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 **Example:** `p036502`

Feature	**Description**
`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 follow enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`
`project_subject_categories`	One or more (comma-separated) su categories for the project from the fo 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, Ma & Science`

Feature	Description
`school_state`	State where school is located (Two-l U.S. postal code). **Example:** `WY`
`project_subject_subcategories`	One or more (comma-separated) su subcategories for the project. **Examp** • `Literacy` • `Literature & Writing, Social Sciences`
`project_resource_summary`	An explanation of the resources nee the project. **Example:** • `My students need hands literacy materials to manage sensory needs!`
`project_essay_1`	First application essay*
`project_essay_2`	Second application essay*
`project_essay_3`	Third application essay*
`project_essay_4`	Fourth application essay*
`project_submitted_datetime`	Datetime when project application w submitted. **Example:** `2016-04-28 12:43:56.245`
`teacher_id`	A unique identifier for the teacher of proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c`

Feature	Description
`teacher_prefix`	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
`teacher_number_of_previously_posted_projects`	Number of project applications previ submitted by the same teacher. **Exa** 2

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

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

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

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

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

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

Notes on the Essay Data

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

- __project_essay_1:__ "Introduce us to your classroom"
- __project_essay_2:__ "Tell us more about your students"
- __project_essay_3:__ "Describe how your students will use the materials you're requesting"
- __project_essay_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 [49]:

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

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

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

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

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

from tqdm import tqdm
import os

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

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.
```

1.1 Reading Data

In [50]:

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

In [51]:

```
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 (32414, 17)
--------------------------------------------------
The attributes of data : ['Unnamed: 0' 'id' 'teacher_id' 'teacher_prefi
x' '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 [52]:

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

```
Number of data points in train data (32414, 4)
['id' 'description' 'quantity' 'price']
```

Out[52]:

	id	description	quantity	price
0	p253737	LC652 - Lakeshore Double-Space Mobile Drying Rack	1	149.00
1	p258326	Bouncy Bands for Desks (Blue support pipes)	3	14.95

1.2 preprocessing of `project_subject_categories`

In [53]:

```
catogories = list(project_data['project_subject_categories'].values)
```

```
# remove special characters from list of strings python: https://stacko
verflow.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 & H
unger"
    for j in i.split(','): # it will split it in three parts ["Math & S
cience", "Warmth", "Care & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory b
ased on space "Math & Science"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are g
oing 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 int
o
    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 [54]:

```python
sub_catogories = list(project_data['project_subject_subcategories'].values)
# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039

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

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

In [55]:

```
# 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 [56]:

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

Out[56]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state
1	140945	p258326	897464ce9ddc600bced1151f324dd63a	Mr.	FL

1.4.2.3 Using Pretrained Models: TFIDF weighted W2V

In [57]:

```
# 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)
```

```
My students are English learners that are working on English as their s
econd or third languages. We are a melting pot of refugees, immigrants,
and native-born Americans bringing the gift of language to our school.
\r\n\r\n We have over 24 languages represented in our English Learner p
rogram with students at every level of mastery.  We also have over 40 c
ountries represented with the families within our school.  Each student
brings a wealth of knowledge and experiences to us that open our eyes t
o new cultures, beliefs, and respect.\"The limits of your language are
the limits of your world.\"-Ludwig Wittgenstein  Our English learner's
have a strong support system at home that begs for more resources.  Man
y times our parents are learning to read and speak English along side o
f their children.  Sometimes this creates barriers for parents to be ab
le to help their child learn phonetics, letter recognition, and other r
```

```
te to help their child learn phonetics, letter recognition, and other r
eading skills.\r\n\r\nBy providing these dvd's and players, students ar
e able to continue their mastery of the English language even if no one
at home is able to assist.  All families with students within the Level
1 proficiency status, will be a offered to be a part of this program.
These educational videos will be specially chosen by the English Learne
r Teacher and will be sent home regularly to watch.  The videos are to
help the child develop early reading skills.\r\n\r\nParents that do not
have access to a dvd player will have the opportunity to check out a dv
d player to use for the year.  The plan is to use these videos and educ
ational dvd's for the years to come for other EL students.\r\nnannan
==================================================
The 51 fifth grade students that will cycle through my classroom this y
ear all love learning, at least most of the time. At our school, 97.3%
of the students receive free or reduced price lunch. Of the 560 student
s, 97.3% are minority students. \r\nThe school has a vibrant community
that loves to get together and celebrate. Around Halloween there is a w
hole school parade to show off the beautiful costumes that students wea
r. On Cinco de Mayo we put on a big festival with crafts made by the st
udents, dances, and games. At the end of the year the school hosts a ca
rnival to celebrate the hard work put in during the school year, with a
dunk tank being the most popular activity.My students will use these fi
ve brightly colored Hokki stools in place of regular, stationary, 4-leg
ged chairs. As I will only have a total of ten in the classroom and not
enough for each student to have an individual one, they will be used in
a variety of ways. During independent reading time they will be used as
special chairs students will each use on occasion. I will utilize them
in place of chairs at my small group tables during math and reading tim
es. The rest of the day they will be used by the students who need the
highest amount of movement in their life in order to stay focused on sc
hool.\r\n\r\nWhenever asked what the classroom is missing, my students
always say more Hokki Stools. They can't get their fill of the 5 stools
we already have. When the students are sitting in group with me on the
Hokki Stools, they are always moving, but at the same time doing their
work. Anytime the students get to pick where they can sit, the Hokki St
ools are the first to be taken. There are always students who head over
to the kidney table to get one of the stools who are disappointed as th
ere are not enough of them. \r\n\r\nWe ask a lot of students to sit for
7 hours a day. The Hokki stools will be a compromise that allow my stud

ents to do desk work and move at the same time. These stools will help
ents to do desk work and move at the same time. These stools will help
students to meet their 60 minutes a day of movement by allowing them to
activate their core muscles for balance while they sit. For many of my
students, these chairs will take away the barrier that exists in school
s for a child who can't sit still.nannan
==================================================
How do you remember your days of school? Was it in a sterile environmen
t with plain walls, rows of desks, and a teacher in front of the room?
A typical day in our room is nothing like that. I work hard to create a
warm inviting themed room for my students look forward to coming to eac
h day.\r\n\r\nMy class is made up of 28 wonderfully unique boys and gir
ls of mixed races in Arkansas.\r\nThey attend a Title I school, which m
eans there is a high enough percentage of free and reduced-price lunch
to qualify. Our school is an \"open classroom\" concept, which is very
unique as there are no walls separating the classrooms. These 9 and 10
year-old students are very eager learners; they are like sponges, absor
bing all the information and experiences and keep on wanting more.With
these resources such as the comfy red throw pillows and the whimsical n
autical hanging decor and the blue fish nets, I will be able to help cr
eate the mood in our classroom setting to be one of a themed nautical e
nvironment. Creating a classroom environment is very important in the s
uccess in each and every child's education. The nautical photo props wi
ll be used with each child as they step foot into our classroom for the
first time on Meet the Teacher evening. I'll take pictures of each chil
d with them, have them developed, and then hung in our classroom ready
for their first day of 4th grade.  This kind gesture will set the tone
before even the first day of school! The nautical thank you cards will
be used throughout the year by the students as they create thank you ca
rds to their team groups.\r\n\r\nYour generous donations will help me t
o help make our classroom a fun, inviting, learning environment from da
y one.\r\n\r\nIt costs lost of money out of my own pocket on resources
to get our classroom ready. Please consider helping with this project t
o make our new school year a very successful one. Thank you!nannan
==================================================
My kindergarten students have varied disabilities ranging from speech a
nd language delays, cognitive delays, gross/fine motor delays, to autis
m. They are eager beavers and always strive to work their hardest worki
ng past their limitations. \r\n\r\nThe materials we have are the ones I
seek out for my students. I teach in a Title I school where most of the

students receive free or reduced price lunch.  Despite their disabiliti
students receive free or reduced price lunch.  Despite their disabiliti
es and limitations, my students love coming to school and come eager to
learn and explore.Have you ever felt like you had ants in your pants an
d you needed to groove and move as you were in a meeting? This is how m
y kids feel all the time. The want to be able to move as they learn or
so they say.Wobble chairs are the answer and I love then because they d
evelop their core, which enhances gross motor and in Turn fine motor sk
ills. \r\nThey also want to learn through games, my kids don't want to
sit and do worksheets. They want to learn to count by jumping and playi
ng. Physical engagement is the key to our success. The number toss and
color and shape mats can make that happen. My students will forget they
are doing work and just have the fun a 6 year old deserves.nannan
==================================================
```

In [58]:

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

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

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

In [59]:

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

```
My kindergarten students have varied disabilities ranging from speech a
nd language delays, cognitive delays, gross/fine motor delays, to autis
```

```
m. They are eager beavers and always strive to work their hardest worki
ng past their limitations. \r\n\r\nThe materials we have are the ones I
seek out for my students. I teach in a Title I school where most of the
students receive free or reduced price lunch.  Despite their disabiliti
es and limitations, my students love coming to school and come eager to
learn and explore.Have you ever felt like you had ants in your pants an
d you needed to groove and move as you were in a meeting? This is how m
y kids feel all the time. The want to be able to move as they learn or
so they say.Wobble chairs are the answer and I love then because they d
evelop their core, which enhances gross motor and in Turn fine motor sk
ills. \r\nThey also want to learn through games, my kids do not want to
sit and do worksheets. They want to learn to count by jumping and playi
ng. Physical engagement is the key to our success. The number toss and
color and shape mats can make that happen. My students will forget they
are doing work and just have the fun a 6 year old deserves.nannan
==================================================
```

In [60]:

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

```
My kindergarten students have varied disabilities ranging from speech a
nd language delays, cognitive delays, gross/fine motor delays, to autis
m. They are eager beavers and always strive to work their hardest worki
ng past their limitations.     The materials we have are the ones I see
k out for my students. I teach in a Title I school where most of the st
udents receive free or reduced price lunch.  Despite their disabilities
and limitations, my students love coming to school and come eager to le
arn and explore.Have you ever felt like you had ants in your pants and
you needed to groove and move as you were in a meeting? This is how my
kids feel all the time. The want to be able to move as they learn or so
they say.Wobble chairs are the answer and I love then because they deve
lop their core, which enhances gross motor and in Turn fine motor skill
s.   They also want to learn through games, my kids do not want to sit
and do worksheets. They want to learn to count by jumping and playing.
Physical engagement is the key to our success. The number toss and colo
```

```
r and shape mats can make that happen. My students will forget they are
doing work and just have the fun a 6 year old deserves.nannan
```

In [61]:

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

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

In [62]:

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

In [63]:
```
# 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):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_essays.append(sent.lower().strip())
```

```
100%|█████████████████████████████████████████████████████████████████
████████| 32414/32414 [00:17<00:00, 1886.24it/s]
```

In [64]:
```
# after preprocesing
preprocessed_essays[20000]
```

```
Out[64]: 'my kindergarten students varied disabilities ranging speech language d
         elays cognitive delays gross fine motor delays autism they eager beaver
         s always strive work hardest working past limitations the materials one
         s i seek students i teach title i school students receive free reduced
         price lunch despite disabilities limitations students love coming schoo
         l come eager learn explore have ever felt like ants pants needed groove
         move meeting this kids feel time the want able move learn say wobble ch
         airs answer i love develop core enhances gross motor turn fine motor sk
         ills they also want learn games kids not want sit worksheets they want
         learn count jumping playing physical engagement key success the number
         toss color shape mats make happen my students forget work fun 6 year ol
         d deserves nannan'
```

1.4 Preprocessing of `project_title`

```
In [65]: # printing some project titles.
         for i in range (0,21):

             print(project_data['project_title'].values[i])
             print("="*50)
```

```
Educational Support for English Learners at Home
==================================================
Wanted: Projector for Hungry Learners
==================================================
Soccer Equipment for AWESOME Middle School Students
==================================================
Techie Kindergarteners
==================================================
Interactive Math Tools
==================================================
Flexible Seating for Mrs. Jarvis' Terrific Third Graders!!
==================================================
Chromebooks for Special Education Reading Program
==================================================
It's the 21st Century
==================================================
Targeting More Success in Class
```

```
==================================================
Just For the Love of Reading--\r\nPure Pleasure
==================================================
Reading Changes Lives
==================================================
Elevating Academics and Parent Rapports Through Technology
==================================================
Building Life Science Experiences
==================================================
Everyone deserves to be heard!
==================================================
TABLETS CAN SHOW US THE WORLD
==================================================
Making Recess Active
==================================================
Making Great LEAP's With Leapfrog!
==================================================
Technology Teaches Tomorrow's Talents Today
==================================================
Test Time
==================================================
Wiggling Our Way to Success
==================================================
Magic Carpet Ride in Our Library
==================================================
```

In [66]:

```
preprocessed_titles = []

for dataset in tqdm(project_data['project_title'].values):
    data = decontracted(dataset)  # Replacing some specific and general
 short form into proper word/stopword.
    data = re.sub(r"it's", "it is", data) # Replacing it's with it is a
s it is not part of function decontracted
    data = data.replace('\\r', ' ') # Replacing \r with space
    data = data.replace('\\"', ' ') # Replacing \ with space
    data = data.replace('\\n', ' ') # Replacing \n with space
    data = re.sub('[^A-Za-z0-9]+', ' ', data) # Replacing special chara
cters with space
    data = re.sub("\S*\d\S*", "", data).strip() # Trimming numbers cont
```

```
aining digits

    data = ' '.join(e for e in data.split() if e not in stopwords) # Re
moving stopwords
    preprocessed_titles.append(data.lower().strip()) # Creating array i
n all the lower cases.
```

