words_cv.append(len(title.split()))

title_no_of_words_cv = np.array(title_no_of_words_cv)

title_no_of_words_test = []
for title in tqdm(X_test_real['clean_titles']):
    title_no_of_words_test.append(len(title.split()))

title_no_of_words_test = np.array(title_no_of_words_test)


print(len(title_no_of_words_train))
print(len(title_no_of_words_cv))
print(len(title_no_of_words_test))
```

```
100%|██████████| 14711/14711 [00:00<00:00, 555912.59it/s]
100%|██████████| 7247/7247 [00:00<00:00, 496644.30it/s]
100%|██████████| 10816/10816 [00:00<00:00, 695343.37it/s]
```

```
14711
```

```
7247
10816
```

Number of words in the essay for every record

In [98]:

```
essay_no_of_words_train = []
for essay in tqdm(X_train_real['clean_essays']):
    essay_no_of_words_train.append(len(essay.split()))

essay_no_of_words_train = np.array(essay_no_of_words_train)

essay_no_of_words_cv = []
for essay in tqdm(X_cv_real['clean_essays']):
    essay_no_of_words_cv.append(len(essay.split()))

essay_no_of_words_cv = np.array(essay_no_of_words_cv)

essay_no_of_words_test = []
for essay in tqdm(X_test_real['clean_essays']):
    essay_no_of_words_test.append(len(essay.split()))

essay_no_of_words_test = np.array(essay_no_of_words_test)


print(len(essay_no_of_words_train))
print(len(essay_no_of_words_cv))
print(len(essay_no_of_words_test))
```

```
100%|██████████| 14711/14711 [00:00<00:00, 84591.75it/s]
100%|██████████| 7247/7247 [00:00<00:00, 90189.51it/s]
100%|██████████| 10816/10816 [00:00<00:00, 96470.61it/s]
```

```
14711
7247
10816
```

Finally merge all data

In [99]:

```
X_train_without_text = hstack((X_train_school_state_ohe, X_train_clean_categories_ohe,
X_train_clean_subcategories_ohe, X_train_project_grade_category_ohe, X_train_teacher_prefix_ohe, X
_train_quantity_norm, X_train_teacher_number_of_previously_posted_projects_norm,
X_train_price_norm, title_no_of_words_train.reshape(-1,1), essay_no_of_words_train.reshape(-1, 1),
sentiment_score_essay_pos_train.reshape(-1, 1), sentiment_score_essay_neg_train.reshape(-1, 1), sen
timent_score_essay_compound_train.reshape(-1, 1), clean_essay_after_svd_train,
clean_title_after_svd_train)).tocsr()
X_cv_without_text = hstack((X_cv_school_state_ohe, X_cv_clean_categories_ohe,
X_cv_clean_subcategories_ohe, X_cv_project_grade_category_ohe, X_cv_clean_teacher_prefix_ohe,
X_cv_quantity_norm, X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_price_norm,
title_no_of_words_cv.reshape(-1,1), essay_no_of_words_cv.reshape(-1, 1),
sentiment_score_essay_pos_cv.reshape(-1, 1), sentiment_score_essay_neg_cv.reshape(-1, 1), sentiment
_score_essay_compound_cv.reshape(-1, 1), clean_essay_after_svd_cv, clean_title_after_svd_cv)).tocsr
()
X_test_without_text = hstack((X_test_school_state_ohe, X_test_clean_categories_ohe,
X_test_clean_subcategories_ohe, X_test_project_grade_category_ohe, X_test_clean_teacher_prefix_ohe
, X_test_quantity_norm, X_test_teacher_number_of_previously_posted_projects_norm,
X_test_price_norm, title_no_of_words_test.reshape(-1,1), essay_no_of_words_test.reshape(-1, 1), sen
timent_score_essay_pos_test.reshape(-1, 1), sentiment_score_essay_neg_test.reshape(-1, 1), sentimen
t_score_essay_compound_test.reshape(-1, 1), clean_essay_after_svd_test, clean_title_after_svd_test
)).tocsr()

print('*'*100)
print('Shape Without Text')
print(X_train_without_text.shape)
print(X_cv_without_text.shape)
print(X_test_without_text.shape)
```

```
***************************************************************************************************

Shape Without Text
(14711, 3107)
(7247, 3107)
(10816, 3107)
```

2.5 Apply XGBoost on the Final Features from the above section

https://xgboost.readthedocs.io/en/latest/python/python_intro.html

In [100]:

```
# No need to split the data into train and test(cv)
# use the Dmatrix and apply xgboost on the whole data
# please check the Quora case study notebook as reference

# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# 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 matplotlib.pyplot as plt
from xgboost import XGBClassifier
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import learning_curve, GridSearchCV

xgb = XGBClassifier(class_weight='balanced', n_jobs=-1)
parameters = {'n_estimators':[10, 50, 100, 150, 200, 300, 500, 1000], 'max_depth': [2, 3, 4, 5, 6, 7
, 8, 9, 10]}
clf=GridSearchCV(xgb, parameters, cv=3, scoring='roc_auc', n_jobs=-1, return_train_score=True)
clf.fit(X_train_without_text, y_train)
```

Out[100]:

```
GridSearchCV(cv=3, error_score='raise-deprecating',
             estimator=XGBClassifier(base_score=0.5, booster='gbtree',
                                     class_weight='balanced',
                                     colsample_bylevel=1, colsample_bynode=1,
                                     colsample_bytree=1, gamma=0,
                                     learning_rate=0.1, max_delta_step=0,
                                     max_depth=3, min_child_weight=1,
                                     missing=None, n_estimators=100, n_jobs=-1,
                                     nthread=None, objective='binary:logistic',
                                     random_state=0, reg_alpha=0, reg_lambda=1,
                                     scale_pos_weight=1, seed=None, silent=None,
                                     subsample=1, verbosity=1),
             iid='warn', n_jobs=-1,
             param_grid={'max_depth': [2, 3, 4, 5, 6, 7, 8, 9, 10],
                         'n_estimators': [10, 50, 100, 150, 200, 300, 500,
                                          1000]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
             scoring='roc_auc', verbose=0)
```

In [101]:

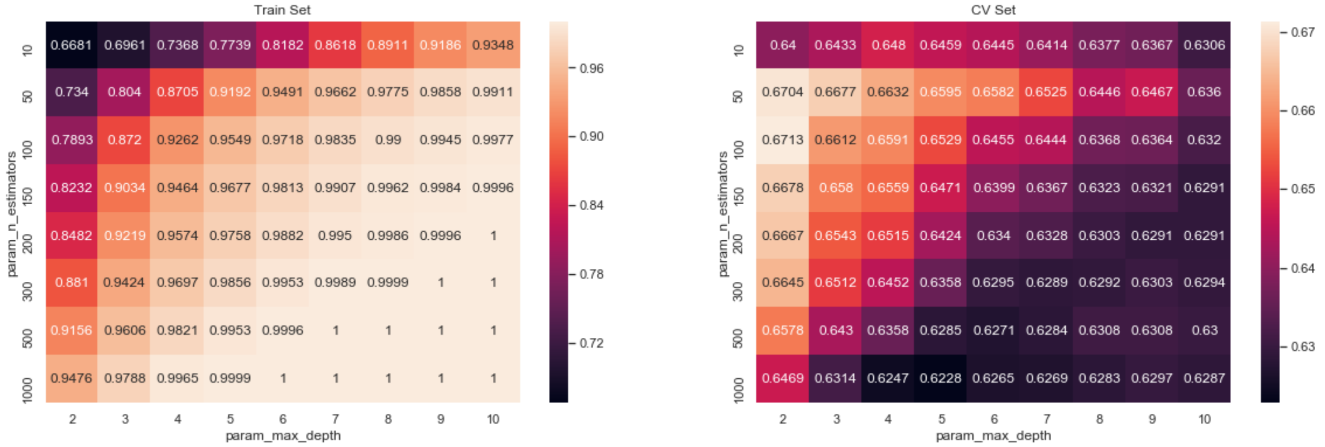
```
# print results
clf.cv_results_
```

Out[101]:

```
{'mean_fit_time': array([2.47225129e+01, 7.09822695e+03, 8.10901316e+03, 2.71076141e+02,
        3.41422669e+02, 5.04593432e+02, 8.31482410e+02, 4.80939944e+04,
        3.04812499e+01, 7.33631486e+03, 8.49803559e+03, 3.95175143e+02,
        5.96966839e+02, 8.65172963e+02, 1.46025668e+03, 2.72822486e+03,
        3.96966859e+02, 8.05172963e+02, 1.16025068e+03, 2.72622136e+03,
        4.16508087e+01, 1.79960250e+02, 3.52768455e+02, 5.21094990e+02,
        7.34189635e+02, 1.21473803e+03, 2.48597235e+03, 4.37441247e+03,
        5.61142547e+01, 2.77075224e+02, 5.52156802e+02, 7.77950274e+02,
        1.05293302e+03, 1.48683424e+03, 2.84014276e+03, 5.57826579e+03,
        7.09665343e+01, 3.64566791e+02, 6.65585373e+02, 1.04044541e+03,
        1.41062237e+03, 1.75227390e+03, 3.59462446e+03, 6.17494752e+04,
        7.73597300e+01, 1.89543728e+04, 6.87129194e+02, 1.00297629e+03,
        1.33431616e+03, 2.36419045e+03, 4.32039944e+03, 7.74229172e+03,
        9.06419886e+01, 4.43677735e+02, 8.79977302e+02, 1.55278115e+03,
        1.70805611e+03, 2.82697391e+03, 4.39352599e+03, 8.31817578e+03,
        9.28330447e+01, 4.65301695e+02, 1.20137279e+03, 1.41404368e+03,
        2.10537759e+03, 2.98031224e+03, 5.20778230e+04, 2.25895785e+04,
        8.97199145e+01, 4.32099843e+02, 8.64825135e+02, 1.31759256e+03,
        1.89744261e+03, 2.55431277e+03, 4.81013427e+03, 5.95984631e+03]),
 'std_fit_time': array([1.34815880e-01, 4.14119012e+03, 6.55813090e+03, 4.12473957e+00,
        4.16214510e+00, 1.23541715e+00, 3.29579490e+00, 2.02188528e+01,
        8.92652689e-01, 1.01900864e+04, 1.16636170e+04, 2.12138890e+01,
        6.86762336e+00, 1.26796539e+01, 6.62008426e+00, 5.68222910e+01,
        1.40508477e-01, 1.10189536e+00, 2.01020118e+00, 8.49792797e+00,
        1.59704942e+01, 5.11184473e+01, 5.89268719e+00, 1.92935075e+02,
        2.51496986e+00, 1.90024075e+01, 5.59560603e+01, 4.04903876e+01,
        2.34606715e+01, 1.64833849e+00, 4.11489664e+01, 3.10196297e+01,
        4.00918425e+00, 1.04701562e+01, 8.59524583e+00, 2.25106152e+01,
        4.32332192e+01, 8.35748756e+01, 1.16900237e+02, 3.40637399e+02,
        5.85586470e-01, 2.63470739e+04, 7.32319893e+01, 5.27412610e+01,
        3.85080388e+01, 8.18639647e+01, 1.25921177e+02, 3.40251579e+01,
        6.36048244e+00, 4.38022288e+01, 1.24419048e+02, 1.25459958e+01,
        6.01824730e+01, 1.17745655e+02, 1.87545910e+02, 1.04641019e+02,
        1.27480838e+00, 2.39158327e+01, 5.71279669e+02, 4.27929813e+01,
        3.78071055e+01, 1.32516969e+02, 1.10338629e+02, 2.25465389e+04,
        9.85796346e-01, 1.02067292e+01, 3.83635032e+00, 5.41942007e+01,
        3.82766368e+01, 3.92587743e+01, 2.46618846e+01, 8.74982738e+02]),
 'mean_score_time': array([2.73436403, 2.68680199, 3.19599454, 2.643116  , 2.5202477 ,
        2.6576906 , 2.60612424, 3.40695977, 2.93901873, 3.21212395,
        2.92077096, 3.11125135, 3.95631893, 3.23412172, 3.56301983,
        3.45072206, 3.23443429, 3.21201412, 3.02552017, 2.89976676,
        3.61083746, 4.645118  , 4.07133428, 5.35868549, 4.05473399,
        3.54159538, 3.62304099, 4.07736731, 3.73592202, 4.74508214,
        4.13507001, 5.0699273 , 4.37523548, 3.82637636, 4.11169672,
        4.11309425, 4.33225926, 3.94714928, 4.16328716, 5.42662811,
        3.63807901, 3.37793326, 3.45090628, 3.44841194, 5.22352036,
        5.21739666, 4.76603643, 5.37799128, 3.5390981 , 3.3945996 ,
        3.766524  , 4.94775899, 4.89610203, 6.98345439, 4.72444924,
        5.25086705, 3.94762142, 3.44754569, 3.41840092, 4.66393399,
        5.30726266, 4.02388   , 3.82061362, 4.4037443 , 3.2766037 ,
        3.91344158, 3.2403415 , 3.82611458, 3.5286653 , 4.39178165,
        4.34848364, 3.03319367]),
 'std_score_time': array([0.04283019, 0.18752814, 0.42952236, 0.12748791, 0.04930236,
        0.03741622, 0.03654883, 0.33968321, 0.45195507, 0.75283007,
        0.48245945, 0.34223273, 1.32089242, 0.21886033, 0.16379486,