```
100%|██████████████████████████████████████████████████████████████████
███████| 32414/32414 [00:00<00:00, 33274.17it/s]
```

In [67]:

```
for i in range (0,21):
    print(preprocessed_titles[i])
    print("="*50)
```

```
educational support english learners home
==================================================
wanted projector hungry learners
==================================================
soccer equipment awesome middle school students
==================================================
techie kindergarteners
==================================================
interactive math tools
==================================================
flexible seating mrs jarvis terrific third graders
==================================================
chromebooks special education reading program
==================================================
it century
==================================================
targeting more success class
==================================================
just for love reading pure pleasure
==================================================
reading changes lives
==================================================
elevating academics parent rapports through technology
==================================================
building life science experiences
==================================================
everyone deserves heard
==================================================
tablets can show us the world
==================================================
making recess active
==================================================
making great leap with leapfrog
==================================================
technology teaches tomorrow talents today
==================================================
test time
==================================================
wiggling our way success
==================================================
magic carpet ride our library
==================================================
```

In [68]:

```
project_data["preprocessed_titles"] = preprocessed_titles

title_word_count = []

for sentence in project_data["preprocessed_titles"] :
    word = len(sentence.split())
    title_word_count.append(word)

project_data["title_word_count"] = title_word_count

project_data.head(5)
```

Out[68]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state

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

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

1.5 Preparing data for models

In [69]:

```
project_data.columns
```

Out[69]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_stat
e',
       'project_submitted_datetime', 'project_grade_category', 'project
_title',
       'project_essay_1', 'project_essay_2', 'project_essay_3',
       'project_essay_4', 'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'project_is_appr
oved',
       'clean_categories', 'clean_subcategories', 'essay',
       'preprocessed_titles', 'title_word_count'],
      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
```

1.5.1 Vectorizing Categorical data

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

In [70]:

```
# we use count vectorizer to convert the values into one
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True)
categories_one_hot = vectorizer.fit_transform(project_data['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds', 'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (32414, 9)
```

In [71]:

```
# we use count vectorizer to convert the values into one
vectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys()), lowercase=False, binary=True)
sub_categories_one_hot = vectorizer.fit_transform(project_data['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot.s
hape)
```

```
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolveme
nt', 'Extracurricular', 'Civics_Government', 'ForeignLanguages', 'Nutri
tionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingA
rts', 'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPre
p', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopme
nt', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Healt
h_Wellness', 'AppliedSciences', 'SpecialNeeds', 'Literature_Writing',
'Mathematics', 'Literacy']
Shape of matrix after one hot encodig  (32414, 30)
```

In [72]:

```
school_state_vectorizer = CountVectorizer(lowercase=False, binary=True)
school_state_vectorizer.fit(project_data['school_state'].values)
print(school_state_vectorizer.get_feature_names())

school_state_one_hot = school_state_vectorizer.transform(project_data[
'school_state'].values)
print("Shape of matrix after one hot encodig ",school_state_one_hot.sha
pe)
print("the type of count vectorizer ",type(school_state_one_hot))
```

```
['AK', 'AL', 'AR', 'AZ', 'CA', 'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'H
I', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'LA', 'MA', 'MD', 'ME', 'MI',
'MN', 'MO', 'MS', 'MT', 'NC', 'ND', 'NE', 'NH', 'NJ', 'NM', 'NV', 'NY',
'OH', 'OK', 'OR', 'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VA', 'VT',
'WA', 'WI', 'WV', 'WY']
Shape of matrix after one hot encodig  (32414, 51)
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
```

In [73]:

```
# https://www.geeksforgeeks.org/python-pandas-dataframe-fillna-to-repla
ce-null-values-in-dataframe/
project_data["teacher_prefix"].fillna("No_Prefix", inplace = True)

teacher_prefix_vectorizer = CountVectorizer(lowercase=False, binary=Tru
e)
teacher_prefix_vectorizer.fit(project_data['teacher_prefix'].values)
print(teacher_prefix_vectorizer.get_feature_names())
```

```
teacher_prefix_one_hot = teacher_prefix_vectorizer.transform(project_data['teacher_prefix'].values)
print("Shape of matrix after one hot encodig ",teacher_prefix_one_hot.shape)
```

```
['Mr', 'Mrs', 'Ms', 'No_Prefix', 'Teacher']
Shape of matrix after one hot encodig  (32414, 5)
```

In [74]:

```
my_grade_counter = Counter()

for project_grade in project_data['project_grade_category'].values:

    if (' ' in project_grade):

        project_grade = project_grade.replace(" ", "~")


    my_grade_counter.update(project_grade.split())


project_grade_cat_dict = dict(my_grade_counter)
sorted_project_grade_cat_dict = dict(sorted(project_grade_cat_dict.items(), key=lambda kv: kv[1]))

grade_cat_vectorizer = CountVectorizer(vocabulary=list(sorted_project_grade_cat_dict.keys()), lowercase=False, binary=True)
grade_cat_vectorizer.fit(project_data['project_grade_category'].values)
print(grade_cat_vectorizer.get_feature_names())

grade_cat_one_hot = grade_cat_vectorizer.transform(project_data['project_grade_category'].values)
print("Shape of matrix after one hot encodig ",grade_cat_one_hot.shape)
```

```
['Grades~9-12', 'Grades~6-8', 'Grades~3-5', 'Grades~PreK-2']
Shape of matrix after one hot encodig  (32414, 4)
```

1.5.2 Vectorizing Text data

1.5.2.1 Bag of words

In [75]:

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

```
Shape of matrix after one hot encodig  (32414, 10299)
```

In [76]:

```
titles_vectorizer = CountVectorizer(min_df=10)
titles_bow = titles_vectorizer.fit_transform(preprocessed_titles)
print("some sample features(unique words in the corpus)",titles_vectorizer.get_feature_names()[0:10])
print("Shape of matrix after one hot encodig ",titles_bow.shape)
print("the type of count vectorizer ",type(titles_bow))
print("the number of unique words ", titles_bow.get_shape()[1])
```

```
some sample features(unique words in the corpus) ['abc', 'about', 'acad
emic', 'access', 'accessible', 'accessing', 'accessories', 'achieve',
'achievement', 'achieving']
Shape of matrix after one hot encodig  (32414, 1601)
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the number of unique words  1601
```

1.5.2.2 TFIDF vectorizer

In [77]:

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

```
Shape of matrix after one hot encodig  (32414, 10299)
```

1.5.2.3 Using Pretrained Models: Avg W2V

In [78]:

```
'''
# Reading glove vectors in python: https://stackoverflow.com/a/3823034
9/4084039
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 o
ur 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-save-and-load-variables-in-python/

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


'''
```

Out[78]:

```
'\n# Reading glove vectors in python: https://stackoverflow.com/a/38230
349/4084039\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        w
ord = splitLine[0]\n        embedding = np.array([float(val) for val in
splitLine[1:]])\n        model[word] = embedding\n    print ("Done.",le
n(model)," words loaded!")\n    return model\nmodel = loadGloveModel
(\'glove.42B.300d.txt\')\n\n# ============================\nOutput:\n
  \nLoading Glove 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 preproce
d_titles:\n    words.extend(i.split(\' \'))\nprint("all the words in th
e coupus", len(words))\nwords = set(words)\nprint("the unique words in
the coupus", len(words))\n\ninter_words = set(model.keys()).intersectio
n(words)\nprint("The number of words that are present in both glove vec
tors and our coupus",       len(inter_words),"(",np.round(len(inter_wor
ds)/len(words)*100,3),"%)")\n\nwords_courpus = {}\nwords_glove = set(mo
del.keys())\nfor i in words:\n    if i in words_glove:\n        words_c
ourpus[i] = model[i]\nprint("word 2 vec length", len(words_courpus))\n
\n\n# stronging variables into pickle files python: http://www.jessicay
```

```
ung.com/how-to-use-pickle-to-save-and-load-variables-in-python/\n\nimpo
rt pickle\nwith open(\'glove_vectors\', \'wb\') as f:\n    pickle.dump
(words_courpus, f)\n\n\n'
```

In [79]:
```
# stronging variables into pickle files python: http://www.jessicayung.
com/how-to-use-pickle-to-save-and-load-variables-in-python/
# make sure you have the glove_vectors file
with open('glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words =  set(model.keys())
```

In [80]:
```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors = []; # the avg-w2v for each sentence/review is stored
 in this list
for sentence in tqdm(preprocessed_essays): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sentence.split(): # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors.append(vector)

print(len(avg_w2v_vectors))
print(len(avg_w2v_vectors[0]))
```

```
100%|██████████████████████████████████████████████████████████████████
████████| 32414/32414 [00:08<00:00, 3719.79it/s]
```

```
32414
300
```

1.5.2.3 Using Pretrained Models: TFIDF weighted W2V

In [81]:

```python
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(preprocessed_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 [82]:

```python
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(preprocessed_essays): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    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 tfidf 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.append(vector)

print(len(tfidf_w2v_vectors))
print(len(tfidf_w2v_vectors[0]))
```

```
100%|██████████████████████████████████████████████████████████████████
█████████| 32414/32414 [01:04<00:00, 500.10it/s]
```

```
32414
300
```

In [83]:

```python
# TFIDF on project titles
titles_tfidf_vectorizer = TfidfVectorizer(min_df=10)
titles_tfidf = titles_tfidf_vectorizer.fit_transform(preprocessed_titles)
print("some sample features(unique words in the corpus)",titles_tfidf_vectorizer.get_feature_names()[10:21])
print("Shape of matrix after one hot encodig ",titles_tfidf.shape)
```

```
some sample features(unique words in the corpus) ['across', 'act', 'act
ion', 'active', 'activities', 'activity', 'add', 'adding', 'adventure',
'adventures', 'after']
Shape of matrix after one hot encodig  (32414, 1601)
```

In [84]:

```python
# AVG W2V on project title
avg_w2v_titles_vectors = [];

for sentence in tqdm(preprocessed_titles):

    vector_titles = np.zeros(300)
    cnt_words_titles = 0;

    for word in sentence.split():

        if word in glove_words:

            vector += model[word]
            cnt_words_titles += 1

    if cnt_words_titles != 0:

        vector_titles /= cnt_words_titles

    avg_w2v_titles_vectors.append(vector_titles)

print(len(avg_w2v_titles_vectors))
print(len(avg_w2v_titles_vectors[0]))
```

```
100%|██████████████████████████████████████████████████
█████| 32414/32414 [00:00<00:00, 73238.03it/s]
```

```
32414
300
```

In [85]:
```
# TFIDF weighted W2V on project_title
titles_tfidf_model = TfidfVectorizer()
titles_tfidf_model.fit(preprocessed_titles)
titles_dictionary = dict(zip(titles_tfidf_model.get_feature_names(), li
st(titles_tfidf_model.idf_)))
titles_tfidf_words = set(titles_tfidf_model.get_feature_names())
```

In [86]:
```
titles_tfidf_w2v_vectors = [];

for titles_sentence in tqdm(preprocessed_titles):

    titles_vector = np.zeros(300)
    titles_tfidf_weight = 0;

    for word in titles_sentence.split():

        if (word in glove_words) and (word in titles_tfidf_words):

            titles_vec = model[word]

            titles_tf_idf = titles_dictionary[word]*(titles_sentence.co
unt(word)/len(titles_sentence.split()))
            titles_vector += (titles_vec * titles_tf_idf)
            titles_tfidf_weight += titles_tf_idf

    if titles_tfidf_weight != 0:

        titles_vector /= titles_tfidf_weight

    titles_tfidf_w2v_vectors.append(titles_vector)

print(len(titles_tfidf_w2v_vectors))
print(len(titles_tfidf_w2v_vectors[0]))
```

```
100%|██████████████████████████████████████████████████████████████████
███████| 32414/32414 [00:00<00:00, 37230.40it/s]
```

```
32414
300
```

1.5.3 Vectorizing Numerical features

In [87]:
```
price_data = resource_data.groupby('id').agg({'price':'sum', 'quantity'
:'sum'}).reset_index()
project_data = pd.merge(project_data, price_data, on='id', how='left')
```

In [88]:
```
# check this one: https://www.youtube.com/watch?v=0HOqOcln3Z4&t=530s
# standardization sklearn: https://scikit-learn.org/stable/modules/gene
rated/sklearn.preprocessing.StandardScaler.html
from sklearn.preprocessing import StandardScaler

# price_standardized = standardScalar.fit(project_data['price'].values)
# this will rise the error
# ValueError: Expected 2D array, got 1D array instead: array=[725.05 21
3.03 329.   ... 399.   287.73   5.5 ].
# Reshape your data either using array.reshape(-1, 1)

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

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

```
Mean : 51.22791725797495, Standard deviation : 149.72149781041014
```

In [89]:
```
price_standardized
```

Out[89]:
```
array([[ 0.65302635],
       [-0.24230266],
       [-0.2857166 ],
       ...,
       [-0.10978996],
       [-0.23014676],
       [-0.10250978]])
```

1.5.4 Merging all the above features

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

In [90]:

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

```
(32414, 9)
(32414, 30)
(32414, 10299)
(32414, 1)
```

In [91]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/40840
39
from scipy.sparse import hstack
# with the same hstack function we are concatinating a sparse matrix an
d a dense matirx :)
X = hstack((categories_one_hot, sub_categories_one_hot, text_bow, price
_standardized))
X.shape
```

Out[91]:

```
(32414, 10339)
```

Computing Sentiment Scores

In [92]:

```
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 te
ach 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 students in my class come
 from a variety of different backgrounds which makes\
for wonderful sharing of experiences and cultures including native amer
icans 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 differen
t 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 m
y 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 an
d writing concepts while cooking delicious healthy \
food for snack time my students will have a grounded appreciation for t
he work that went into making the food and knowledge \
of where the ingredients came from as well as how it is healthy for the
ir bodies this project would expand our learning of \
nutrition and agricultural cooking recipes by having us peel our own ap
ples to make homemade applesauce make our own bread \
and mix up healthy plants from our classroom garden in the spring we wi
ll also create our own cookbooks to be printed and \
shared with families students will gain math and literature skills as w
ell as a life long enjoyment 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, comp
ound)
# 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 10: Clustering

- step 1: Choose any vectorizer (data matrix) that you have worked in any of the assignments, and got the best AUC value.
- step 2: Choose any of the feature selection/reduction algorithms ex: selectkbest features, pretrained word vectors, model based feature selection etc and reduce the number of features to 5k features
- step 3: Apply all three kmeans, Agglomerative clustering, DBSCAN
  - **K-Means Clustering:**
    ● Find the best ‘k’ using the elbow-knee method (plot k vs inertia_)
  - **Agglomerative Clustering:**
    ● Apply agglomerative algorithm and try a different number of clusters like 2,5 etc.
    ● You can take less data points (as this is very computationally expensive one) to perform hierarchical clustering because they do take a considerable amount of time to run.
  - **DBSCAN Clustering:**
    ● Find the best ‘eps’ using the elbow-knee method.
    ● You can take a smaller sample size for this as well.
- step 4: Summarize each cluster by manually observing few points from each cluster.
- step 5: You need to plot the word cloud with essay text for each cluster for each of algorithms mentioned in step 3.

2. Clustering

Splitting data into Train and cross validation(or test): Stratified Sampling

In [93]:

```
approved_project = project_data['project_is_approved'].values
project_data.drop(['project_is_approved'], axis=1, inplace=True)
project_data.head(1)
```

Out[93]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	p
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN	(

1 rows × 21 columns

In [94]:

```
# Data splitting

from sklearn.model_selection import train_test_split

# Splitting in train and test
X_train, X_test, y_train, y_test = train_test_split(project_data, appro
ved_project, test_size=0.33, stratify=approved_project)

# Splitting in Train Test and Cross Validation
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_
size=0.33, stratify=y_train)
```

2.1 Choose the best data matrix on which you got the

best AUC

Naive Bayes Model with bow/tfidf feature had best AUC value 0.71. For this assignment we will use bow feature.

2.2 Make Data Model Ready: encoding numerical, categorical features

In [95]:

```
# Vectorizing Categories on Train, Test and CV data
from sklearn.feature_extraction.text import CountVectorizer

ccvectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True)

# Fit only to train data
ccvectorizer.fit(X_train['clean_categories'].values)

# Transform to train,test and CV data
X_Train_categories_one_hot = ccvectorizer.transform(X_train['clean_categories'].values)
X_Test_categories_one_hot = ccvectorizer.transform(X_test['clean_categories'].values)
X_CV_categories_one_hot = ccvectorizer.transform(X_cv['clean_categories'].values)

print("Shape of train matrix after one hot encodig ",X_Train_categories_one_hot.shape)
print("Shape of test matrix after one hot encodig ",X_Test_categories_one_hot.shape)
print("Shape of cv matrix after one hot encodig ",X_CV_categories_one_hot.shape)
```

```
Shape of train matrix after one hot encodig  (14550, 9)
Shape of test matrix after one hot encodig  (10697, 9)
Shape of cv matrix after one hot encodig  (7167, 9)
```

In [96]:

```
# Vectorizing subcategories on train, test and cv

csvectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys
()), lowercase=False, binary=True)
csvectorizer.fit(X_train['clean_subcategories'].values)

X_Train_sub_categories_one_hot = csvectorizer.transform(X_train['clean_
subcategories'].values)
X_Test_sub_categories_one_hot = csvectorizer.transform(X_test['clean_su
bcategories'].values)
X_CV_sub_categories_one_hot = csvectorizer.transform(X_cv['clean_subcat
egories'].values)

print("Shape of train matrix after one hot encodig ",X_Train_sub_catego
ries_one_hot.shape)
print("Shape of test matrix after one hot encodig ",X_Test_sub_categori
es_one_hot.shape)
print("Shape of cv matrix after one hot encodig ",X_CV_sub_categories_o
ne_hot.shape)
```

```
Shape of train matrix after one hot encodig  (14550, 30)
Shape of test matrix after one hot encodig  (10697, 30)
Shape of cv matrix after one hot encodig  (7167, 30)
```

In [97]:

```
# Vectorizing school state on train , test and cv

school_state_vectorizer = CountVectorizer(lowercase=False, binary=True)

school_state_vectorizer.fit(X_train['school_state'].values)
print(school_state_vectorizer.get_feature_names())

X_Train_school_state_one_hot = school_state_vectorizer.transform(X_trai
n['school_state'].values)
X_Test_school_state_one_hot = school_state_vectorizer.transform(X_test[
'school_state'].values)
X_CV_school_state_one_hot = school_state_vectorizer.transform(X_cv['sch
ool_state'].values)

print("Shape of train matrix after one hot encodig ",X_Train_school_sta
te_one_hot.shape)
print("Shape of test matrix after one hot encodig ",X_Test_school_state
_one_hot.shape)
print("Shape of cv matrix after one hot encodig ",X_CV_school_state_one
_hot.shape)

print("the type of count vectorizer ",type(X_Train_school_state_one_hot
))
print("the type of count vectorizer ",type(X_Test_school_state_one_hot
))
print("the type of count vectorizer ",type(X_CV_school_state_one_hot))
```

```
['AK', 'AL', 'AR', 'AZ', 'CA', 'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'H
I', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'LA', 'MA', 'MD', 'ME', 'MI',
'MN', 'MO', 'MS', 'MT', 'NC', 'ND', 'NE', 'NH', 'NJ', 'NM', 'NV', 'NY',
'OH', 'OK', 'OR', 'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VA', 'VT',
'WA', 'WI', 'WV', 'WY']
Shape of train matrix after one hot encodig  (14550, 51)
Shape of test matrix after one hot encodig  (10697, 51)
Shape of cv matrix after one hot encodig  (7167, 51)
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
```

In [98]:

```
# Vectorizing teacher prefix on train , test and cv

project_data["teacher_prefix"].fillna("No_Prefix", inplace = True)

teacher_prefix_vectorizer = CountVectorizer(lowercase=False, binary=Tru
e)
teacher_prefix_vectorizer.fit(X_train['teacher_prefix'].values)

print(teacher_prefix_vectorizer.get_feature_names())

X_Train_teacher_prefix_one_hot = teacher_prefix_vectorizer.transform(X_
train['teacher_prefix'].values)
X_Test_teacher_prefix_one_hot = teacher_prefix_vectorizer.transform(X_t
est['teacher_prefix'].values)
X_CV_teacher_prefix_one_hot = teacher_prefix_vectorizer.transform(X_cv[
'teacher_prefix'].values)

print("Shape of train matrix after one hot encodig ",X_Train_teacher_pr
efix_one_hot.shape)
print("Shape of test matrix after one hot encodig ",X_Test_teacher_pref
ix_one_hot.shape)
print("Shape of cv matrix after one hot encodig ",X_CV_teacher_prefix_o
ne_hot.shape)
```

```
['Mr', 'Mrs', 'Ms', 'Teacher']
Shape of train matrix after one hot encodig  (14550, 4)
Shape of test matrix after one hot encodig  (10697, 4)
Shape of cv matrix after one hot encodig  (7167, 4)
```

In [99]:

```
# Vectorizing grade category on train , test and cv

my_grade_counter = Counter()

for project_grade in project_data['project_grade_category'].values:

    if (' ' in project_grade):

        project_grade = project_grade.replace(" ", "~")

    my_grade_counter.update(project_grade.split())

project_grade_cat_dict = dict(my_grade_counter)
sorted_project_grade_cat_dict = dict(sorted(project_grade_cat_dict.item
s(), key=lambda kv: kv[1]))

grade_cat_vectorizer = CountVectorizer(vocabulary=list(sorted_project_g
rade_cat_dict.keys()), lowercase=False, binary=True)
grade_cat_vectorizer.fit(X_train['project_grade_category'].values)
print(grade_cat_vectorizer.get_feature_names())

X_Train_grade_cat_one_hot = grade_cat_vectorizer.transform(X_train['pro
ject_grade_category'].values)
X_Test_grade_cat_one_hot = grade_cat_vectorizer.transform(X_test['proje
ct_grade_category'].values)
```

```
X_CV_grade_cat_one_hot = grade_cat_vectorizer.transform(X_cv['project_g
rade_category'].values)

print("Shape of train matrix after one hot encodig ",X_Train_grade_cat_
one_hot.shape)
print("Shape of test matrix after one hot encodig ",X_Test_grade_cat_on
e_hot.shape)
print("Shape of cv matrix after one hot encodig ",X_CV_grade_cat_one_ho
t.shape)
```

```
['Grades~9-12', 'Grades~6-8', 'Grades~3-5', 'Grades~PreK-2']
Shape of train matrix after one hot encodig  (14550, 4)
Shape of test matrix after one hot encodig  (10697, 4)
Shape of cv matrix after one hot encodig  (7167, 4)
```

2.3 Make Data Model Ready: encoding eassay, and project_title

In [100]:

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

In [101]:

```
# preprocessing essay train data
from tqdm import tqdm
X_Train_preprocessed_essays = []
# tqdm is for printing the status bar
for sentence in tqdm(X_train['essay'].values):
    X_Train_essay_sent = decontracted(sentance)
    X_Train_essay_sent = X_Train_essay_sent.replace('\\r', ' ')
    X_Train_essay_sent = X_Train_essay_sent.replace('\\"', ' ')
    X_Train_essay_sent = X_Train_essay_sent.replace('\\n', ' ')
    X_Train_essay_sent = re.sub('[^A-Za-z0-9]+', ' ', X_Train_essay_sen
t)

    X_Train_essay_sent = ' '.join(e for e in X_Train_essay_sent.split()
 if e.lower() not in stopwords)
    X_Train_preprocessed_essays.append(X_Train_essay_sent.lower().strip
())
```

```
100%|██████████████████████████████████████████████████████████████████
████████| 14550/14550 [00:12<00:00, 1212.25it/s]
```

In [102]:

```
# preprocessing essay test data
from tqdm import tqdm
X_Test_preprocessed_essays = []
# tqdm is for printing the status bar
for sentence in tqdm(X_test['essay'].values):
    X_Test_essay_sent = decontracted(sentence)
    X_Test_essay_sent = X_Test_essay_sent.replace('\\r', ' ')
    X_Test_essay_sent = X_Test_essay_sent.replace('\\"', ' ')
    X_Test_essay_sent = X_Test_essay_sent.replace('\\n', ' ')
    X_Test_essay_sent = re.sub('[^A-Za-z0-9]+', ' ', X_Test_essay_sent)

    X_Test_essay_sent = ' '.join(e for e in X_Test_essay_sent.split() i
f e.lower() not in stopwords)
    X_Test_preprocessed_essays.append(X_Test_essay_sent.lower().strip
())
```

```
100%|██████████████████████████████████████████████████████████████████
████████| 10697/10697 [00:05<00:00, 1889.83it/s]
```

In [103]:

```
# preprocessing essay cv data
from tqdm import tqdm
X_CV_preprocessed_essays = []
# tqdm is for printing the status bar
for sentence in tqdm(X_cv['essay'].values):
    X_CV_essay_sent = decontracted(sentence)
    X_CV_essay_sent = X_CV_essay_sent.replace('\\r', ' ')
    X_CV_essay_sent = X_CV_essay_sent.replace('\\"', ' ')
    X_CV_essay_sent = X_CV_essay_sent.replace('\\n', ' ')
    X_CV_essay_sent = re.sub('[^A-Za-z0-9]+', ' ', X_CV_essay_sent)

    X_CV_essay_sent = ' '.join(e for e in X_CV_essay_sent.split() if e.
lower() not in stopwords)
    X_CV_preprocessed_essays.append(X_CV_essay_sent.lower().strip())
```

```
100%|██████████████████████████████████████████████████████████████████
██████████| 7167/7167 [00:03<00:00, 1805.17it/s]
```

In [104]:

```
# preprocessing project title train data
X_Train_preprocessed_titles = []

for dataset in tqdm(X_train['project_title'].values):
    data = decontracted(dataset)  # Replacing some specific and general
 short form into proper word/stopword.
    data = re.sub(r"it's", "it is", data) # Replacing it's with it is a
s it is not part of function decontracted
    data = data.replace('\\r', ' ') # Replacing \r with space
    data = data.replace('\\"', ' ') # Replacing \ with space
    data = data.replace('\\n', ' ') # Replacing \n with space
    data = re.sub('[^A-Za-z0-9]+', ' ', data) # Replacing special chara
cters with space
    data = re.sub("\S*\d\S*", "", data).strip() # Trimming numbers cont
aining digits

    data = ' '.join(e for e in data.split() if e not in stopwords) # Re
moving stopwords
    X_Train_preprocessed_titles.append(data.lower().strip()) # Creating
 array in all the lower cases.
```

```
100%|██████████████████████████████████████████████████████████████████
███████| 14550/14550 [00:00<00:00, 31590.38it/s]
```

In [105]:

```
# preprocessing project title test data
X_Test_preprocessed_titles = []

for dataset in tqdm(X_test['project_title'].values):
    data = decontracted(dataset)  # Replacing some specific and general
 short form into proper word/stopword.
    data = re.sub(r"it's", "it is", data) # Replacing it's with it is a
s it is not part of function decontracted
    data = data.replace('\\r', ' ') # Replacing \r with space
    data = data.replace('\\"', ' ') # Replacing \ with space
    data = data.replace('\\n', ' ') # Replacing \n with space
    data = re.sub('[^A-Za-z0-9]+', ' ', data) # Replacing special chara
cters with space
    data = re.sub("\S*\d\S*", "", data).strip() # Trimming numbers cont
aining digits

    data = ' '.join(e for e in data.split() if e not in stopwords) # Re
moving stopwords
    X_Test_preprocessed_titles.append(data.lower().strip()) # Creating
 array in all the lower cases.
```

```
100%|██████████████████████████████████████████████████████████████████
███████| 10697/10697 [00:00<00:00, 33760.97it/s]
```