        0.29317658, 0.28278008, 0.26276546, 0.02265397, 0.06385464,
        0.4782119 , 0.69467154, 0.73120188, 1.06867453, 0.67072384,
        0.22294655, 0.21968302, 0.50639623, 0.23247889, 0.72587858,
        0.1702641 , 0.38055831, 1.10725821, 0.30884253, 0.27007554,
        0.83411582, 0.65442813, 0.60237138, 0.35678945, 0.23571061,
        0.25421399, 0.45932624, 0.60817344, 0.31858481, 2.05084123,
        0.899683  , 0.63668001, 1.02497126, 0.17551256, 0.17862323,
        0.29398484, 1.02187057, 1.46617672, 3.40814464, 0.69859128,
        0.68402816, 0.51307841, 0.12406895, 0.55784392, 0.60375816,
        0.39508324, 0.23311214, 0.06694128, 0.24647399, 0.19303401,
        0.34518051, 0.06767634, 0.63668953, 0.06366813, 1.05983827,
        0.47506945, 0.46192287]),
 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4,
                    4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6,
                    6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8,
                    8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10,
                    10, 10],
              mask=[False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False],
        fill_value='?'
        fill_value='?',
             dtype=object),
 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 1000, 10, 50, 100,
                    150, 200, 300, 500, 1000, 10, 50, 100, 150, 200, 300,
                    500, 1000, 10, 50, 100, 150, 200, 300, 500, 1000, 10,
                    50, 100, 150, 200, 300, 500, 1000, 10, 50, 100, 150,
                    200, 300, 500, 1000, 10, 50, 100, 150, 200, 300, 500,
                    1000, 10, 50, 100, 150, 200, 300, 500, 1000, 10, 50,
                    100, 150, 200, 300, 500, 1000],
              mask=[False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False,
                    False, False, False, False, False, False, False, False],
        fill_value='?',
             dtype=object),
 'params': [{'max_depth': 2, 'n_estimators': 10},
  {'max_depth': 2, 'n_estimators': 50},
  {'max_depth': 2, 'n_estimators': 100},
  {'max_depth': 2, 'n_estimators': 150},
  {'max_depth': 2, 'n_estimators': 200},
  {'max_depth': 2, 'n_estimators': 300},
  {'max_depth': 2, 'n_estimators': 500},
  {'max_depth': 2, 'n_estimators': 1000},
  {'max_depth': 3, 'n_estimators': 10},
  {'max_depth': 3, 'n_estimators': 50},
  {'max_depth': 3, 'n_estimators': 100},
  {'max_depth': 3, 'n_estimators': 150},
  {'max_depth': 3, 'n_estimators': 200},
  {'max_depth': 3, 'n_estimators': 300},
  {'max_depth': 3, 'n_estimators': 500},
  {'max_depth': 3, 'n_estimators': 1000},
  {'max_depth': 4, 'n_estimators': 10},
  {'max_depth': 4, 'n_estimators': 50},
  {'max_depth': 4, 'n_estimators': 100},
  {'max_depth': 4, 'n_estimators': 150},
  {'max_depth': 4, 'n_estimators': 200},
  {'max_depth': 4, 'n_estimators': 300},
  {'max_depth': 4, 'n_estimators': 500},
  {'max_depth': 4, 'n_estimators': 1000},
  {'max_depth': 5, 'n_estimators': 10},
  {'max_depth': 5, 'n_estimators': 50},
  {'max_depth': 5, 'n_estimators': 100},
  {'max_depth': 5, 'n_estimators': 150},
  {'max_depth': 5, 'n_estimators': 200},
  {'max_depth': 5, 'n_estimators': 300},
  {'max_depth': 5, 'n_estimators': 500},
  {'max_depth': 5, 'n_estimators': 1000},
  {'max_depth': 6, 'n_estimators': 10},
  {'max_depth': 6, 'n_estimators': 50},
  {'max_depth': 6, 'n_estimators': 100},
  {'max_depth': 6, 'n_estimators': 150},
  {'max_depth': 6, 'n_estimators': 200},
  {'max_depth': 6, 'n_estimators': 300},
  {'max_depth': 6, 'n_estimators': 500},
  {'max_depth': 6, 'n_estimators': 1000},
  {'max_depth': 7, 'n_estimators': 10},
  {'max_depth': 7, 'n_estimators': 50},
  {'max_depth': 7, 'n_estimators': 100},
  {'max_depth': 7, 'n_estimators': 150},
  {'max_depth': 7, 'n_estimators': 200},
  {'max_depth': 7, 'n_estimators': 300},
  {'max_depth': 7, 'n_estimators': 500},
  {'max_depth': 7, 'n_estimators': 1000},
  {'max_depth': 8, 'n_estimators': 10},