In [106]:

```
# preprocessing project title cv data
X_CV_preprocessed_titles = []

for dataset in tqdm(X_cv['project_title'].values):
    data = decontracted(dataset)  # Replacing some specific and general
 short form into proper word/stopword.
    data = re.sub(r"it's", "it is", data) # Replacing it's with it is a
s it is not part of function decontracted
    data = data.replace('\\r', ' ') # Replacing \r with space
    data = data.replace('\\"', ' ') # Replacing \ with space
    data = data.replace('\\n', ' ') # Replacing \n with space
    data = re.sub('[^A-Za-z0-9]+', ' ', data) # Replacing special chara
cters with space
    data = re.sub("\S*\d\S*", "", data).strip() # Trimming numbers cont
aining digits

    data = ' '.join(e for e in data.split() if e not in stopwords) # Re
moving stopwords
    X_CV_preprocessed_titles.append(data.lower().strip()) # Creating ar
ray in all the lower cases.
```

```
100%|██████████████████████████████████████████████████████████████████
█████████| 7167/7167 [00:00<00:00, 27793.03it/s]
```

In [107]:

```
# BOW Essay train, test and cv data
# We are considering only the words which appeared in at least 10 docum
ents(rows or projects).
bow_essay_vectorizer = CountVectorizer(min_df=10, max_features = 5000)
bow_essay_vectorizer.fit(X_Train_preprocessed_essays)

X_Train_essay_bow = bow_essay_vectorizer.transform(X_Train_preprocessed
_essays)
X_Test_essay_bow = bow_essay_vectorizer.transform(X_Test_preprocessed_e
ssays)
X_CV_essay_bow = bow_essay_vectorizer.transform(X_CV_preprocessed_essay
s)

print("Shape of train matrix after one hot encodig ",X_Train_essay_bow.
shape)
print("Shape of test matrix after one hot encodig ",X_Test_essay_bow.sh
ape)
print("Shape of CV matrix after one hot encodig ",X_CV_essay_bow.shape)
```

```
Shape of train matrix after one hot encodig  (14550, 133)
Shape of test matrix after one hot encodig  (10697, 133)
Shape of CV matrix after one hot encodig  (7167, 133)
```

In [108]:

```
# BOW title train,test and cv data

titles_vectorizer = CountVectorizer(min_df=10, max_features = 5000)
titles_vectorizer.fit(X_Train_preprocessed_titles)

X_Train_titles_bow = titles_vectorizer.transform(X_Train_preprocessed_t
itles)
X_Test_titles_bow = titles_vectorizer.transform(X_Test_preprocessed_tit
les)
X_CV_titles_bow = titles_vectorizer.transform(X_CV_preprocessed_titles)

print("some sample features(unique words in the corpus)",titles_vectori
zer.get_feature_names()[0:10])
print("Shape of train matrix after one hot encodig ",X_Train_titles_bow
.shape)
print("Shape of test matrix after one hot encodig ",X_Test_titles_bow.s
hape)
print("Shape of CV matrix after one hot encodig ",X_CV_titles_bow.shape)
```

```
some sample features(unique words in the corpus) ['about', 'academic', 'access', 'action', 'active', 'activities', 'activity', 'add', 'adding', 'adventure']
Shape of train matrix after one hot encodig  (14550, 879)
Shape of test matrix after one hot encodig  (10697, 879)
Shape of CV matrix after one hot encodig  (7167, 879)
```

In [111]:

```
# Vectorizing numerical feature

# Merging price data with train, test and cv
X_train = pd.merge(X_train, price_data, on='id', how='left')
X_test = pd.merge(X_test, price_data, on='id', how='left')
X_cv = pd.merge(X_cv, price_data, on='id', how='left')
```

In [112]:

```
# Standardizing price train test and cv data

from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
# normalizer.fit(X_train['price'].values)
# this will rise an error Expected 2D array, got 1D array instead:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['price'].values.reshape(-1,1))

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

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

```
After vectorizations
(14550, 1) (14550,)
(10697, 1) (10697,)
(7167, 1) (7167,)
=======================================================================
=============================
```

In [113]:

```
normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].
values.reshape(-1,1))

X_train_price_norm = normalizer.transform(X_train['teacher_number_of_pr
eviously_posted_projects'].values.reshape(-1,1))
X_test_price_norm = normalizer.transform(X_test['teacher_number_of_prev
iously_posted_projects'].values.reshape(-1,1))
X_cv_price_norm = normalizer.transform(X_cv['teacher_number_of_previous
ly_posted_projects'].values.reshape(-1,1))

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

```
After vectorizations
(14550, 1) (14550,)
(10697, 1) (10697,)
(7167, 1) (7167,)
=======================================================================
=============================
```

In [114]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/40840
39
from scipy.sparse import hstack

# Train data stack
```

```
X_tr = hstack((X_Train_categories_one_hot,X_Train_sub_categories_one_ho
t,X_Train_school_state_one_hot,
               X_Train_teacher_prefix_one_hot,X_Train_grade_cat_one_hot
,X_Train_essay_bow,X_Train_titles_bow,
               X_train_price_norm)).tocsr()

# CV data Stack
X_cr = hstack((X_CV_categories_one_hot,X_CV_sub_categories_one_hot,X_CV
_school_state_one_hot,
               X_CV_teacher_prefix_one_hot,X_CV_grade_cat_one_hot,X_CV_
essay_bow,X_CV_titles_bow,
               X_cv_price_norm)).tocsr()

# Test Data Stack
X_te = hstack((X_Test_categories_one_hot,X_Test_sub_categories_one_hot,
X_Test_school_state_one_hot,
               X_Test_teacher_prefix_one_hot,X_Test_grade_cat_one_hot,X
_Test_essay_bow,X_Test_titles_bow,
               X_test_price_norm)).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(14550, 1111) (14550,)
(7167, 1111) (7167,)
(10697, 1111) (10697,)
========================================================================
============================
```

2.4 Dimensionality Reduction on the selected features

In [115]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.feature_sel
ection.SelectKBest.html#sklearn.feature_selection.SelectKBest
# https://scikit-learn.org/stable/modules/generated/sklearn.feature_sel
```

```
ection.f_classif.html#sklearn.feature_selection.f_classif

from sklearn.feature_selection import SelectKBest, chi2, f_classif

new_feature = SelectKBest(f_classif, k=1000)      # reduced to 1000 fea
tures as total features are 1111.

new_feature.fit(X_tr,y_train)

X_tr_new = new_feature.transform(X_tr)
X_te_new = new_feature.transform(X_te)
X_cr_new = new_feature.transform(X_cr)

print("After Dimensionality Reduction")
print(X_tr_new.shape, y_train.shape)
print(X_cr_new.shape, y_cv.shape)
print(X_te_new.shape, y_test.shape)
print("="*100)
```

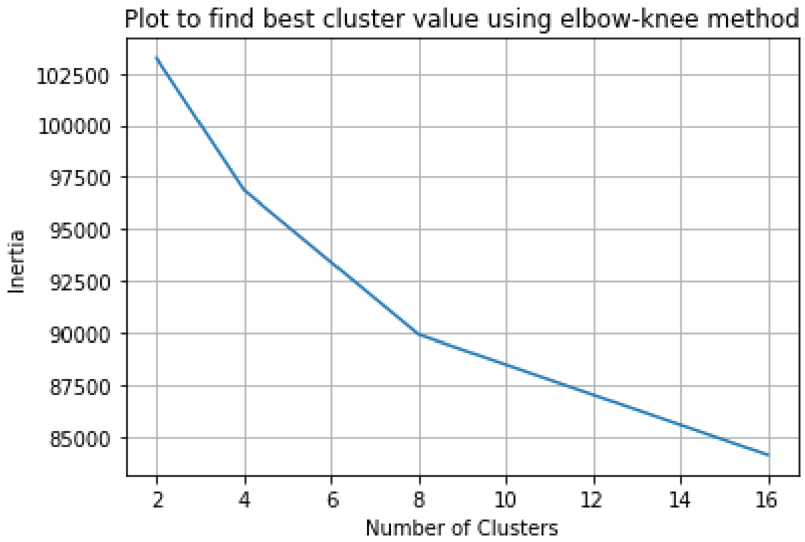
```
After Dimensionality Reduction
(14550, 1000) (14550,)
(7167, 1000) (7167,)
(10697, 1000) (10697,)
========================================================================
=============================
```

2.5 Apply Kmeans

In [116]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMe
ans.html
# https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-
kmeans/

from sklearn.cluster import KMeans
hyperparameter = [2, 4, 8, 16]
inertias = []

for n_cluster in tqdm(hyperparameter):
    kmeans = KMeans(n_clusters=n_cluster, random_state=0).fit(X_tr_new)
    inertias.append(kmeans.inertia_)

plt.plot(hyperparameter, inertias)
plt.xlabel("Number of Clusters")
plt.ylabel("Inertia")
plt.title("Plot to find best cluster value using elbow-knee method")
plt.grid()
plt.show()
```

```
100%|████████████████████████████████████████████████████████████████
██████████████████| 4/4 [02:51<00:00, 42.28s/it]
```



In [117]:
```
kmeans_best = KMeans(n_clusters=8, random_state=0).fit(X_tr_new)
```

In [118]:
```
kmeans_best.labels_
```

Out[118]:
```
array([5, 5, 5, ..., 4, 3, 3])
```

In [119]:
```
len(kmeans_best.labels_)
```

Out[119]: 14550

In [218]:
```
# Creating different clusters

cluster_0 = []
cluster_1 = []
cluster_2 = []
cluster_3 = []
cluster_4 = []
cluster_5 = []
cluster_6 = []
cluster_7 = []

for val in range(0, len(kmeans_best.labels_)):

    if (kmeans_best.labels_[val] == 0):
        cluster_0.append(preprocessed_essays[val])

    if (kmeans_best.labels_[val] == 1):
        cluster_1.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 2:
        cluster_2.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 3:
        cluster_3.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 4:
        cluster_4.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 5:
        cluster_5.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 6:
        cluster_6.append(preprocessed_essays[val])

    if kmeans_best.labels_[val] == 7:
        cluster_7.append(preprocessed_essays[val])
```

In [220]:

```
from wordcloud import WordCloud


def create_wordcloud(cluster, cluster_num, algo_name):

    unique_string=(" ").join(cluster)
    wordcloud = WordCloud(width = 800, height = 800, background_color =
'black', min_font_size = 6).generate(unique_string)

    print("Word cloud plot for essay text for cluster {} for {} algorit
hm".format(cluster_num, algo_name))
    plt.figure(figsize = (8, 6))
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.tight_layout(pad = 0)
    plt.show()

    plt.close()
```

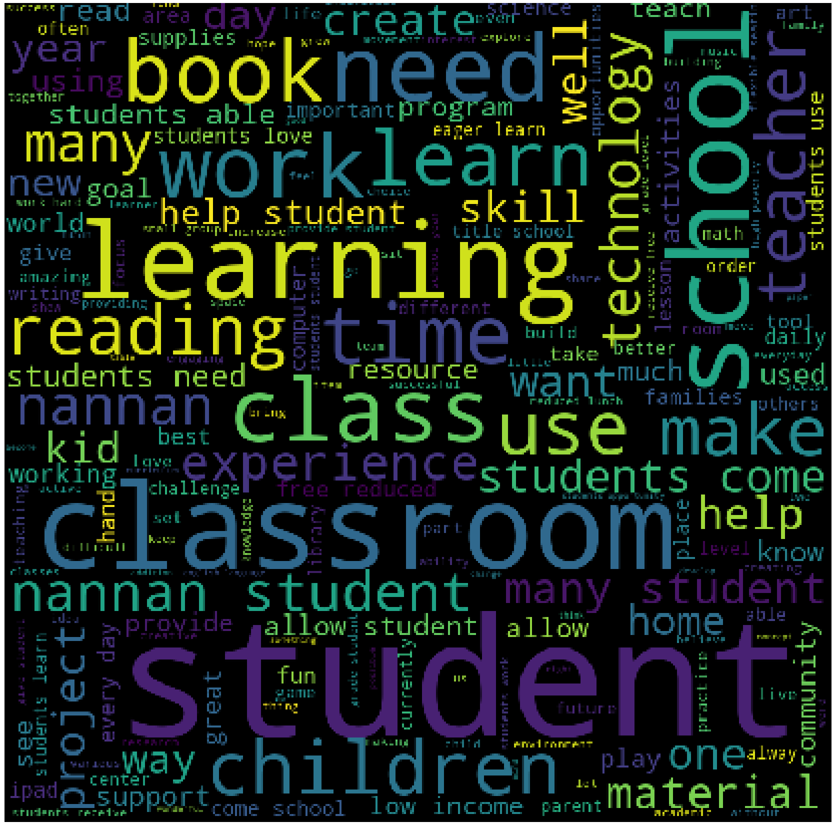
In [221]:

```
# https://www.geeksforgeeks.org/python-program-to-count-words-in-a-sent
ence/
# https://www.geeksforgeeks.org/find-k-frequent-words-data-set-python/

def wordcloud_analysis(cluster, cluster_num):

    from collections import Counter

    word_data = []
    final_data = []

    for i in range(len(cluster)):

        word_data.append(cluster[i])

    word_data = " ".join(word_data)

    cluster_word_split = word_data.split()

    #print(cluster_word_split)
    tot_words = len(cluster_word_split)

    print("Total number of words in cluster {} are {}".format(cluster_num, tot_words))

    Counter = Counter(cluster_word_split)

    cluster_most_common_words = Counter.most_common(10)

    print("Top 10 most common words in cluster {} are {}".format(cluster_num, cluster_most_common_words))
```

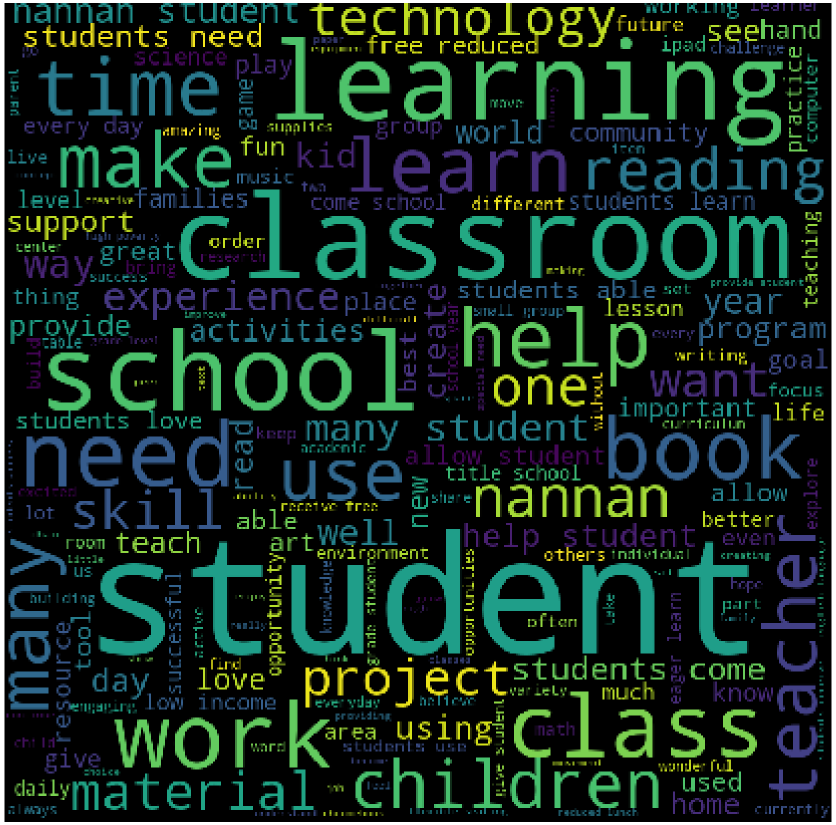
In [222]:

```
cluster = [cluster_0, cluster_1, cluster_2, cluster_3, cluster_4, cluster_5, cluster_6, cluster_7]
count = 0

for val in cluster:

    create_wordcloud(val, count, "K-Means")
    wordcloud_analysis(val, count)
    count +=1
```

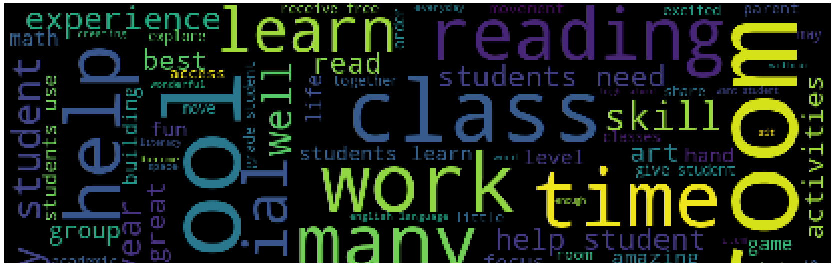
```
Word cloud plot for essay text for cluster 0 for K-Means algorithm
```

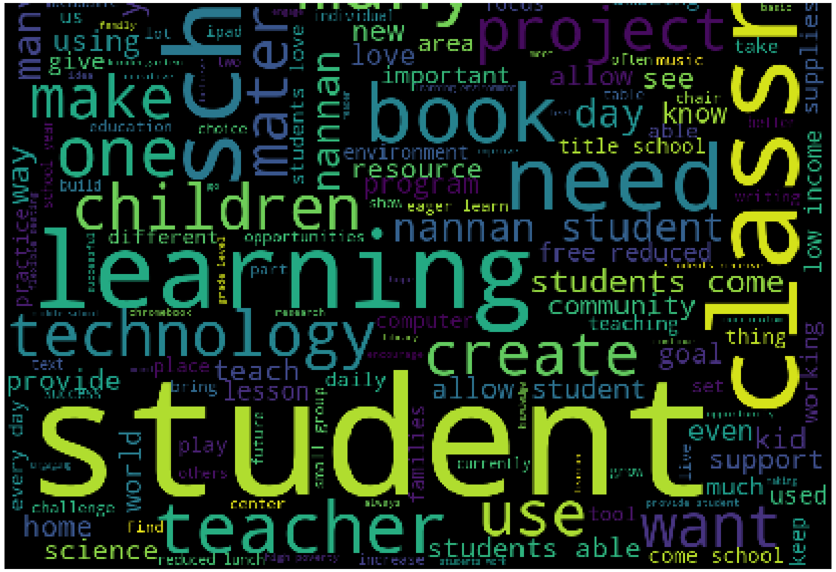


```
Total number of words in cluster 0 are 275009
Top 10 most common words in cluster 0 are [('students', 13165), ('i', 5
410), ('school', 4202), ('my', 3036), ('learning', 2924), ('classroom',
2802), ('they', 2210), ('not', 2209), ('the', 2162), ('learn', 2082)]
Word cloud plot for essay text for cluster 1 for K-Means algorithm
```

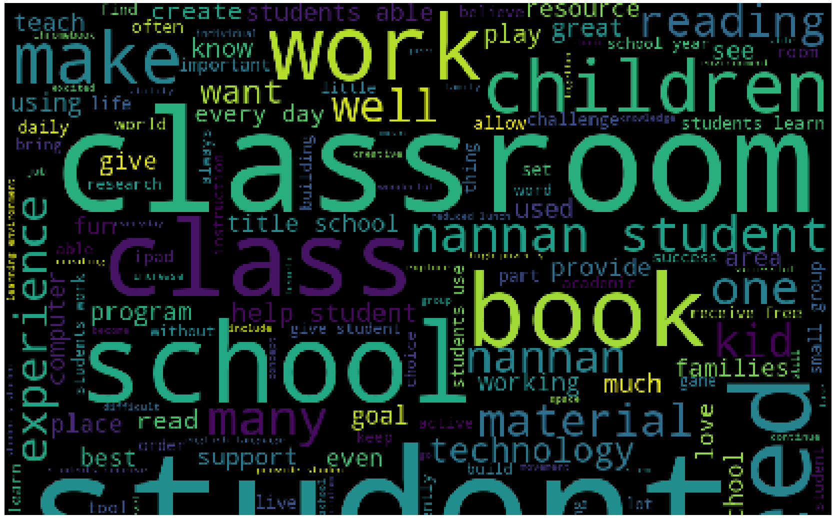


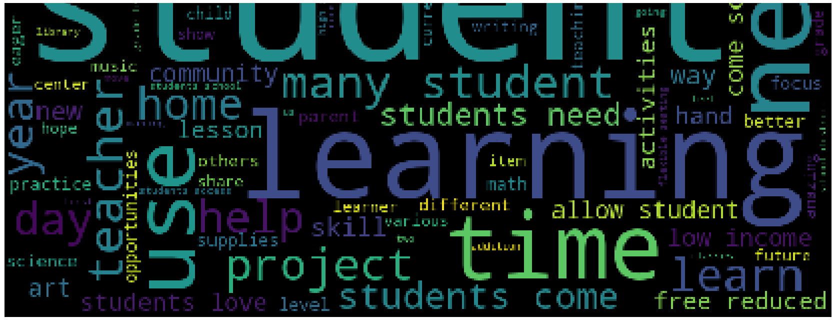
```
Total number of words in cluster 1 are 233192
Top 10 most common words in cluster 1 are [('students', 11201), ('i', 4
741), ('school', 3596), ('my', 2658), ('learning', 2485), ('classroom',
2461), ('learn', 1885), ('the', 1854), ('they', 1823), ('help', 1765)]
Word cloud plot for essay text for cluster 2 for K-Means algorithm
```



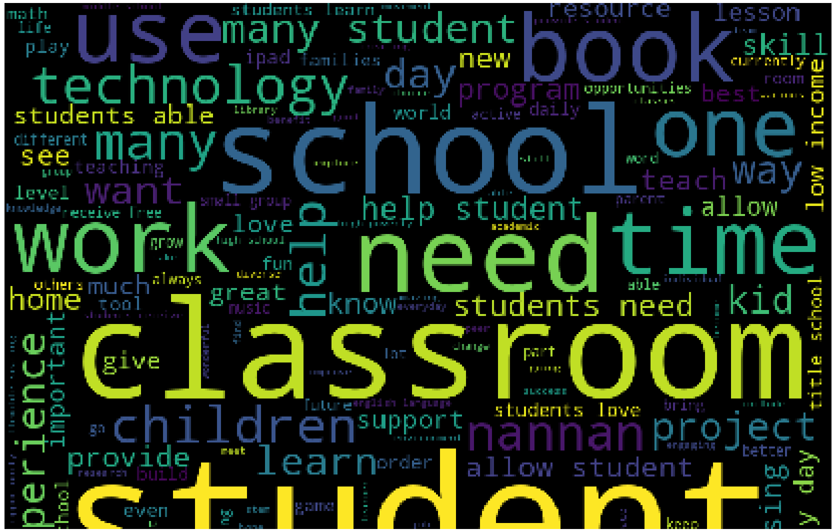


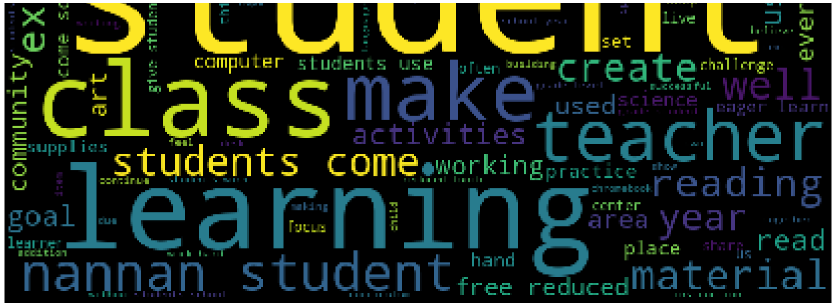
```
Total number of words in cluster 2 are 242928
Top 10 most common words in cluster 2 are [('students', 11806), ('i', 5
000), ('school', 3776), ('my', 2662), ('learning', 2599), ('classroom',
2481), ('the', 2032), ('not', 1971), ('learn', 1878), ('they', 1855)]
Word cloud plot for essay text for cluster 3 for K-Means algorithm
```



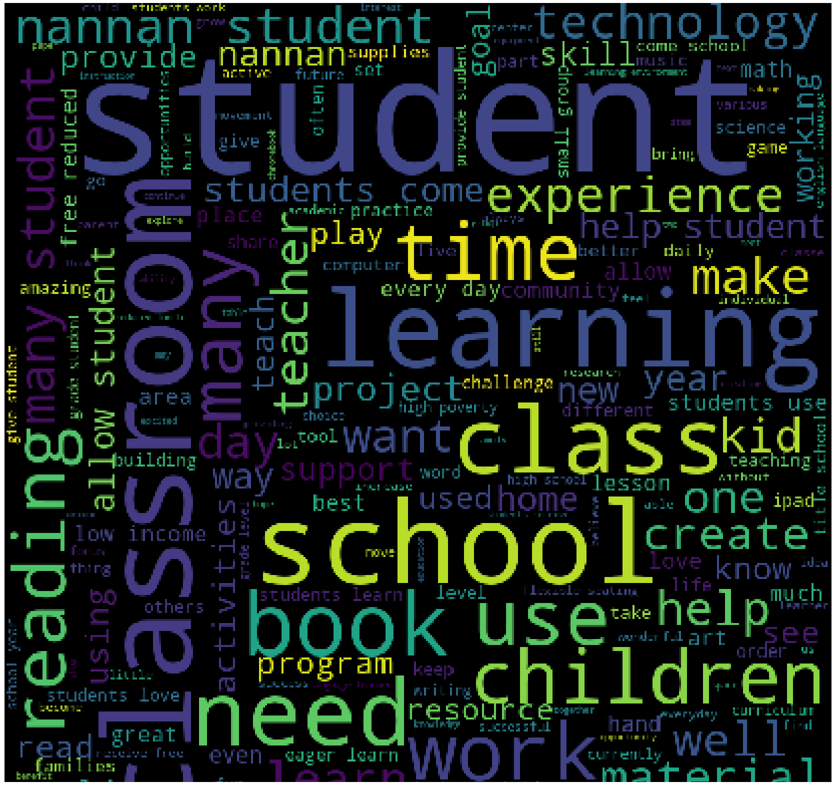