  {'max_depth': 8, 'n_estimators': 50},
  {'max_depth': 8, 'n_estimators': 100},
  {'max_depth': 8, 'n_estimators': 150},
  {'max_depth': 8, 'n_estimators': 200},
  {'max_depth': 8, 'n_estimators': 300},
  {'max_depth': 8, 'n_estimators': 500},
  {'max_depth': 8, 'n_estimators': 1000},
  {'max_depth': 9, 'n_estimators': 10},
  {'max_depth': 9, 'n_estimators': 50},
  {'max_depth': 9, 'n_estimators': 50},
  {'max_depth': 9, 'n_estimators': 100},
  {'max_depth': 9, 'n_estimators': 150},
  {'max_depth': 9, 'n_estimators': 200},
  {'max_depth': 9, 'n_estimators': 300},
  {'max_depth': 9, 'n_estimators': 500},
  {'max_depth': 9, 'n_estimators': 1000},
  {'max_depth': 10, 'n_estimators': 10},
  {'max_depth': 10, 'n_estimators': 50},
  {'max_depth': 10, 'n_estimators': 100},
  {'max_depth': 10, 'n_estimators': 150},
  {'max_depth': 10, 'n_estimators': 200},
  {'max_depth': 10, 'n_estimators': 300},
  {'max_depth': 10, 'n_estimators': 500},
  {'max_depth': 10, 'n_estimators': 1000}],
 'split0_test_score': array([0.633449  , 0.66682639, 0.6688381 , 0.66676024, 0.66467441,
        0.66277206, 0.65520011, 0.64536348, 0.64255778, 0.66598951,
        0.65859466, 0.65743817, 0.65532238, 0.65518162, 0.64638305,
        0.63432286, 0.64927053, 0.6626855 , 0.6582148 , 0.659546  ,
        0.65528093, 0.64645924, 0.63439826, 0.62087441, 0.6443063 ,
        0.65988936, 0.65715953, 0.65239138, 0.65110593, 0.64212435,
        0.62945699, 0.62251215, 0.64210746, 0.65746304, 0.64259237,
        0.63522988, 0.6300731 , 0.62612078, 0.62463161, 0.62230349,
        0.64597672, 0.65209074, 0.64304429, 0.63392212, 0.62858106,
        0.62439043, 0.62408708, 0.62310402, 0.64385263, 0.64940012,
        0.64270332, 0.63451861, 0.63361606, 0.63107528, 0.63244761,
        0.63090918, 0.63352201, 0.64475216, 0.63280675, 0.62855348,
        0.62566823, 0.62544363, 0.62441626, 0.62227734, 0.62560447,
        0.64569106, 0.64003868, 0.63361239, 0.6315535 , 0.63233682,
        0.63268512, 0.62872309]),
 'split1_test_score': array([0.64068122, 0.6655147 , 0.66422005, 0.66005671, 0.66093992,
        0.65760376, 0.65481633, 0.64777296, 0.64331538, 0.6655891 ,
        0.65716199, 0.65511201, 0.65139842, 0.64940784, 0.6429255 ,
        0.63659595, 0.6473478 , 0.66046495, 0.6522521 , 0.64963551,
        0.64605108, 0.6424312 , 0.63657248, 0.63227008, 0.64731938,
        0.65796011, 0.64089596, 0.63780215, 0.63413661, 0.63247093,
        0.63000632, 0.62609683, 0.64858465, 0.65386462, 0.64230683,
        0.64139648, 0.63815147, 0.63536006, 0.6339776 , 0.63356313,
        0.64672658, 0.65131429, 0.6453628 , 0.6408776 , 0.63937795,
        0.63565111, 0.63614907, 0.63147771, 0.63600363, 0.63920568,
        0.63637403, 0.63626035, 0.63434225, 0.63093296, 0.6315114 ,
        0.62835532, 0.63927146, 0.64871541, 0.64132335, 0.63507395,
        0.63104584, 0.63333498, 0.63345584, 0.63165445, 0.64124528,
        0.63350917, 0.63278305, 0.63389889, 0.63353471, 0.63113668,
        0.63073579, 0.62848464]),
 'split2_test_score': array([0.64590962, 0.67879198, 0.68079837, 0.67647027, 0.67438022,
        0.67309132, 0.66331289, 0.64752406, 0.64415006, 0.67147225,
        0.6677312 , 0.66139255, 0.65630496, 0.6490542 , 0.63970604,
        0.62321345, 0.64751895, 0.66644948, 0.66691631, 0.65858788,
        0.65325219, 0.64685287, 0.63651149, 0.62082916, 0.64607918,
        0.66070954, 0.66055627, 0.65118656, 0.64197714, 0.63271696,
        0.62592361, 0.61979507, 0.6429362 , 0.66322316, 0.65153732,
        0.64310878, 0.63362747, 0.62696616, 0.62271852, 0.62364293,
        0.63136419, 0.65420437, 0.64491481, 0.63530481, 0.63036108,
        0.62651785, 0.62495051, 0.62600998, 0.63331934, 0.64518143,
        0.63134168, 0.62618241, 0.62306657, 0.62555816, 0.62835516,
        0.62570711, 0.63741067, 0.64669561, 0.63506852, 0.63257151,
        0.630469  , 0.63226258, 0.63463777, 0.63523281, 0.6250851 ,
        0.62881002, 0.62325081, 0.61985605, 0.62228601, 0.62461108,