```
Total number of words in cluster 3 are 335322
Top 10 most common words in cluster 3 are [('students', 15916), ('i', 6
826), ('school', 5133), ('my', 3715), ('learning', 3624), ('classroom',
3555), ('the', 2813), ('they', 2653), ('not', 2601), ('help', 2494)]
Word cloud plot for essay text for cluster 4 for K-Means algorithm
```

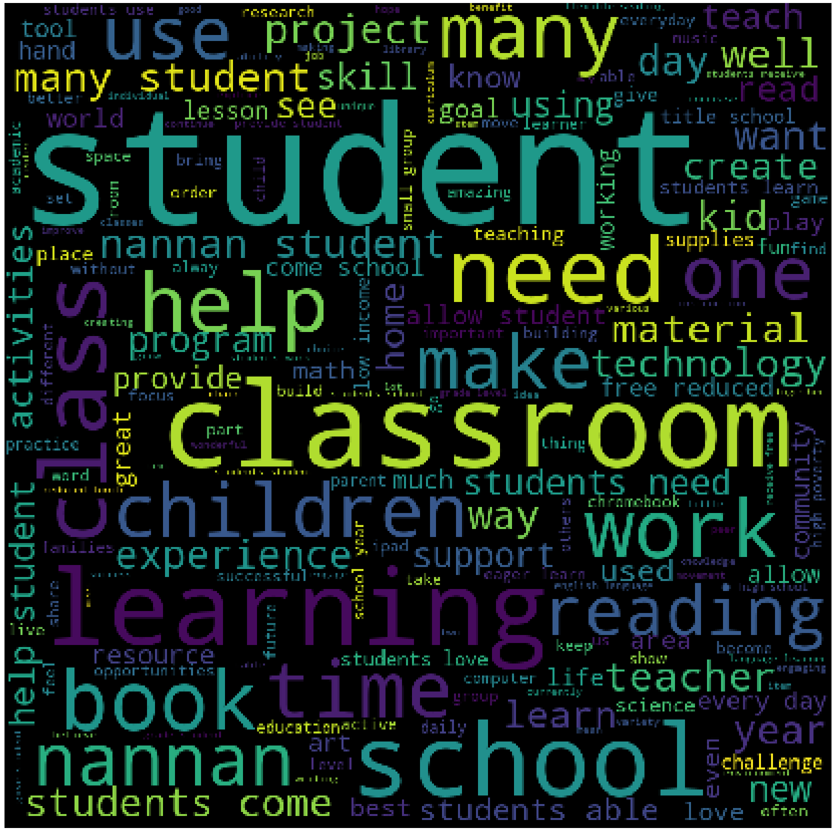




```
Total number of words in cluster 4 are 326325
Top 10 most common words in cluster 4 are [('students', 15621), ('i', 6
452), ('school', 5047), ('my', 3684), ('learning', 3491), ('classroom',
3370), ('the', 2648), ('they', 2635), ('not', 2606), ('learn', 2520)]
Word cloud plot for essay text for cluster 5 for K-Means algorithm
```

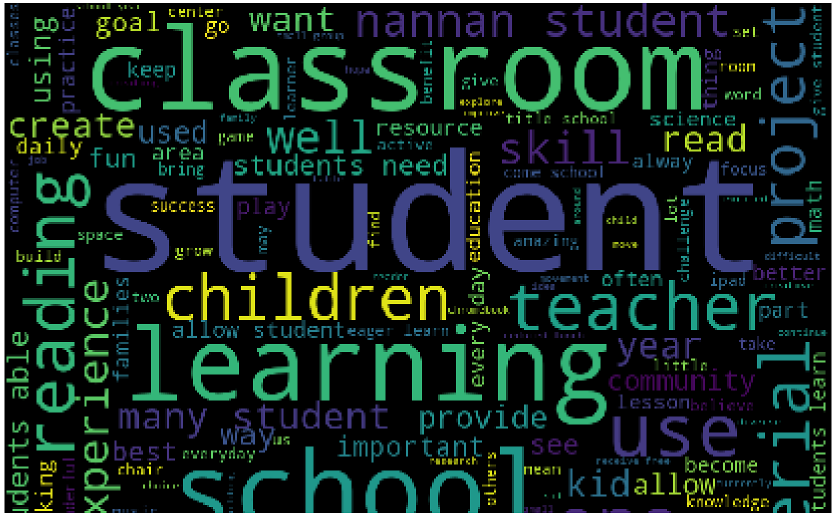


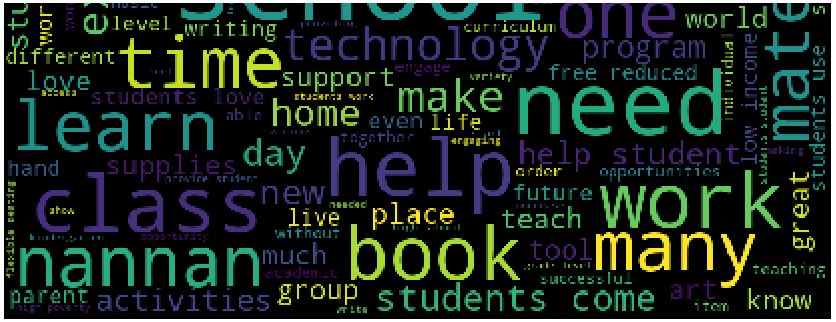
```
Total number of words in cluster 5 are 319401
Top 10 most common words in cluster 5 are [('students', 15312), ('i', 6
463), ('school', 4984), ('my', 3501), ('learning', 3494), ('classroom',
3334), ('the', 2655), ('not', 2513), ('learn', 2438), ('help', 2437)]
Word cloud plot for essay text for cluster 6 for K-Means algorithm
```



```
Total number of words in cluster 6 are 289543
Top 10 most common words in cluster 6 are [('students', 13771), ('i', 5
947), ('school', 4587), ('my', 3127), ('learning', 3057), ('classroom',
2949), ('the', 2317), ('they', 2315), ('not', 2312), ('learn', 2199)]
```

Word cloud plot for essay text for cluster 7 for K-Means algorithm





```
Total number of words in cluster 7 are 176070
Top 10 most common words in cluster 7 are [('students', 8489), ('i', 35
60), ('school', 2700), ('my', 2023), ('learning', 1860), ('classroom',
1792), ('the', 1483), ('not', 1413), ('they', 1344), ('help', 1313)]
```

2.6 Apply AgglomerativeClustering

In [223]:

```
# https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
# https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
```

```
from sklearn.cluster import AgglomerativeClustering

clustering = AgglomerativeClustering(n_clusters=2).fit(X_tr_new.toarray
())
```

In [224]:

```
clustering.labels_
```

Out[224]:

```
array([1, 1, 1, ..., 0, 0, 0], dtype=int64)
```

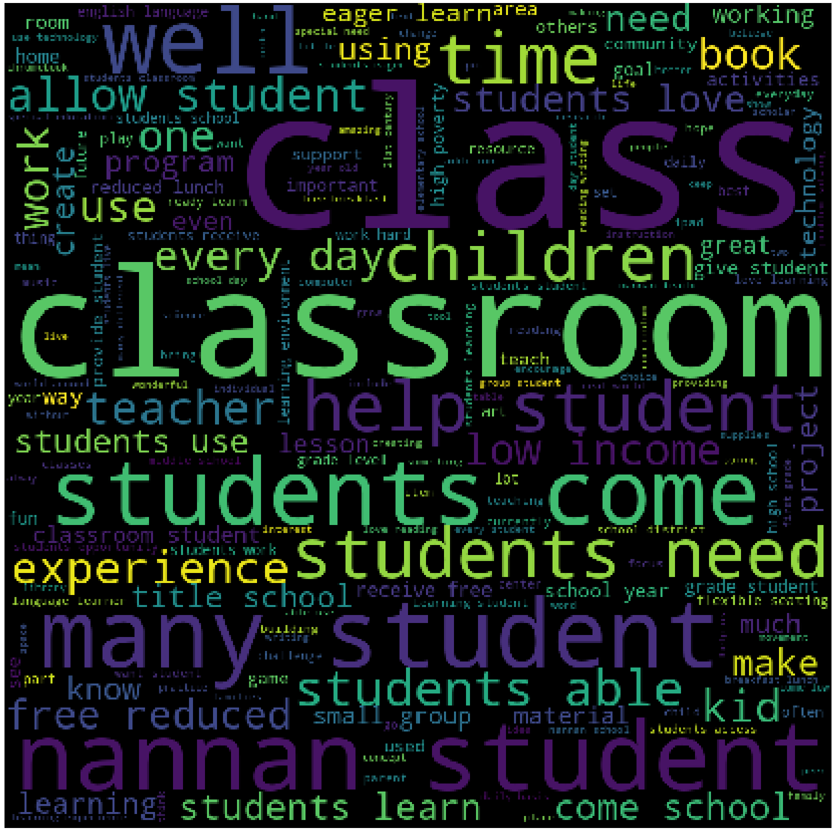
In [225]:

```
# Creating different clusters

cluster_0 = []
cluster_1 = []

for val in range(0, len(clustering.labels_)):

    if (clustering.labels_[val] == 0):
        cluster_0.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 1):
        cluster_1.append(preprocessed_essays[val])
```

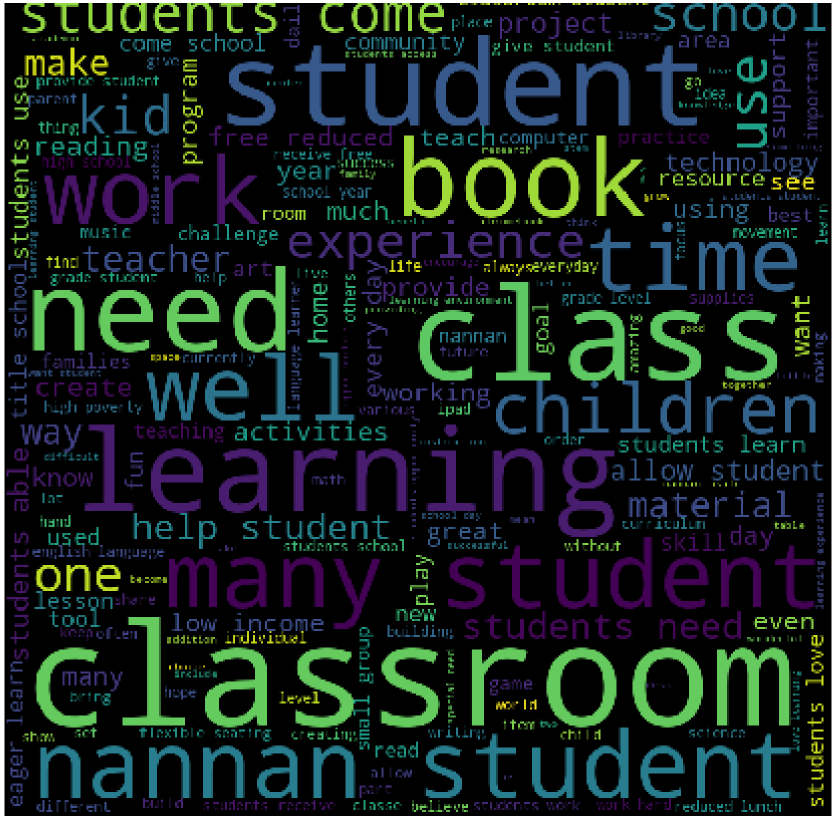
In [226]:

```
cluster = [cluster_0, cluster_1]
count = 0

for val in cluster:

    create_wordcloud(val, count, "Agglomerative")
    wordcloud_analysis(val, count)
    count +=1
```

```
Word cloud plot for essay text for cluster 0 for Agglomerative algorith
m
```



```
Total number of words in cluster 0 are 1542836
Top 10 most common words in cluster 0 are [('students', 73906), ('i', 3
1235), ('school', 23899), ('my', 17300), ('learning', 16540), ('classro
om', 15965), ('the', 12590), ('not', 12166), ('they', 12122), ('learn',
11893)]
Word cloud plot for essay text for cluster 1 for Agglomerative algorith
m
```



```
Total number of words in cluster 1 are 654954
Top 10 most common words in cluster 1 are [('students', 31375), ('i', 1
3164), ('school', 10126), ('my', 7106), ('learning', 6994), ('classroo
m', 6779), ('the', 5374), ('not', 5218), ('they', 5117), ('help', 497
0)]
```

In [227]:

```
# https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
# https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

from sklearn.cluster import AgglomerativeClustering

clustering = AgglomerativeClustering(n_clusters=5).fit(X_tr_new.toarray())
```

In [228]:

```
clustering.labels_
```

Out[228]:

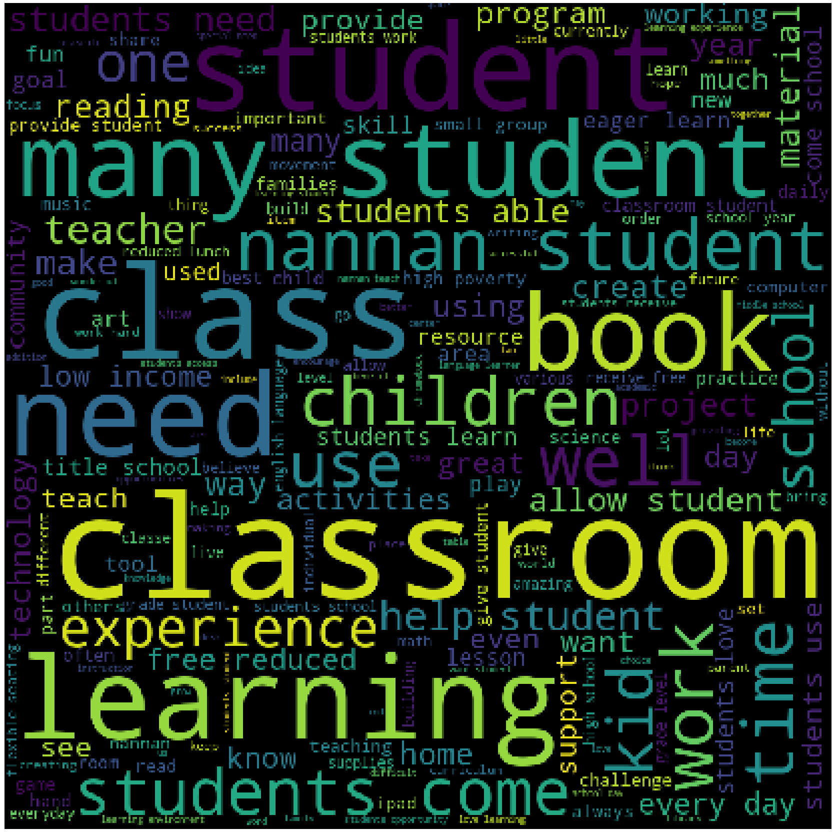
```
array([0, 0, 0, ..., 2, 2, 2], dtype=int64)
```

In [229]:

```
# Creating different clusters

cluster_0 = []
cluster_1 = []
cluster_2 = []
cluster_3 = []
cluster_4 = []

for val in range(0, len(clustering.labels_)):

    if (clustering.labels_[val] == 0):
        cluster_0.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 1):
        cluster_1.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 2):
        cluster_2.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 3):
        cluster_3.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 4):
        cluster_4.append(preprocessed_essays[val])
```

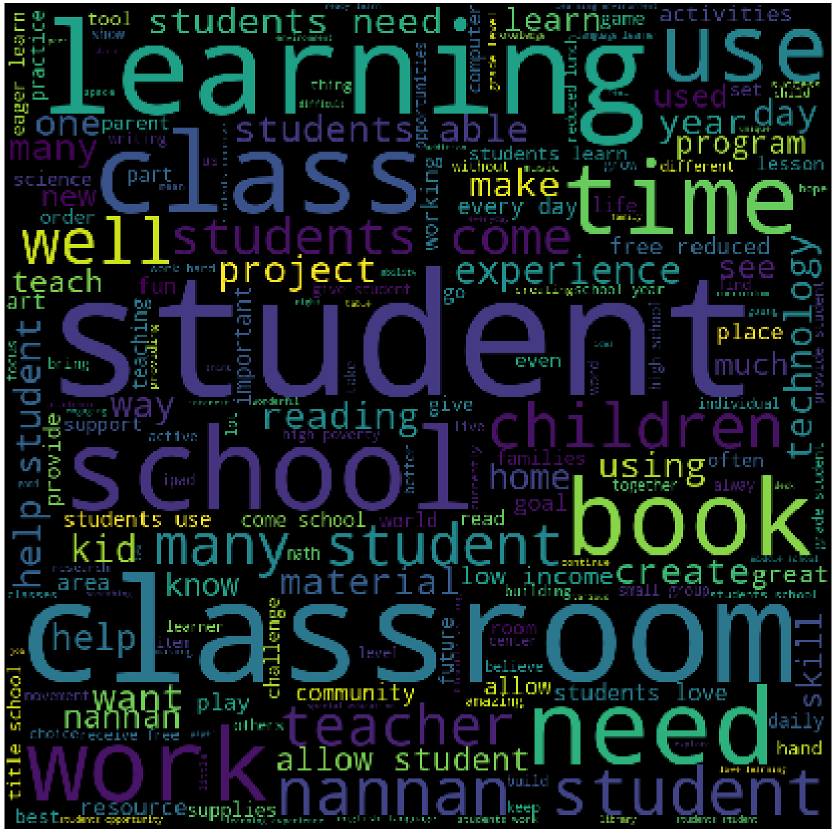
In [230]:
```python
cluster = [cluster_0, cluster_1, cluster_2, cluster_3, cluster_4]
count = 0

for val in cluster:

    create_wordcloud(val, count, "Agglomerative")
    wordcloud_analysis(val, count)
    count +=1
```

```
Word cloud plot for essay text for cluster 0 for Agglomerative algorith
m
```



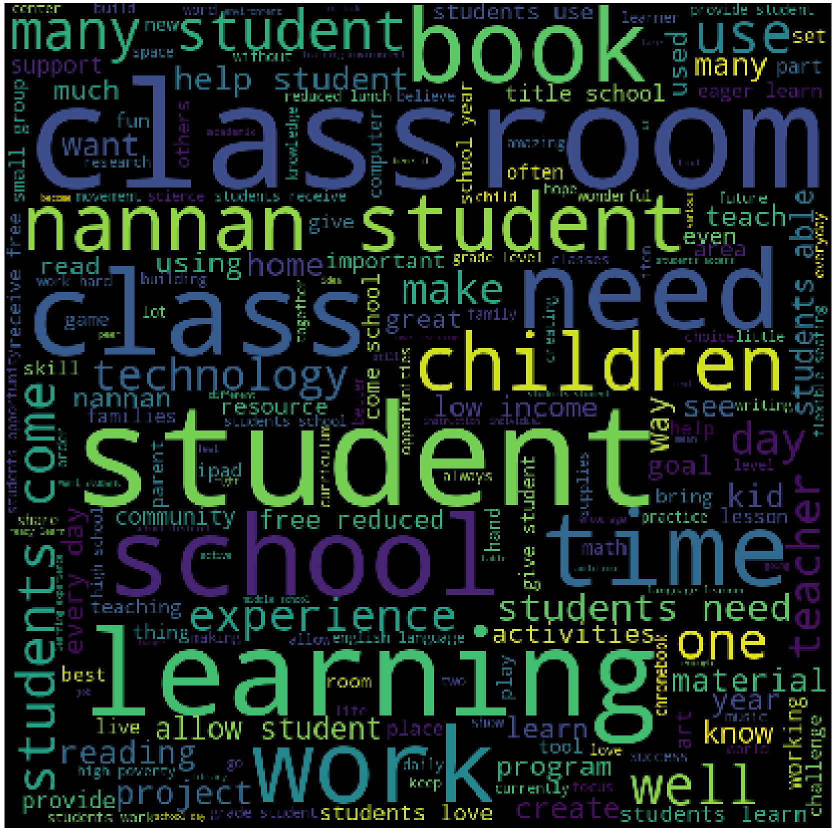
```
Total number of words in cluster 0 are 654954
```

```
Top 10 most common words in cluster 0 are [('students', 31375), ('i', 1
3164), ('school', 10126), ('my', 7106), ('learning', 6994), ('classroo
m', 6779), ('the', 5374), ('not', 5218), ('they', 5117), ('help', 497
0)]
Word cloud plot for essay text for cluster 1 for Agglomerative algorith
m
```

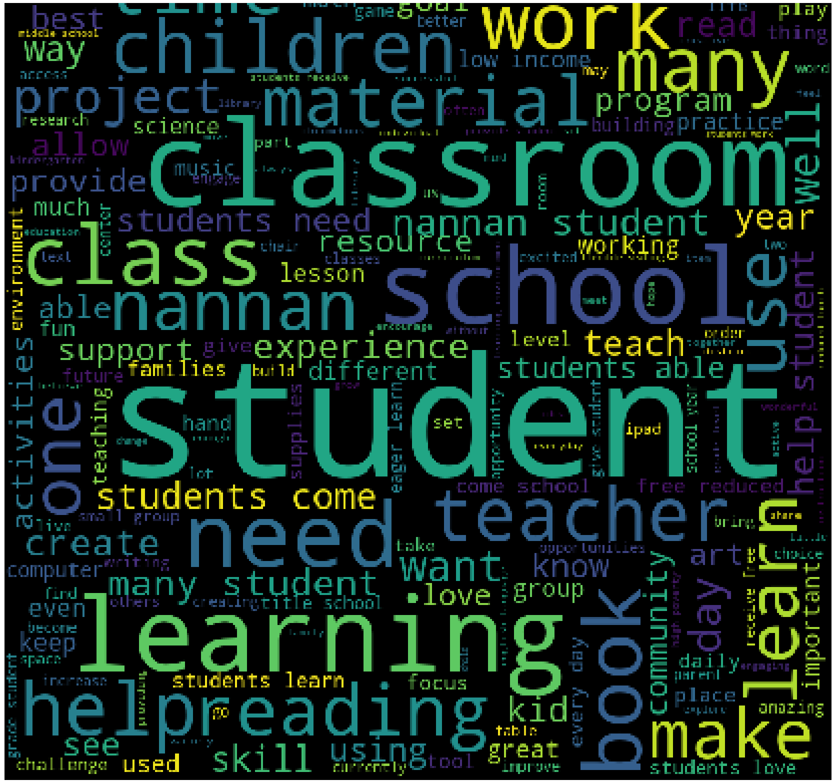


```
Total number of words in cluster 1 are 462009
Top 10 most common words in cluster 1 are [('students', 22110), ('i', 9
223), ('school', 7115), ('my', 5188), ('learning', 4964), ('classroom',
4703), ('the', 3748), ('they', 3641), ('not', 3620), ('learn', 3535)]
```

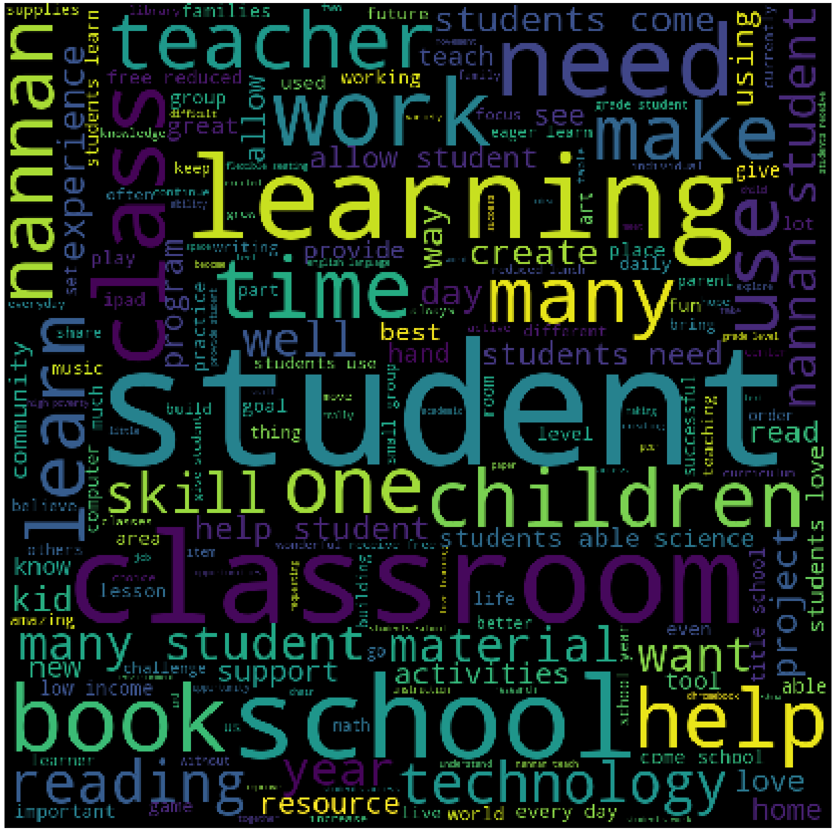
```
Word cloud plot for essay text for cluster 2 for Agglomerative algorith
m
```



```
Total number of words in cluster 2 are 583018
Top 10 most common words in cluster 2 are [('students', 27791), ('i', 1
1746), ('school', 8997), ('my', 6507), ('learning', 6264), ('classroo
m', 6069), ('the', 4761), ('not', 4664), ('they', 4655), ('learn', 442
8)]
Word cloud plot for essay text for cluster 3 for Agglomerative algorith
m
```



```
Total number of words in cluster 3 are 222219
Top 10 most common words in cluster 3 are [('students', 10757), ('i', 4
587), ('school', 3438), ('my', 2445), ('learning', 2409), ('classroom',
2305), ('the', 1840), ('not', 1808), ('learn', 1723), ('they', 1676)]
Word cloud plot for essay text for cluster 4 for Agglomerative algorith
m
```

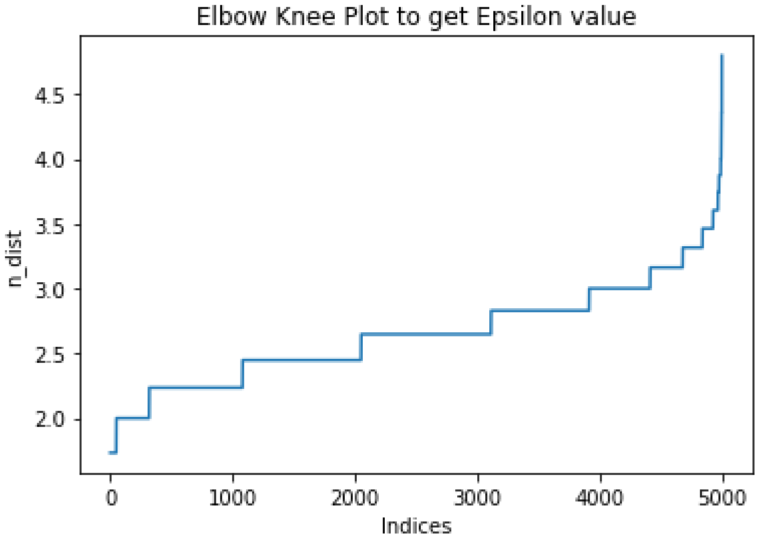


```
Total number of words in cluster 4 are 275590
Top 10 most common words in cluster 4 are [('students', 13248), ('i', 5
679), ('school', 4349), ('my', 3160), ('learning', 2903), ('classroom',
2888), ('the', 2241), ('learn', 2207), ('they', 2150), ('help', 2125)]
```

2.7 Apply DBSCAN

In [231]:

```
# https://stackoverflow.com/questions/12893492/choosing-eps-and-minpts-for-dbscan-r/48558030#48558030
# https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
# https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html
# https://ipfs-sec.stackexchange.cloudflare-ipfs.com/datascience/A/question/10162.html