        0.62665419, 0.62885712]),
 'mean_test_score': array([0.64001239, 0.67037721, 0.67128517, 0.66776227, 0.66666458,
        0.66448881, 0.65777609, 0.64688663, 0.64334096, 0.66768339,
        0.66116227, 0.65798084, 0.65434205, 0.65121509, 0.64300532,
        0.63137782, 0.64804593, 0.66319991, 0.65912761, 0.65592362,
        0.65152857, 0.64524794, 0.63582722, 0.62465737, 0.6459014 ,
        0.65951972, 0.65287117, 0.64712741, 0.64240774, 0.63577161,
        0.62846244, 0.62280131, 0.64454244, 0.65818351, 0.64547845,
        0.63991108, 0.63395015, 0.62948188, 0.62710891, 0.62650261,
        0.64135646, 0.6525364 , 0.64444044, 0.63670113, 0.63277279,
        0.62885252, 0.62839497, 0.62686339, 0.63772603, 0.6445964 ,
        0.63680714, 0.63232076, 0.63034207, 0.62918906, 0.63077162,
        0.62832422, 0.63673428, 0.64672079, 0.63639905, 0.63206584,
        0.62906056, 0.6303464 , 0.63083575, 0.62972052, 0.63064426,
        0.63600473, 0.63202527, 0.62912305, 0.62912507, 0.62936193,
        0.6300254 , 0.62868829]),
 'std_test_score': array([0.00510913, 0.00597359, 0.00698541, 0.00673773, 0.00566427,
        0.00643787, 0.00391784, 0.00108203, 0.00065032, 0.00268384,
        0.00468115, 0.00259241, 0.00211956, 0.00280904, 0.00272655,
        0.00468115, 0.00259241, 0.00211956, 0.00280904, 0.00272655,
        0.00584661, 0.00086891, 0.00246995, 0.00602094, 0.00446309,
        0.00396036, 0.001998  , 0.00101094, 0.00538248, 0.00123653,
        0.00115242, 0.00857971, 0.00661161, 0.00693461, 0.00449411,
        0.001809  , 0.00258063, 0.00287794, 0.00385418, 0.00428542,
        0.00338376, 0.00330598, 0.00417038, 0.00491881, 0.00502189,
        0.00707151, 0.00122117, 0.00100423, 0.00300638, 0.00472628,
        0.00488467, 0.00549374, 0.0034715 , 0.00446945, 0.00418251,
        0.00464862, 0.00439789, 0.00515257, 0.00256783, 0.00175075,
        0.00212392, 0.00239551, 0.00161814, 0.00360202, 0.00268596,
        0.00241076, 0.00349501, 0.00456576, 0.00546311, 0.00749829,
        0.00711417, 0.00687476, 0.00655313, 0.00490264, 0.00339457,
        0.00251291, 0.00015403]),
 'rank_test_score': array([34,  2,  1,  3,  5,  6, 13, 22, 30,  4,  8, 12, 15, 19, 31, 49, 20,
         7, 10, 14, 18, 26, 42, 71, 24,  9, 16, 21, 32, 43, 65, 72, 28, 11,
        25, 35, 44, 57, 68, 70, 33, 17, 29, 39, 45, 63, 66, 69, 36, 27, 37,
        46, 54, 59, 51, 67, 38, 23, 40, 47, 62, 53, 50, 56, 52, 41, 48, 61,
        60, 58, 55, 64], dtype=int32),
 'split0_train_score': array([0.66545814, 0.73229016, 0.78687275, 0.82033267, 0.84520794,
        0.87925191, 0.91520804, 0.94844424, 0.6969778 , 0.8049791 ,
        0.87449573, 0.90575503, 0.92310569, 0.94268478, 0.96168232,
        0.97995052, 0.74292446, 0.87409324, 0.92854056, 0.94715877,
        0.95849532, 0.97018662, 0.98215178, 0.99768627, 0.78098597,
        0.92447653, 0.95754768, 0.96904684, 0.97681425, 0.98594168,
        0.99618551, 0.9999632 , 0.82447234, 0.95381461, 0.97426594,
        0.98224637, 0.98873338, 0.99623428, 0.99979788, 1.        ,
        0.8694174 , 0.97022147, 0.98357463, 0.99118377, 0.99525799,
        0.99887563, 0.99999792, 1.        , 0.90021118, 0.98027838,
        0.99067747, 0.99711606, 0.9992038 , 0.99995498, 1.        ,
        1.        , 0.92315219, 0.98765838, 0.99474541, 0.99850618,
        0.99970136, 0.99999737, 1.        , 1.        , 0.9423679 ,
        0.99228608, 0.99771477, 0.99956669, 0.99998132, 1.        ,
        1.        , 1.        ]),
 'split1_train_score': array([0.66717237, 0.73541512, 0.79183449, 0.82740131, 0.85134038,
        0.8834154 , 0.91707878, 0.94811649, 0.6948898 , 0.80468721,
        0.87121779, 0.90139699, 0.92341837, 0.94388713, 0.96082714,
        0.97796625, 0.73542529, 0.86953702, 0.92554883, 0.94714152,
        0.95841675, 0.97090023, 0.98177098, 0.99565358, 0.77128382,
        0.91843715, 0.95648515, 0.96941787, 0.97697867, 0.98588017,