X_tr_new = X_tr_new[:5000]

min_pts=10
n_dist=[]
distance = []

for val in X_tr_new:

    val_dist = np.sort(np.sum((X_tr_new.toarray()-val.toarray())**2, axis=1), axis=None)

    distance.append(val_dist[min_pts])

n_dist = np.sort(np.sqrt(np.array(distance)))

indices = [ind for ind in range(X_tr_new.shape[0])]

plt.plot(indices, n_dist)
plt.xlabel('Indices')
plt.ylabel('n_dist')
plt.title('Elbow Knee Plot to get Epsilon value')
plt.show()
```



```
In [283]: from sklearn.cluster import DBSCAN

          clustering = DBSCAN(eps=2.3)
          clustering.fit(X_tr_new)

Out[283]: DBSCAN(algorithm='auto', eps=2.3, leaf_size=30, metric='euclidean',
              metric_params=None, min_samples=5, n_jobs=1, p=None)

In [284]: clustering.labels_

Out[284]: array([-1,  0,  0, ...,  0, -1, -1], dtype=int64)

In [285]: # https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.htm
          l
          # Number of clusters in labels, ignoring noise if present.

          n_clusters_ = len(set(clustering.labels_)) - (1 if -1 in clustering.lab
          els_ else 0)
          n_noise_ = list(clustering.labels_).count(-1)

          print('Estimated number of clusters: %d' % n_clusters_)
          print('Estimated number of noise points: %d' % n_noise_)
```

```
Estimated number of clusters: 4
Estimated number of noise points: 2177
```

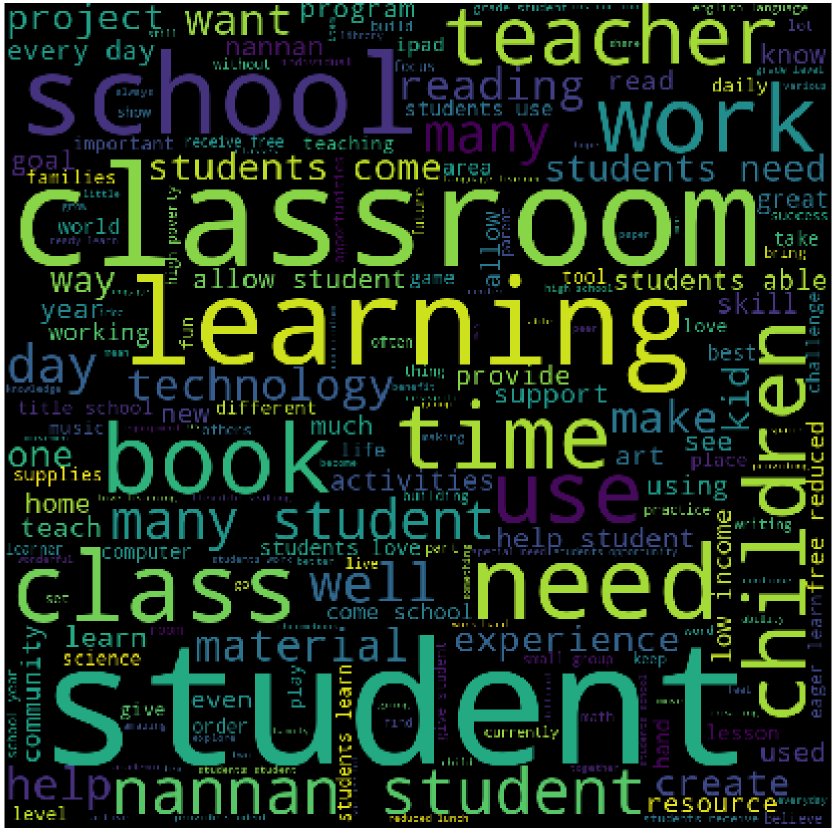
In [286]:

```
# Creating different clusters

cluster_0 = []
cluster_1 = []
cluster_2 = []
cluster_3 = []

for val in range(0, len(clustering.labels_)):

    if (clustering.labels_[val] == 0):
        cluster_0.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 1):
        cluster_1.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 2):
        cluster_2.append(preprocessed_essays[val])

    if (clustering.labels_[val] == 3):
        cluster_3.append(preprocessed_essays[val])
```

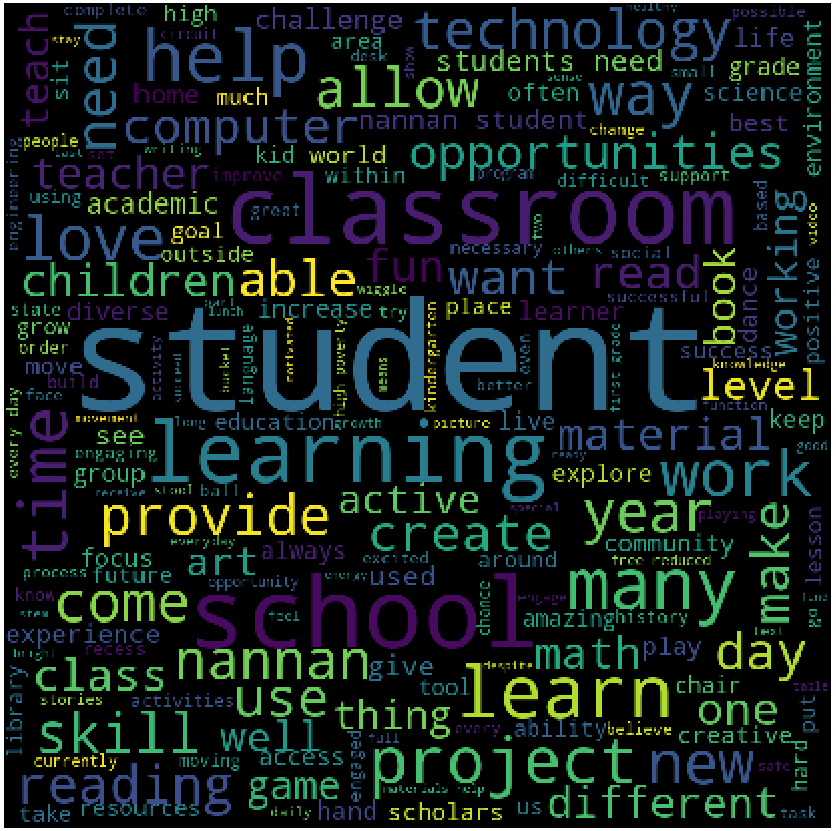
In [288]:

```
cluster = [cluster_0, cluster_1, cluster_2, cluster_3]
count = 0

for val in cluster:

    create_wordcloud(val, count, "DBSCAN")
    wordcloud_analysis(val, count)
    count +=1
```

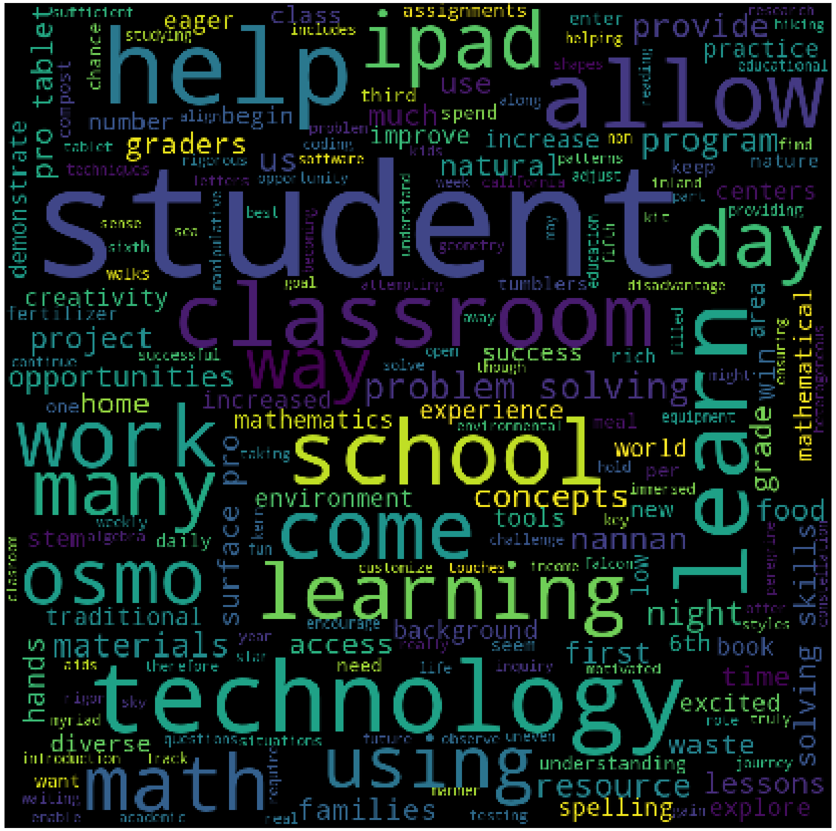
```
Word cloud plot for essay text for cluster 0 for DBSCAN algorithm
```



```
Total number of words in cluster 0 are 420306
Top 10 most common words in cluster 0 are [('students', 20005), ('i',
8370), ('school', 6655), ('my', 4675), ('classroom', 4486), ('learnin
g', 4460), ('the', 3532), ('they', 3340), ('not', 3228), ('learn', 31
76)]
Word cloud plot for essay text for cluster 1 for DBSCAN algorithm
```

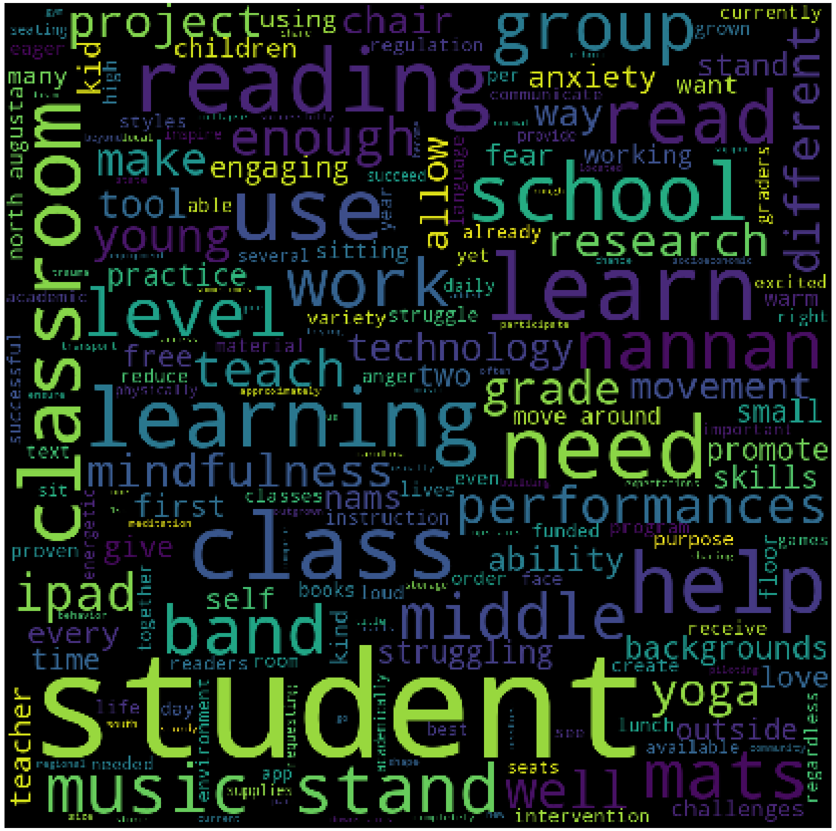


```
Total number of words in cluster 1 are 7191
Top 10 most common words in cluster 1 are [('students', 341), ('i', 1
43), ('school', 98), ('classroom', 81), ('my', 80), ('learning', 69),
('they', 65), ('the', 62), ('not', 55), ('learn', 55)]
Word cloud plot for essay text for cluster 2 for DBSCAN algorithm
```



```
Total number of words in cluster 2 are 593
Top 10 most common words in cluster 2 are [('students', 31), ('help',
11), ('technology', 10), ('learn', 10), ('allow', 9), ('the', 8), ('c
lassroom', 8), ('school', 8), ('come', 7), ('many', 6)]
```

```
lassroom', 8), ('school', 8), ('come', 7), ('many', 6)]
Word cloud plot for essay text for cluster 3 for DBSCAN algorithm
```



```
Total number of words in cluster 3 are 931
Top 10 most common words in cluster 3 are [('students', 43), ('i', 3
1), ('class', 11), ('learn', 11), ('reading', 10), ('learning', 9),
('use', 9), ('need', 9), ('help', 9), ('school', 8)]
```

3. Conclusions

Please write down few lines of your observations on this assignment.

In [19]:

```
# http://zetcode.com/python/prettytable/

from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["Model Name", "Hyperparameters", "Hyperparameter Value
 1", "Hyperparameter Value 2", "Number of Clusters"]

x.add_row(["Kmeans", "n_clusters/max_iter", 8, 300, 8])
x.add_row(["AgglomerativeClustering", "n_clusters", 2, "na", 2])
x.add_row(["AgglomerativeClustering", "n_clusters", 5, "na", 5])
x.add_row(["DBSCAN", "minpts/eps", 10, 2.3, 4])
print(x)
```

```
+-------------------------+---------------------+----------------------
--+------------------------+--------------------+
|        Model Name       |   Hyperparameters   | Hyperparameter Value
1 | Hyperparameter Value 2 | Number of Clusters |
+-------------------------+---------------------+----------------------
--+------------------------+--------------------+
|          Kmeans         | n_clusters/max_iter |           8          
  |          300           |         8          |
| AgglomerativeClustering |      n_clusters     |           2          
  |           na           |         2          |
| AgglomerativeClustering |      n_clusters      |           5
  |           na           |         5          |
|          DBSCAN          |      minpts/eps      |           10
  |          2.3           |         4          |
+-------------------------+---------------------+----------------------
--+------------------------+--------------------+
```