        0.99502284, 0.99995334, 0.81606279, 0.94874133, 0.9717394 ,
        0.98174107, 0.98809948, 0.99492477, 0.9994576 , 0.99999992,
        0.85973617, 0.96562535, 0.98422001, 0.99036426, 0.99500868,
        0.99864463, 0.99996538, 1.        , 0.88896775, 0.97729291,
        0.99064169, 0.99500677, 0.99815415, 0.99991561, 1.        ,
        1.        , 0.92180113, 0.98702157, 0.99466481, 0.99856079,
        0.99957054, 0.99999011, 1.        , 1.        , 0.93601337,
        0.99166356, 0.9977101 , 0.99970383, 0.99997272, 1.        ,
        1.        , 1.        ]),
 'split2_train_score': array([0.67157996, 0.73423571, 0.78921819, 0.8219884 , 0.84810454,
        0.88030395, 0.91461124, 0.94631937, 0.69647498, 0.80239184,
        0.87020749, 0.90298934, 0.91924653, 0.94058018, 0.95924735,
        0.97836607, 0.73212579, 0.8679303 , 0.92457429, 0.94494793,
        0.95526474, 0.96803658, 0.98233141, 0.99627599, 0.76944088,
        0.91457144, 0.95071883, 0.96473983, 0.97363543, 0.98509672,
        0.99463538, 0.99990404, 0.81402071, 0.9446029 , 0.96951766,
        0.98000138, 0.98765742, 0.9947492 , 0.99953513, 1.        ,
        0.85626274, 0.96264192, 0.98258084, 0.99042887, 0.99487715,
        0.99907815, 0.99998157, 1.        , 0.88413125, 0.97504081,
        0.98853907, 0.99634395, 0.99857212, 0.99993523, 1.        ,
        1.        , 0.91090004, 0.98284753, 0.99410709, 0.99802078,
        0.99951638, 0.99999154, 1.        , 1.        , 0.92604002,
        0.98934582, 0.99770363, 0.99953688, 0.99997671, 1.        ,
        1.        , 1.        ]),
 'mean_train_score': array([0.66807016, 0.73398033, 0.78930848, 0.82324079, 0.84821762,
        0.88099042, 0.91563269, 0.9476267 , 0.69611419, 0.80401938,
        0.87197367, 0.90338045, 0.92192353, 0.94238403, 0.9605856 ,
        0.97876095, 0.73682518, 0.87052019, 0.92622123, 0.94641607,
        0.95739227, 0.96970781, 0.98208472, 0.99653862, 0.77390355,
        0.9191617 , 0.95491722, 0.96773485, 0.97580945, 0.98563952,
        0.99528124, 0.99994019, 0.81818528, 0.94905295, 0.971841  ,
        0.98132961, 0.98816343, 0.99530275, 0.99959687, 0.99999997,
        0.86180544, 0.96616291, 0.98345849, 0.99065897, 0.99504794,
        0.99886613, 0.99998163, 1.        , 0.89110339, 0.97753737,
        0.98995274, 0.99615559, 0.99864336, 0.99993527, 1.        ,
        1.        , 0.91861778, 0.98584249, 0.99450577, 0.99836259,
        0.9995961 , 0.99999301, 1.        , 1.        , 0.9348071 ,
        0.99109849, 0.9977095 , 0.99960247, 0.99997692, 1.        ,
        1.        , 1.        ]),
        1.        , 1.        ]]),
 'std_train_score': array([2.57859072e-03, 1.28847777e-03, 2.02662650e-03, 3.01858905e-03,
        2.50483712e-03, 1.76768925e-03, 1.05116916e-03, 9.34051706e-04,
        8.89778912e-04, 1.15700074e-03, 1.83043948e-03, 1.80052850e-03,
        1.89722662e-03, 1.36670475e-03, 1.00863979e-03, 8.56846996e-04,
        4.51830133e-03, 2.61028660e-03, 1.68758520e-03, 1.03815750e-03,
        1.50473127e-03, 1.21712175e-03, 2.33656550e-04, 8.50364974e-04,
        5.06422469e-03, 4.07606463e-03, 3.00023261e-03, 2.12320603e-03,
        1.53873052e-03, 3.84642981e-04, 6.58687030e-04, 2.58775417e-05,
        4.52311500e-03, 3.76711464e-03, 1.93980945e-03, 9.61584372e-04,
        4.41578256e-04, 6.62579468e-04, 1.45617026e-04, 3.76017613e-08,
        5.56612751e-03, 3.11759511e-03, 6.74211603e-04, 3.72029500e-04,
        1.57935732e-04, 1.77112964e-04, 1.32855342e-05, 0.00000000e+00,
        6.73606239e-03, 2.14520120e-03, 9.99722737e-04, 8.71354923e-04,
        4.31467326e-04, 1.60723042e-05, 0.00000000e+00, 0.00000000e+00,
        5.48507498e-03, 2.13365619e-03, 2.83823077e-04, 2.42718032e-04,
        7.76499879e-05, 3.13761042e-06, 0.00000000e+00, 0.00000000e+00,
        6.72017938e-03, 1.26511242e-03, 4.56506567e-06, 7.26996633e-05,
        3.51402969e-06, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00])}
```

In [102]:

```
# Find best hyper parameter max_depth and min_samples_split
import seaborn as sns; sns.set()
max_scores = pd.DataFrame(clf.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max(
).unstack()[['mean_test_score', 'mean_train_score']]
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [103]:

```
# Print params
print(clf.best_estimator_)
print(clf.score(X_train_without_text, y_train))
print(clf.score(X_test_without_text, y_test))
```

```
XGBClassifier(base_score=0.5, booster='gbtree', class_weight='balanced',
              colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
              gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=2,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=-1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
0.7563664058308456
0.6771070859606634
```

In [104]:

```
n_estimators = 100
max_depth = 2
```

In [106]:

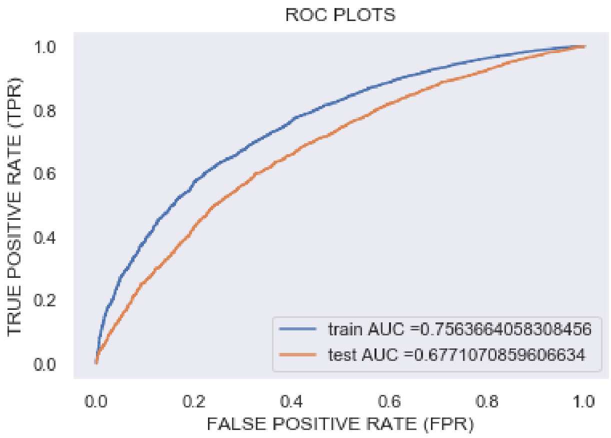
```
%%time

# Create ROC Plot for Test Set
parameters = {'max_depth':[max_depth], 'n_estimators':[n_estimators]}
xgb=GridSearchCV(XGBClassifier(class_weight='balanced', n_estimators=n_estimators,
max_depth=max_depth, n_jobs=-1), parameters, cv=3, scoring='roc_auc', n_jobs=-1, return_train_score
=True)
xgb.fit(X_train_without_text, y_train);

y_train_pred = clf.predict_proba(X_train_without_text)[:,1]
y_test_pred = clf.predict_proba(X_test_without_text)[:,1]

train_fpr, train_tpr, tr_thresholds = roc_curve(y_train, y_train_pred)
test_fpr, test_tpr, te_thresholds = roc_curve(y_test, y_test_pred)

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("FALSE POSITIVE RATE (FPR)")
plt.ylabel("TRUE POSITIVE RATE (TPR)")
plt.title("ROC PLOTS")
plt.grid()
plt.show()
```



```
CPU times: user 2min 11s, sys: 2.09 s, total: 2min 13s
Wall time: 4min 45s
```

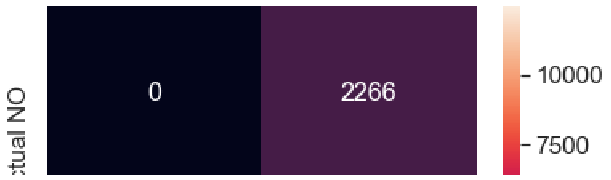
In [109]:

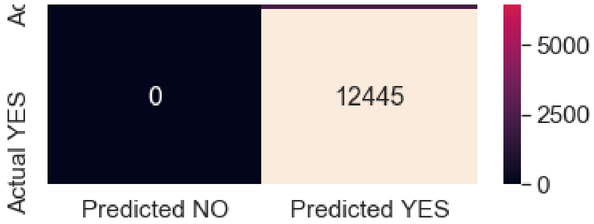
```
#function to get heatmap confusion matrix
def get_confusion_matrix(clf,X_te,y_test):
    y_pred = clf.predict(X_te)
    df_cm = pd.DataFrame(confusion_matrix(y_test, y_pred), range(2),range(2))
    df_cm.columns = ['Predicted NO','Predicted YES']
    df_cm = df_cm.rename({0: 'Actual NO', 1: 'Actual YES'})
    sns.set(font_scale=1.4)#for label size
    sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
```

In [110]:

```
%%time
get_confusion_matrix(xgb,X_train_without_text,y_train)
```

```
CPU times: user 3.77 s, sys: 303 ms, total: 4.07 s
Wall time: 4.18 s
```

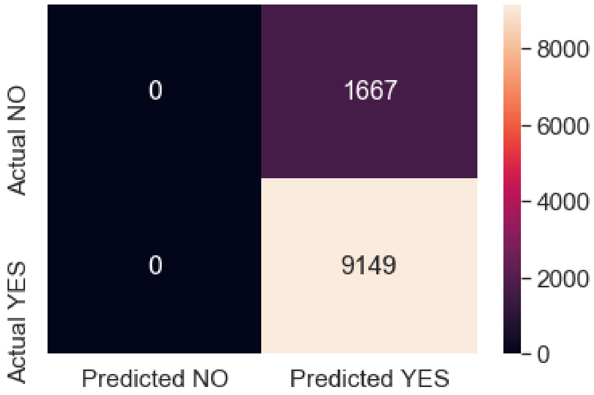




In [111]:

```
%%time
get_confusion_matrix(xgb,X_test_without_text,y_test)
```

```
CPU times: user 2.71 s, sys: 163 ms, total: 2.87 s
Wall time: 2.91 s
```



3. Conclusion

In [112]:

```
# Please write down few lines about what you observed from this assignment.
# Please compare all your models using Prettytable library
# Please compare all your models using Prettytable library
#http://zetcode.com/python/prettytable/
from prettytable import PrettyTable

x = PrettyTable()
x.field_names = ["Vectorizer", "Model", "N Estimators", "Max Depth", "AUC"]
x.add_row(["Bag of Words", "Random Forest", 100, 2, 0.68])


print(x)
```

```
+--------------+---------------+--------------+-----------+------+
|  Vectorizer  |     Model     | N Estimators | Max Depth | AUC  |
+--------------+---------------+--------------+-----------+------+
| Bag of Words | Random Forest |     100      |     2     | 0.68 |
+--------------+---------------+--------------+-----------+------+
```

In [ ]: