DonorsChoose

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

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

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

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

About the DonorsChoose Data Set

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

Feature	Description
`project_id`	A unique identifier for the proposed project. **Example:** `p036502`
`project_title`	Title of the project. **Examples:** • `Art Will Make You Happy!` • `First Grade Fun`
`project_grade_category`	Grade level of students for which the project is targeted. One of the following enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`
`project_subject_categories`	One or more (comma-separated) subject categories for the project from the following enumerated list of values: • `Applied Learning` • `Care & Hunger` • `Health & Sports` • `History & Civics` • `Literacy & Language` • `Math & Science` • `Music & The Arts` • `Special Needs` • `Warmth` **Examples:** • `Music & The Arts` • `Literacy & Language, Math & Science`
`school_state`	State where school is located (Two-letter U.S. postal code (https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations#Postal_codes)). **Example:** `WY`
`project_subject_subcategories`	One or more (comma-separated) subject subcategories for the project. **Examples:** • `Literacy` • `Literature & Writing, Social Sciences`
`project_resource_summary`	An explanation of the resources needed for the project. **Example:** • `My students need hands on literacy materials to manage sensory needs!`
`project_essay_1`	First application essay*
`project_essay_2`	Second application essay*
`project_essay_3`	Third application essay*
`project_essay_4`	Fourth application essay*
`project_submitted_datetime`	Datetime when project application was submitted. **Example:** `2016-04-28 12:43:56.245`
`teacher_id`	A unique identifier for the teacher of the proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c15c56`
`teacher_prefix`	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
`teacher_number_of_previously_posted_projects`	Number of project applications previously submitted by the same teacher. **Example:** `2`

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

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

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

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

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

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

Notes on the Essay Data

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

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

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

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

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

In [1]:
```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
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
!pip install chart_studio
from chart_studio import plotly
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
from collections import Counter
```

```
Requirement already satisfied: chart_studio in /usr/local/lib/python3.6/dist-packages (1.1.0)
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from chart_studio) (1.12.0)
Requirement already satisfied: plotly in /usr/local/lib/python3.6/dist-packages (from chart_studio) (4.4.1)
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (from chart_studio) (2.23.0)
Requirement already satisfied: retrying>=1.3.3 in /usr/local/lib/python3.6/dist-packages (from chart_studio) (1.3.
3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packages (from requests->chart_
studio) (2020.4.5.1)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.6/dist-packages
(from requests->chart_studio) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-packages (from requests->chart_studi
o) (2.9)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/dist-packages (from requests->chart_s
tudio) (3.0.4)
```

1.1 Loading Data

In [0]:
```
project_data = pd.read_csv('/content/drive/My Drive/Assignments_DonorsChoose_2018/train_data.csv')
resource_data = pd.read_csv('/content/drive/My Drive/Assignments_DonorsChoose_2018/resources.csv')
```

In [3]:

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

```
Number of data points in train data (109248, 17)
--------------------------------------------------
The attributes of data : ['Unnamed: 0' 'id' 'teacher_id' 'teacher_prefix' 'school_state'
 'project_submitted_datetime' 'project_grade_category'
 'project_subject_categories' 'project_subject_subcategories'
 'project_title' 'project_essay_1' 'project_essay_2' 'project_essay_3'
 'project_essay_4' 'project_resource_summary'
 'teacher_number_of_previously_posted_projects' 'project_is_approved']
```

In [4]:

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

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

Out[4]:

	id	description	quantity	price
0	p233245	LC652 - Lakeshore Double-Space Mobile Drying Rack	1	149.00
1	p069063	Bouncy Bands for Desks (Blue support pipes)	3	14.95

1.2 Preprocessing Categorical Data

1.2.1 preprocessing `project_subject_categories`

In [0]:

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

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python
cat_list = []
for i in catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & 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 'Th
e')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Scienc
e"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())

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

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

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

In [6]:

```
sorted_cat_dict.keys()
```

Out[6]:

```
dict_keys(['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds', 'Health_Spo
rts', 'Math_Science', 'Literacy_Language'])
```

1.2.2 preprocessing of `project_subject_subcategories`

```
In [0]: 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 'Th
        e')
                j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Scienc
        e"
                temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
                temp = temp.replace('&','_') # we are replacing the & value into
            sub_cat_list.append(temp.strip())

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

        from collections import Counter
        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]))
```

```
In [8]: sorted_sub_cat_dict.keys()
```

```
Out[8]: dict_keys(['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular', 'Civics_G
        overnment', 'ForeignLanguages', 'NutritionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingArts',
        'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPrep', 'Music', 'History_Geography', 'Health_LifeScien
        ce', 'EarlyDevelopment', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedSc
        iences', 'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Literacy'])
```

1.2.3 preprocessing of `School State`

```
In [9]: project_data['school_state'].unique()
```

```
Out[9]: array(['IN', 'FL', 'AZ', 'KY', 'TX', 'CT', 'GA', 'SC', 'NC', 'CA', 'NY',
               'OK', 'MA', 'NV', 'OH', 'PA', 'AL', 'LA', 'VA', 'AR', 'WA', 'WV',
               'ID', 'TN', 'MS', 'CO', 'UT', 'IL', 'MI', 'HI', 'IA', 'RI', 'NJ',
               'MO', 'DE', 'MN', 'ME', 'WY', 'ND', 'OR', 'AK', 'MD', 'WI', 'SD',
               'NE', 'NM', 'DC', 'KS', 'MT', 'NH', 'VT'], dtype=object)
```

```
In [10]: project_data['school_state'][project_data['school_state'].isnull()==True]
```

```
Out[10]: Series([], Name: school_state, dtype: object)
```

```
In [0]: # count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
        my_counter = Counter()
        for word in project_data['school_state'].values:
            my_counter.update(word.split())

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

```
In [12]: sorted_school_state_dict.keys()
```

```
Out[12]: dict_keys(['VT', 'WY', 'ND', 'MT', 'RI', 'SD', 'NE', 'DE', 'AK', 'NH', 'WV', 'ME', 'HI', 'DC', 'NM', 'KS', 'IA',
         'ID', 'AR', 'CO', 'MN', 'OR', 'KY', 'MS', 'NV', 'MD', 'CT', 'TN', 'UT', 'AL', 'WI', 'VA', 'AZ', 'NJ', 'OK', 'WA',
         'MA', 'LA', 'OH', 'MO', 'IN', 'PA', 'MI', 'SC', 'GA', 'IL', 'NC', 'FL', 'NY', 'TX', 'CA'])
```

1.2.4 preprocessing of `Teacher Prefix`

```
In [13]: project_data.groupby(['teacher_prefix'])['teacher_prefix'].count()
```

```
Out[13]: teacher_prefix
         Dr.            13
         Mr.         10648
         Mrs.        57269
         Ms.         38955
         Teacher      2360
         Name: teacher_prefix, dtype: int64
```

In [14]:

```
project_data['teacher_prefix'][project_data['teacher_prefix'].isnull()==True]
```

Out[14]:

```
7820     NaN
30368    NaN
57654    NaN
Name: teacher_prefix, dtype: object
```

In [0]:

```
project_data['teacher_prefix'].fillna(project_data['teacher_prefix'].mode()[0],inplace=True)
```

In [16]:

```
project_data['teacher_prefix'][project_data['teacher_prefix'].isnull()==True]
```

Out[16]:

```
Series([], Name: teacher_prefix, dtype: object)
```

In [17]:

```
project_data['teacher_prefix'].unique()
```

Out[17]:

```
array(['Mrs.', 'Mr.', 'Ms.', 'Teacher', 'Dr.'], dtype=object)
```

In [0]:

```
teacher_prefix = list(project_data['teacher_prefix'].values)

teacher_prefix_list = []
for i in teacher_prefix:
    temp = ""
    temp = i.split('.')
    temp = i.replace('.','')
    teacher_prefix_list.append(temp)

project_data['clean_teacher_prefix'] = teacher_prefix_list
project_data.drop(['teacher_prefix'], 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_teacher_prefix'].values:
    my_counter.update(word.split())

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

In [19]:

```
sorted_teacher_prefix_dict.keys()
```

Out[19]:

```
dict_keys(['Dr', 'Teacher', 'Mr', 'Ms', 'Mrs'])
```

In [20]:

```
project_data.groupby(['clean_teacher_prefix'])['clean_teacher_prefix'].count()
```

Out[20]:

```
clean_teacher_prefix
Dr            13
Mr         10648
Mrs        57272
Ms         38955
Teacher     2360
Name: clean_teacher_prefix, dtype: int64
```

1.2.5 preprocessing of `Project Grade Category`

In [21]:

```
project_data.groupby(['project_grade_category'])['project_grade_category'].count()
```

Out[21]:

```
project_grade_category
Grades 3-5       37137
Grades 6-8       16923
Grades 9-12      10963
Grades PreK-2    44225
Name: project_grade_category, dtype: int64
```

In [22]:

```
project_data['project_grade_category'][project_data['project_grade_category'].isnull()==True]
```

Out[22]:

```
Series([], Name: project_grade_category, dtype: object)
```

In [0]:

```
project_grade_category = list(project_data['project_grade_category'].values)

project_grade_category_list = []
for i in project_grade_category:
    temp = ""
    temp = i.split(' ')
    temp = i.replace('Grades ','')
    project_grade_category_list.append(temp)

project_data['clean_project_grade_category'] = project_grade_category_list
project_data.drop(['project_grade_category'], 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_project_grade_category'].values:
    my_counter.update(word.split())

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

In [24]:

```
sorted_project_grade_category_dict.keys()
```

Out[24]:

```
dict_keys(['9-12', '6-8', '3-5', 'PreK-2'])
```

In [25]:

```
project_data.groupby(['clean_project_grade_category'])['clean_project_grade_category'].count()
```

Out[25]:

```
clean_project_grade_category
3-5       37137
6-8       16923
9-12      10963
PreK-2    44225
Name: clean_project_grade_category, dtype: int64
```

1.3 Text Preprocessing of project_essay

In [0]:

```
# 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 [27]:

```
project_data.head(1)
```

Out[27]:

	Unnamed: 0	id	teacher_id	school_state	project_submitted_datetime	project_title	project_essay_1	project_essay
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	IN	2016-12-05 13:43:57	Educational Support for English Learners at Home	My students are English learners that are work...	\"The limits your langua are the limits

In [0]:

```
# 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 [29]:

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

```
My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gros
s/fine motor delays, to autism. They are eager beavers and always strive to work their hardest working past their
limitations. \r\n\r\nThe materials we have are the ones I seek out for my students. I teach in a Title I school wh
ere most of the students receive free or reduced price lunch.  Despite their disabilities and limitations, my stud
ents love coming to school and come eager to learn and explore.Have you ever felt like you had ants in your pants
and you needed to groove and move as you were in a meeting? This is how my kids feel all the time. 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, which enhances gross motor and in Turn fine motor skills. \r\nThey also want to learn through games, my kids
do not want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is t
he 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 [30]:

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

```
My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gros
s/fine motor delays, to autism. They are eager beavers and always strive to work their hardest working past their
limitations.     The materials we have are the ones I seek out for my students. I teach in a Title I school where
most of the students receive free or reduced price lunch.  Despite their disabilities and limitations, my students
love coming to school and come eager to learn and explore.Have you ever felt like you had ants in your pants and y
ou 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 develop their core,
which enhances gross motor and in Turn fine motor skills.   They also want to learn through games, my kids do not
want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is the key
to our success. The number toss and color and shape mats can make that happen. My students will forget they are do
ing work and just have the fun a 6 year old deserves.nannan
```

In [31]:

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

```
My kindergarten students have varied disabilities ranging from speech and language delays cognitive delays gross f
ine motor delays to autism They are eager beavers and always strive to work their hardest working past their limit
ations The materials we have are the ones I seek out for my students I teach in a Title I school where most of the
students receive free or reduced price lunch Despite their disabilities and limitations my students love coming to
school and come eager to learn and explore Have you ever felt like you had ants in your pants and you needed to gr
oove 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 l
earn or so they say Wobble chairs are the answer and I love then because they develop their core which enhances gr
oss motor and in Turn fine motor skills They also want to learn through games my kids do not want to sit and do wo
rksheets They want to learn to count by jumping and playing Physical engagement is the key to our success The numb
er toss and color and shape mats can make that happen My students will forget they are doing work and just have th
e fun a 6 year old deserves nannan
```

In [0]:

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

In [33]:
```
# 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%|██████████| 109248/109248 [01:02<00:00, 1735.30it/s]
```

In [34]:
```
# after preprocesing
preprocessed_essays[20000]
```

Out[34]:
```
'my kindergarten students varied disabilities ranging speech language delays cognitive delays gross fine motor del
ays autism they eager beavers always strive work hardest working past limitations the materials ones i seek studen
ts i teach title i school students receive free reduced price lunch despite disabilities limitations students love
coming school come eager learn explore have ever felt like ants pants needed groove move meeting this kids feel ti
me the want able move learn say wobble chairs answer i love develop core enhances gross motor turn fine motor skil
ls they also want learn games kids not want sit worksheets they want learn count jumping playing physical engageme
nt key success the number toss color shape mats make happen my students forget work fun 6 year old deserves nanna
n'
```

In [0]:
```
project_data['preprocessed_essays'] = preprocessed_essays
project_data.drop(['essay'], axis=1, inplace=True)
```

1.4 Preprocessing of `project_title`

In [36]:
```
project_data['project_title'][2000:2010]
```

Out[36]:
```
2000                    Steady Stools for Active Learning
2001                                   Classroom Supplies
2002    Kindergarten Students Deserve Quality  Books a...
2003                                 Listen to Understand!
2004                              iPads to iGnite Learning
2005                                  Tablets For Learning
2006                                              Go P.E.!
2007                                    Making Learning Fun!
2008    Empowerment Through Silk Screen Designed Tee S...
2009                                   Let's Play Together!
Name: project_title, dtype: object
```

In [37]:
```
# Combining all the above statemennts
from tqdm import tqdm
preprocessed_titles = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['project_title'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_titles.append(sent.lower().strip())
```

```
100%|██████████| 109248/109248 [00:02<00:00, 41572.97it/s]
```

In [38]:
```
preprocessed_titles[2000:2010]
```

Out[38]:
```
['steady stools active learning',
 'classroom supplies',
 'kindergarten students deserve quality books vibrant rug',
 'listen understand',
 'ipads ignite learning',
 'tablets for learning',
 'go p e',
 'making learning fun',
 'empowerment through silk screen designed tee shirts',
 'let play together']
```

In [0]:
```
project_data['preprocessed_titles'] = preprocessed_titles
project_data.drop(['project_title'], axis=1, inplace=True)
```

1.5 Merging Numerical data in Resources to project_data

In [0]:

```python
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 [41]:

```python
project_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109248 entries, 0 to 109247
Data columns (total 20 columns):
 #   Column                                        Non-Null Count   Dtype  
---  ------                                        --------------   -----  
 0   Unnamed: 0                                    109248 non-null  int64  
 1   id                                            109248 non-null  object 
 2   teacher_id                                    109248 non-null  object 
 3   school_state                                  109248 non-null  object 
 4   project_submitted_datetime                    109248 non-null  object 
 5   project_essay_1                               109248 non-null  object 
 6   project_essay_2                               109248 non-null  object 
 7   project_essay_3                               3758 non-null    object 
 8   project_essay_4                               3758 non-null    object 
 9   project_resource_summary                      109248 non-null  object 
 10  teacher_number_of_previously_posted_projects  109248 non-null  int64  
 11  project_is_approved                           109248 non-null  int64  
 12  clean_categories                              109248 non-null  object 
 13  clean_subcategories                           109248 non-null  object 
 14  clean_teacher_prefix                          109248 non-null  object 
 15  clean_project_grade_category                  109248 non-null  object 
 16  preprocessed_essays                           109248 non-null  object 
 17  preprocessed_titles                           109248 non-null  object 
 18  price                                         109248 non-null  float64
 19  quantity                                      109248 non-null  int64  
dtypes: float64(1), int64(4), object(15)
memory usage: 17.5+ MB
```

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
- Essay : text data

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

In [0]:

```python
data1 = project_data.drop(['Unnamed: 0', 'id','project_submitted_datetime','project_essay_1','project_essay_2','project_essay_3','project_essay_4','project_resource_summary','teacher_id'], axis = 1)
```

In [43]:

```python
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109248 entries, 0 to 109247
Data columns (total 11 columns):
 #   Column                                        Non-Null Count   Dtype  
---  ------                                        --------------   -----  
 0   school_state                                  109248 non-null  object 
 1   teacher_number_of_previously_posted_projects  109248 non-null  int64  
 2   project_is_approved                           109248 non-null  int64  
 3   clean_categories                              109248 non-null  object 
 4   clean_subcategories                           109248 non-null  object 
 5   clean_teacher_prefix                          109248 non-null  object 
 6   clean_project_grade_category                  109248 non-null  object 
 7   preprocessed_essays                           109248 non-null  object 
 8   preprocessed_titles                           109248 non-null  object 
 9   price                                         109248 non-null  float64
 10  quantity                                      109248 non-null  int64  
dtypes: float64(1), int64(3), object(7)
memory usage: 10.0+ MB
```

Train test split

In [0]:

```
# train test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data1, data1['project_is_approved'], test_size=0.33, stratify=d
ata1['project_is_approved'])
```

In [0]:

```
#Features
X_train.drop(['project_is_approved'], axis=1, inplace=True)

X_test.drop(['project_is_approved'], axis=1, inplace=True)
```

In [46]:

```
X_train.head()
```

Out[46]:

	school_state	teacher_number_of_previously_posted_projects	clean_categories	clean_subcategories	clean_teacher_prefix	clean_project
51345	IL	3	Literacy_Language History_Civics	Literature_Writing SocialSciences	Mrs	
89329	TX	0	Literacy_Language	Literature_Writing	Mrs	
100922	NJ	0	SpecialNeeds	SpecialNeeds	Mrs	
70847	MN	6	Math_Science Literacy_Language	EnvironmentalScience Literacy	Mrs	
55269	MA	0	SpecialNeeds	SpecialNeeds	Ms	

1.6 Make Data Model Ready: encoding essay, and project_title

In [0]:

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

# when you plot any graph make sure you use
    # a. Title, that describes your plot, this will be very helpful to the reader
    # b. Legends if needed
    # c. X-axis label
    # d. Y-axis label
```

1.6.1 TF IDF Essay and Title

1.6.1.1 TF IDF Essay

In [48]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

print("="*100)


vectorizer = TfidfVectorizer()
vectorizer.fit(X_train['preprocessed_essays'].values) # fit has to happen only on train data

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

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

```
(73196, 10) (73196,)
(36052, 10) (36052,)
====================================================================================================
After vectorizations
(73196, 48269) (73196,)
(36052, 48269) (36052,)
====================================================================================================
```

1.6.1.2 TF IDF Title

In [49]:

```
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

print("="*100)


vectorizer = TfidfVectorizer()
vectorizer.fit(X_train['preprocessed_titles'].values) # fit has to happen only on train data

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

print("After vectorizations")
print(X_train_title_tfidf.shape, y_train.shape)
print(X_test_title_tfidf.shape, y_test.shape)
print("="*100)
```

```
(73196, 10) (73196,)
(36052, 10) (36052,)
====================================================================================================
After vectorizations
(73196, 14121) (73196,)
(36052, 14121) (36052,)
====================================================================================================
```

1.7 Make Data Model Ready: encoding numerical, categorical features

In [0]:

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

# when you plot any graph make sure you use
    # a. Title, that describes your plot, this will be very helpful to the reader
    # b. Legends if needed
    # c. X-axis label
    # d. Y-axis label
```

1.7.1 Numerical features

1. teacher_number_of_previously_posted_projects
2. price
3. quantity

1.7.1.1 Teacher number of previously posted projects

In [51]:

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

X_train_TPPP_norm = normalizer.transform(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-
1))
X_test_TPPP_norm = normalizer.transform(X_test['teacher_number_of_previously_posted_projects'].values.reshape(1,-1
))

print("After vectorizations")
print(X_train_TPPP_norm.shape, y_train.shape)
print(X_test_TPPP_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(1, 73196) (73196,)
(1, 36052) (36052,)
====================================================================================================
```

In [52]:

```
print("Transpose of teacher number of previously posted projects")

X_train_TPPP_norm = X_train_TPPP_norm.transpose()
X_test_TPPP_norm = X_test_TPPP_norm.transpose()

print("After transpose")
print(X_train_TPPP_norm.shape, y_train.shape)
print(X_test_TPPP_norm.shape, y_test.shape)
print("="*100)
```

```
Transpose of teacher number of previously posted projects
After transpose
(73196, 1) (73196,)
(36052, 1) (36052,)
====================================================================================================
```

1.7.1.2 price

In [53]:

```
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))

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

```
After vectorizations
(1, 73196) (73196,)
(1, 36052) (36052,)
====================================================================================================
```

In [54]:

```
print("Transpose of price")

X_train_price_norm = X_train_price_norm.transpose()
X_test_price_norm = X_test_price_norm.transpose()


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

```
Transpose of price
After vectorizations
(73196, 1) (73196,)
(36052, 1) (36052,)
====================================================================================================
```

1.7.1.3 quantity

In [55]:

```
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['quantity'].values.reshape(1,-1))

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

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

```
After vectorizations
(1, 73196) (73196,)
(1, 36052) (36052,)
====================================================================================================
```

In [56]:

```
print("Transpose of Quantity")

X_train_quantity_norm = X_train_quantity_norm.transpose()
X_test_quantity_norm = X_test_quantity_norm.transpose()


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

```
Transpose of Quantity
After vectorizations
(73196, 1) (73196,)
(36052, 1) (36052,)
====================================================================================================
```

1.7.2 Categorical Data

Categorical Features for vectorization

1. Clean Categories
2. Clean Sub Categories
3. School State
4. Teacher Prefix
5. Project grade category

1.7.2.1 Clean Categories

In [57]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(X_train['clean_categories'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_CC_ohe = vectorizer.transform(X_train['clean_categories'].values)
X_test_CC_ohe = vectorizer.transform(X_test['clean_categories'].values)

print("After vectorizations")
print(X_train_CC_ohe.shape, y_train.shape)
print(X_test_CC_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(73196, 9) (73196,)
(36052, 9) (36052,)
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds', 'Health_Sports', 'Mat
h_Science', 'Literacy_Language']
====================================================================================================
```

1.7.2.2 Clean Sub Categories

In [58]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(X_train['clean_subcategories'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_CSC_ohe = vectorizer.transform(X_train['clean_subcategories'].values)
X_test_CSC_ohe = vectorizer.transform(X_test['clean_subcategories'].values)

print("After vectorizations")
print(X_train_CSC_ohe.shape, y_train.shape)
print(X_test_CSC_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(73196, 30) (73196,)
(36052, 30) (36052,)
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular', 'Civics_Governmen
t', 'ForeignLanguages', 'NutritionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingArts', 'Charac
terEducation', 'TeamSports', 'Other', 'College_CareerPrep', 'Music', 'History_Geography', 'Health_LifeScience', 'E
arlyDevelopment', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedScience
s', 'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Literacy']
====================================================================================================
```

1.7.2.3 School State

In [59]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_school_state_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(X_train['school_state'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_state_ohe = vectorizer.transform(X_train['school_state'].values)
X_test_state_ohe = vectorizer.transform(X_test['school_state'].values)

print("After vectorizations")
print(X_train_state_ohe.shape, y_train.shape)
print(X_test_state_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(73196, 51) (73196,)
(36052, 51) (36052,)
['VT', 'WY', 'ND', 'MT', 'RI', 'SD', 'NE', 'DE', 'AK', 'NH', 'WV', 'ME', 'HI', 'DC', 'NM', 'KS', 'IA', 'ID', 'AR',
'CO', 'MN', 'OR', 'KY', 'MS', 'NV', 'MD', 'CT', 'TN', 'UT', 'AL', 'WI', 'VA', 'AZ', 'NJ', 'OK', 'WA', 'MA', 'LA',
'OH', 'MO', 'IN', 'PA', 'MI', 'SC', 'GA', 'IL', 'NC', 'FL', 'NY', 'TX', 'CA']
====================================================================================================
```

1.7.2.4 Teacher prefix

In [60]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_teacher_prefix_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(X_train['clean_teacher_prefix'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_teacher_ohe = vectorizer.transform(X_train['clean_teacher_prefix'].values)
X_test_teacher_ohe = vectorizer.transform(X_test['clean_teacher_prefix'].values)

print("After vectorizations")
print(X_train_teacher_ohe.shape, y_train.shape)
print(X_test_teacher_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(73196, 5) (73196,)
(36052, 5) (36052,)
['Dr', 'Teacher', 'Mr', 'Ms', 'Mrs']
====================================================================================================
```

1.7.2.5 Project Grade category

In [61]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_project_grade_category_dict.keys()), lowercase=False, binary=True)
vectorizer.fit(X_train['clean_project_grade_category'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_grade_ohe = vectorizer.transform(X_train['clean_project_grade_category'].values)
X_test_grade_ohe = vectorizer.transform(X_test['clean_project_grade_category'].values)

print("After vectorizations")
print(X_train_grade_ohe.shape, y_train.shape)
print(X_test_grade_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(73196, 4) (73196,)
(36052, 4) (36052,)
['9-12', '6-8', '3-5', 'PreK-2']
====================================================================================================
```

Concatinating all the features

In [62]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr_TFIDF = hstack((X_train_essay_tfidf, X_train_title_tfidf, X_train_state_ohe, X_train_teacher_ohe, X_train_grade_ohe, X_train_CSC_ohe, X_train_CC_ohe, X_train_price_norm, X_train_quantity_norm, X_train_TPPP_norm)).tocsr()
X_te_TFIDF = hstack((X_test_essay_tfidf, X_test_title_tfidf, X_test_state_ohe, X_test_teacher_ohe, X_test_grade_ohe, X_test_CSC_ohe, X_test_CC_ohe, X_test_price_norm, X_test_quantity_norm, X_test_TPPP_norm)).tocsr()

print("Final Data matrix")
print(X_tr_TFIDF.shape, y_train.shape)
print(X_te_TFIDF.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(73196, 62492) (73196,)
(36052, 62492) (36052,)
====================================================================================================
```

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 (https://scikit-learn.org/stable/modules/feature_selection.html)/reduction algorithms (https://scikit-learn.org/stable/modules/decomposition.html) 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 (https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/) 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 (https://stackoverflow.com/a/48558030/4084039).
      ● 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.

Clustering

2.1 Choose the best data matrix on which you got the best AUC

In [1]:

```
# I used set1 features where text is represented as TFIDF Vectorization of essay and title
```

2.2 Make Data Model Ready: encoding numerical, categorical features</h2>

```
# Already done at top
```

2.3 Make Data Model Ready: encoding eassay, and project_title

```
# Already done at top
```

2.4 Selecting Best-k features (or) Dimensionality Reduction to get k-features

use only top 5000 Features using selectKbest

In [0]:

```
from sklearn.feature_selection import SelectKBest, chi2
t = SelectKBest(chi2,k=5000).fit(X_tr_TFIDF, y_train)
X_tr = t.transform(X_tr_TFIDF)
X_te = t.transform(X_te_TFIDF)
```

In [64]:

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

```
Final Data matrix on TFIDF
(73196, 5000) (73196,)
(36052, 5000) (36052,)
====================================================================================================
```

2.5 Apply Kmeans

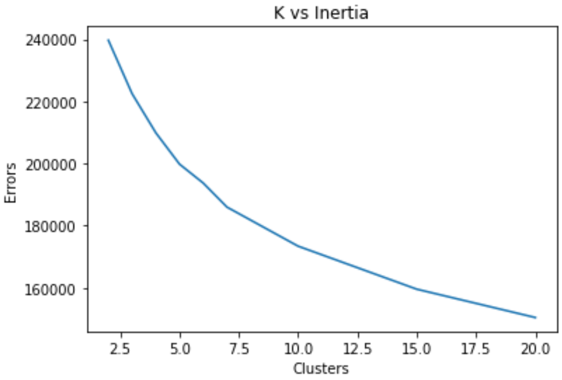
In [72]:

```
from sklearn.cluster import KMeans
clusters = [2, 3, 4, 5, 6, 7, 10, 15, 20]
loss = []
for i in tqdm(clusters):
    Kmean = KMeans(n_clusters=i,n_jobs=-1).fit(X_tr)
    loss.append(Kmean.inertia_)
```

```
100%|██████████| 9/9 [1:39:28<00:00, 663.21s/it]
```

In [74]:

```
plt.plot(clusters, loss)
plt.xlabel('Clusters')
plt.ylabel('Errors')
plt.title('K vs Inertia')
plt.show()
```



Optimal n_clusters

```
In [111]: from sklearn.cluster import KMeans
          optimal_k = 7
          kmeans = KMeans(n_clusters=optimal_k, n_jobs=-1)
          kmeans.fit(X_tr)

Out[111]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
                 n_clusters=7, n_init=10, n_jobs=-1, precompute_distances='auto',
                 random_state=None, tol=0.0001, verbose=0)
```

```
In [112]: kmeans.n_clusters

Out[112]: 7
```

```
In [113]: kmeans.labels_

Out[113]: array([0, 0, 3, ..., 5, 6, 6], dtype=int32)
```

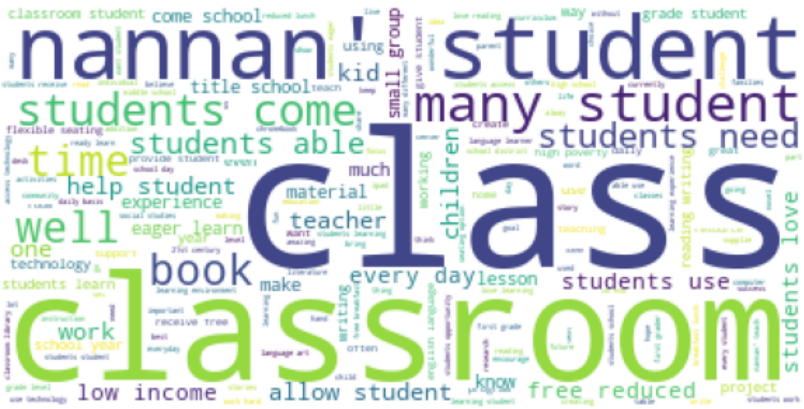
```
In [124]: cluster_dataset = {i:[] for i in range(optimal_k)}

          for index in tqdm(range(kmeans.labels_.shape[0])):
              cluster_dataset[kmeans.labels_[index]].append(X_train["preprocessed_essays"].iloc[index])

          print("length of each cluster:")
          for i in cluster_dataset:
              print(len(cluster_dataset[i]))

          100%|██████████| 73196/73196 [00:01<00:00, 55764.10it/s]

          length of each cluster:
          8801
          6115
          12572
          9160
          11143
          12011
          13394
```

Examining each cluster and its wordcloud

```
In [0]: from wordcloud import WordCloud, STOPWORDS
        stopwords = set(STOPWORDS)

        def examine_cluster_with_wordcloud(cluster_data):
            for paragraph in cluster_data[:3]:
                print(paragraph)
                print("-"*80)

            wordcloud = WordCloud(background_color='white', stopwords=stopwords, random_state=28).generate(str(cluster_data
        ))
            plt.figure(figsize = (8, 8), facecolor = None)
            plt.imshow(wordcloud)
            plt.axis("off")
            plt.tight_layout(pad = 0)
            plt.show()
            print("="*80)
```

```
In [126]: examine_cluster_with_wordcloud(cluster_dataset[0])
```

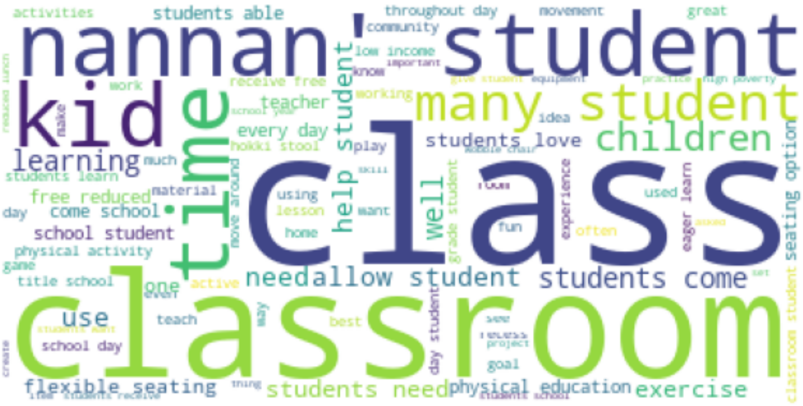
```
have ever made plans try new recipe create something you envisioned serving eating dish using new creation you may
even started creating dish project found pause project not ingredients materials needed do remember disappointment
frustration felt could not follow plan that sometimes happens students resources need succeed it amazing watch mid
dle school students grow see evidence growth collaborative conversations writing student products data they learni
ng set academic goals create plans utilize strategies meet goals unfortunately sometimes start moving full speed a
head toward achieving goals find not supplies resources need sadly lose momentum get back track success seventh ei
ghth grade students use interactive notebooks learn constitution prepare federal constitution test the notebooks u
sed take notes create study guides house foldable student products organizers student creations work social studie
s stations not interactive notebooks serve students well study constitution also serve portfolios the notebooks sh
owcase provide evidence students learning similarly sixth grade students use notebooks learning stations social st
udies learn writing traits process nannan
--------------------------------------------------------------------------------
my students kids when walk door reading language arts classroom i take ownership education well part lives they co
me variety levels finished harry potter transitioning picture books chapter books even reluctant readers avoid boo
ks costs my students best best or least i tell first day school i tell i secret share not tell anyone else school
sitting stool front room seeing attentive eyes eagerly waiting secret i tell hushed voice principal put best third
graders classroom i get privilege teaching best classes year it never fails live expectation any one one small gro
up instruction plus students but teacher 16 kids leading small group the answer engaging activity stations engaged
students busy focused learning become distraction small group students individual learners the reading writing act
ivity stations provide clear instructions easy student needed materials easy teacher after concept station introdu
ced class station added small group rotation time thus keeping 16 kids happily learning without teacher i instruct
small group 6 nannan
--------------------------------------------------------------------------------
i 25 inquisitive second graders we start day little breakfast great book who not love great book in order capture
interest attention students need books level we school three weeks many students read majority books level unfortu
nately we low income school could use help building library they great love learning given every opportunity achie
ve my 2nd graders able settle great book first thing morning they also able check book sparks interest enjoy outsi
de school reading needs part everyday home school lives i believe help build great connection all students enjoy r
eading books able read interested reading core learning therefore variety leveled books read create encouraging le
arning environment help us create stronger readers nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

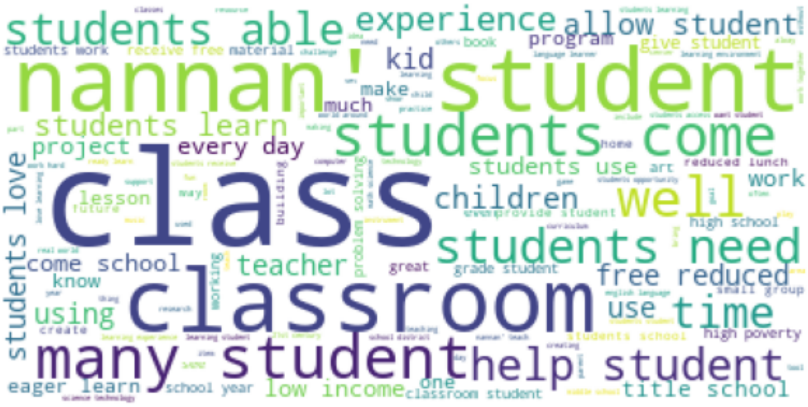
In [127]:
```
examine_cluster_with_wordcloud(cluster_dataset[1])
```

```
my students eclectic mix movers shakers they full energy energy not always reserved playground this active group 3
4 students insists learning easier entire body involved they like freedom wiggle work my students come variety soc
ioeconomic backgrounds ethnicities many school one constant lives no matter home life like one deserves opportunit
y learn environment welcoming encouraging as students become focused comfortable become productive reach highest p
otential potential sometimes surprises even my students told hard sit still work some described human bouncy balls
wanted know important sit still i told 34 kids classroom not much sitting still part important seat seat respect s
pace i thought discussion days later class meeting group students asked could get wiggle seats like seen classroom
s i replied great idea asked one going pay chairs because great items classroom generously donated donorschoose su
ggested submit project asking chairs share brilliant idea i said we spent time looking amazing chairs allow people
move work decided buoy chairs sturdy enough nine year old move like wind it proven active seating improves posture
strengthens core abdominal trunk muscles students constantly motion this also beneficial increasing strength muscl
e tone promotes motor control use pencil scissors classroom tools it wonderful give student opportunity get bodies
healthy minds work nannan
--------------------------------------------------------------------------------
my students diverse group i many kids would benefit movement class special education students adhd students medica
l needs students often students cannot stay focused movement keep going i often students not speak english movemen
t way connect some students cannot afford join sports project help day day reading writing often kids sitting long
periods time my class came idea get equipment room would help move healthy reading writing musical mats anyone we
saw mats thought fun would musical mat activity get moving thinking we could play music music stops whatever mat t
ells it might jog place hop my kids would never sitting long my kids bike pedals would love try elliptical ones th
is awesome way keep legs moving imagine kids room instead sitting chairs those students need hard space work added
trays work i cannot wait see working mats though please consider helping us add items room nannan
--------------------------------------------------------------------------------
my students kids i like call come everyday chip shoulder daring teach learning today important life so everyday i
motivate inspire kids all i want kids say done best my students majority native hawaiians pacific islanders the co
mmunities students come considered low socio economic status i feel students mentally tough come school ready work
they little rough around edges show teacher they anything many come rough tough home i teaching school since 2005
2006 school year our school operation since 2002 2003 school year still growing we consistently trying improve bet
ter school can i drink water normal question students i receive daily my students walk around hydro flasks filled
water carry around powerade drinks when students classroom ask drink water options limited they either drink water
sink classroom sink bathroom water fountain couple minutes away class all three options not viable i student i wou
ld not drink sink bathroom for project i requesting bottom load water cooler this water cooler located classroom s
tudents allowed fill containers water everyday i hoping water cooler encourage students drink water daily help cut
back sugary drinks nannan
--------------------------------------------------------------------------------
```



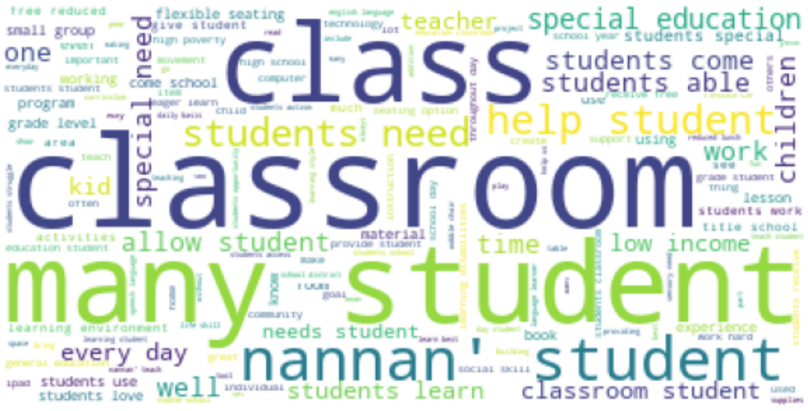
```
================================================================================
In [128]: examine_cluster_with_wordcloud(cluster_dataset[2])
```

```
as teacher low income high poverty school district students faced many challenges classroom despite many challenge
s face i looking help engage learning exploring creativity us history they learn best active participation movemen
t around classroom hands activities many challenges face may prevent getting ahead early life necessary supplies f
rom minute walk classroom i focus potential growth i may not able control home lives however i control experience 
school day help ignite love learning especially american history by creative positive way i hopeful inspire even e
arliest learners continue path academic excellence my students materials needed participate active engaged learnin
g activities the students use paper creating manipulatives help learn us history the paper also used differentiate
instruction according student needs the markers glue scissors enable develop projects enable explore creativity br
ing history life the pens pencils provide students may not necessary materials class ability fully participate as
malala yousafzai says one child one teacher one book one pen change world but first need book pens nannan
--------------------------------------------------------------------------------
21st century students need access 21st century technology we not resources need provide challenged risk students t
echnology individualize learning increase student engagement the majority students black latino english language l
earners mild severe learning disabilities 90 receive free reduced price lunch they live shelters parents incarcera
ted siblings gangs adversity aspects lives yet despite obstacles students attendance daily ready motivated learn s
ucceed desire provide opportunities change lives we use imac classroom write record short films math common core m
ath concepts in videos students act real world math problems they also create videos teach classmates major math c
oncepts emphasizing ways talk math problems strategies used find solutions students also use imac access digital g
oogle classroom students access complete assignments online watch video lessons extra support sign small group hel
p needed students able use wireless mouse presenting class time student engagement improve use technology sparks i
nterest emphasizes student voice student thinking student centered classroom students learn researching skills app
ropriate internet use proper technology use keyboarding development using technology educational purposes well int
ernet safety rules nannan
--------------------------------------------------------------------------------
my students come classroom eager energetic ready learn everyday i spend 180 days precious children they hold speci
al part heart never leaves after 8 years teaching i heart full many students many students come low income home go
al provide meaningful comfortable learning experience these students faced several challenges classroom through do
norschoose i hope able provide experiences our classroom flexible seating classroom in flexible seating classroom 
students many different seating options rather typical desk chair we currently rocker chairs mats pillows lap desk
s coffee tables bilibo seats yoga mats using with implementation flexible seating students requested chairs wiggle
after researching seating options felt wobble stools would best support flexible seating plan place through use it
ems seen great success stamina completing classwork by donating project students would able wiggle wobble staying 
task these stools would make difference allowing active even working this especially helpful struggle staying stil
l working we hope consider helping fund project nannan
--------------------------------------------------------------------------------
```



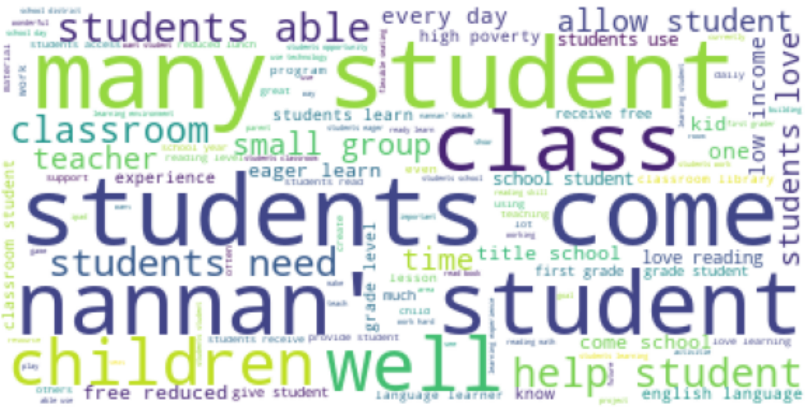
```
================================================================================
In [129]: examine_cluster_with_wordcloud(cluster_dataset[3])
```

```
my 1st grade students attend small school strong community our elementary school provides learning experience meet
s needs students instruction differentiated unique learner they like move love read love lots positive attention f
lexible seating choice provided students allows work around room comfortably focused it provides students environm
ent need best every year students best get wiggles students best standing kneeling number positions little bodies
find comfy the choices students feel invested responsible learning when i give students choices i see greater enga
gement excitement higher desire learn their effort increases certain amount pride comes work some immediate benefi
ts flexible seating include burning calories using excess energy increased motivation engagement improving core st
rength overall posture the materials students need already home need help getting door the large carpet placed fro
nt classroom used gathering place mini lessons share time the bean bags carry around cushions used around classroo
m flexible seating choice reading writing the wobbly stools exercise balls used students may need get wiggles work
ing by donating project not help improve increase student attention focus ultimately help increase academic achiev
ement nannan
--------------------------------------------------------------------------------
i work amazing students springfield holyoke chicopee ages 14 many students struggled traditional public school set
ting luckily college prep school takes unique needs account pca works tfa americorps local community colleges help
students grow fully prepared college my study skills students diving subjects across spectrum including everything
geometry current affairs my students wide variety learning styles lot energy many students come community struggle
s limited resources they share dreams creating new businesses growing communities graduating college they passiona
te making world better place all students benefit regular opportunities move whether need move around room tap des
ks movement helps think with expo markers students freedom grow ideas large dry erase friendly desks without anxie
ty permanence paper this project help students grown minds executive functioning skills adding organization fidget
s expo markers help students learn organize parts school lives prepare organizing materials college nannan
--------------------------------------------------------------------------------
my students coming classroom eager begin educational journey one first experience my classroom one firsts journey
continue reach potential student person productive member society i offer first reading experiences friendships ex
posure technology true socialization peers i also try make meaningful positive possible young my students need leg
o table explore learning center time our students use materials explore ability building creating planning manipul
ating using fine gross motor skills these activities used learning center times throughout day read write learning
together they able use manipulatives whole group small group time your generous donation project improve pre k cla
ssroom building stronger environment learning growing this change students lives better they love school feel succ
ess early age fun learning your generous donation project improve pre k classroom building stronger readers writer
s this change students lives better love school feel success early age fun learning nannan
--------------------------------------------------------------------------------
```



```
================================================================================
In [130]: examine_cluster_with_wordcloud(cluster_dataset[4])
```

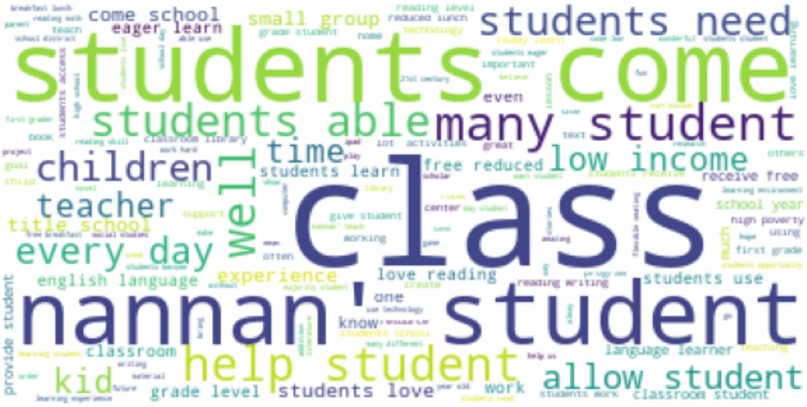
```
in first grade classroom students come variety backgrounds experiences two thirds students come families speak ano
ther language home we first graders fluent six different languages born asia africa latin america united states ma
ny students also come families experiencing poverty they love coming school love my first graders come school exci
ted learn new things grow make friends play they active learners need movement throughout day maximize learning he
althy jumping running place dancing stretching helps grow fun ready learning i engage families first graders home
visits academic parent teacher teams the families students excited helping first graders become readers reading bo
oks home helping develop deeper understanding book talking books read home in class i students new united states l
earning speak english well learning read english their reading skills kindergarten pre kindergarten level they nee
d interesting books beginning reading level take home read families while learning read also need introduced conte
nt first grade science standards they need non fiction books help learn read simple repetitive text beautiful pict
ures match words first graders interested science especially animals my first graders beginning believe readers ne
ed continue reading home well school these books help successful make want keep reading your support project help
students become readers thinkers students families enjoy reading books together talking nannan
--------------------------------------------------------------------------------
my students special they hard workers eager learn every day our school located highway 80 rural alabama unfortunat
ely many students not working internet homes i would love get experience learning basic literacy mathematical skil
ls easy use handheld tablet small group table centers there many apps would benefit students my kindergarten class
benefit tablets classroom many ways i plan using tablets independent center classroom sometimes small group table
they access accelerated reading program students read book take test apps include starfall epic reading reads book
s students many i also plan installing app math called xtra math drills students adding subtraction facts the tabl
ets also provide practice kindergarten class ever changing technological world nannan
--------------------------------------------------------------------------------
my sixth graders go title i school come diverse homes i students live parents single parents students live grandpa
rents but vast majority students low socio economic housing cannot afford school supplies it incredibly sad childr
en deprived opportunities my students 85 range free reduced lunch try provide many materials children not means ge
t my students challenged school not get lot home support job make sure need make learning priority what makes year
difficult others oklahoma education budget lowest ever materials students not available year do not underestimate
power great vocabulary the items project give students differentiated ways learn new words every week keep organiz
ed practice the dry erase tape go directly desktops used daily quick review word definition the cards used create
dictionary words color coded markers finally cases keep dictionary organized portable students practice anywhere a
nytime anybody these items empower kids increase vocabulary end increase reading skills nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

In [131]:
```
examine_cluster_with_wordcloud(cluster_dataset[5])
```

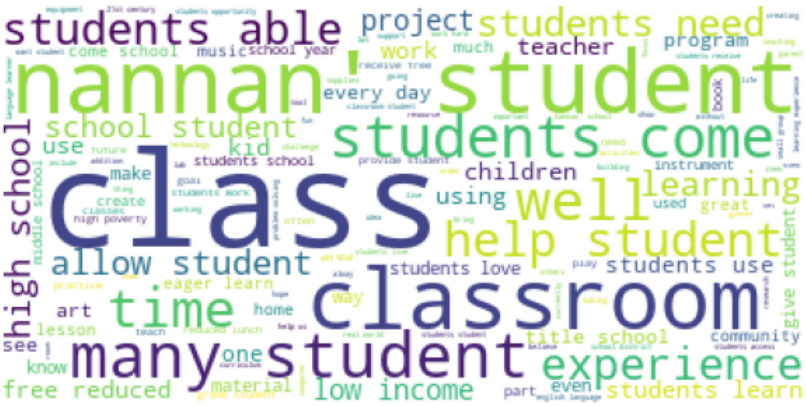
```
my students smartest greatest 3rd graders ever they set goals strive reach every day my kiddos challenge make laug
h teach valuable lessons they balance engaging learning environment technology could always use i want students us
e many forms media technology order enhance learning success please support wonderful students watch reach stars y
ear my students need two laptops classroom students use many forms media technology throughout day along given cur
riculum some computer programs used blended learning st math study island accelerated reader these programs allow
differentiation students learn pace however grade levels share computer cart students wait turn use laptops more l
aptops would enhance learning environment giving students opportunities complete blended learning programs researc
h projects educational activities classroom nannan
--------------------------------------------------------------------------------
my students eighth graders urban california public school they represent cultural socio economic diversity city my
middle school students arrive classroom wide range skills highly gifted second language learners some english stud
ents still read many lost never found love reading these wide ranging abilities interest levels makes challenge fi
nd engaging reading material draw learners literature help kids understand lives others reading real life worldly
situations helps students think circumstances new ways these great titles characters world worlds beyond help buil
d classroom library students book hand without school library students need access wide variety engaging titles an
d reading fiction shown help young people build empathy when students read critically relate characters situations
better able understand people work real world they interact people compassion empathy nannan
--------------------------------------------------------------------------------
as teacher title one school students faced many different challenges the school provides students breakfast lunch
based socioeconomic status area regardless students crave learn literature rich environment i two classes 22 enthu
siastic learners year total 44 amazing students our focus engaging students reading literature our students excite
d every year participate various book clubs competitions the range readers classroom range 3rd grade 5th grade i w
ould like close achievement gap using interesting text students the best way student learn find interesting book r
ead that i want provide students my students come different countries socioeconomic backgrounds these students dif
ferent reading levels interests having books allow students read level read something enjoy it allow students feel
successful level pushing next one all students need exposed literature see wondrous worlds books create the studen
ts using literature books various ways besides reading students create hands activities based books read make crea
tive book reports share others they also make book corner spotlight books take ownership read these books allow te
acher reach individual student students become independent readers they finally get understand concept reading enj
oyment nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

In [132]:
```
examine_cluster_with_wordcloud(cluster_dataset[6])
```

```
my yearbook class place students learning express creativity share creativity fellow classmates community the clas
s composed high school students across many socio economic levels they primarily hispanic students several english
language learners they excited expand skills explore new tools develop skills it desire encourage explore photojou
rnalism graphic design career option i no budget purchase material class i took program year trying revitalize yea
rbook class fallen poor state i spent year trying fund raise new camera equipment order give students best experie
nce possible when i took program beginning year one camera several years old poor condition through outside suppor
t i able obtain two new cameras the final piece i would like add good lens students take better pictures extracurr
icular activities sporting events in order give students better insight photojournalism i would like access equipm
ent would likely use job through project i hope give students opportunity my district community not nearly enough
resources provide students experiences wealthy school district able provide this small step improving quality educ
ation training nannan
--------------------------------------------------------------------------------
my students kind loving energetic zesty first grade special age group students show much growth one year time span
every day come school positive attitude despite challenges may face my students excited new discovery room initiat
ive allows complete stem projects explore wonderful world science social studies they eager learn love working tog
ether group projects especially science curiosity makes day fun challenging full learning my students creating bre
akfast restaurant next unit goods services standard since creative already named restaurant delight coinsides scho
ol motto light this project allow students menu designs come life allow take pride creativity we printing order fo
rms menus students design they also able take ownership taking orders adding costs materials providing change gues
ts our students already excited project i cannot wait possibility using amazing piece technology make vision happe
n ynannan
--------------------------------------------------------------------------------
our students come variety backgrounds we 100 free breakfast lunch make sure kids getting brains bodies fed we almo
st 1500 students 3 buildings middle school campus even though large spread campus sports clubs help bring us toget
her our students get excited extra curricular activities our students always carrying around passion love school c
lothes posters hearts our volleyball players wear spandex shorts games practices having pants games practices help
girls stay warm cold volleyball season being low economic school district not girls afford buy pants providing gir
ls warm pants also help look like team creating positive school pride representation team we teaching volleyball p
layers positively represent school not good sportsmanship court also young ladies court nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

2.6 Apply AgglomerativeClustering</h2>

In [3]:
```
#Considering only 5k pts due to computational limits
```

In [0]:
```
X_tr_5k = X_tr[:5000]
X_train_5k = X_train[:5000]
```

In [66]:
```
X_tr.shape
X_tr_5k.shape
```

Out[66]:
```
(5000, 5000)
```

k=2 (clusters)

In [68]:
```
from sklearn.cluster import AgglomerativeClustering
X_tr_TFIDF_aggl = X_tr_5k.toarray()
aggl_cluster_2 = AgglomerativeClustering(n_clusters=2)
aggl_cluster_2.fit(X_tr_TFIDF_aggl)
```

Out[68]:
```
AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                        connectivity=None, distance_threshold=None,
                        linkage='ward', memory=None, n_clusters=2)
```

In [133]:
```
aggl_cluster_2.n_clusters
```

Out[133]:
```
2
```

In [70]:
```
aggl_cluster_2.labels_
```

Out[70]:
```
array([1, 1, 0, ..., 0, 1, 1])
```

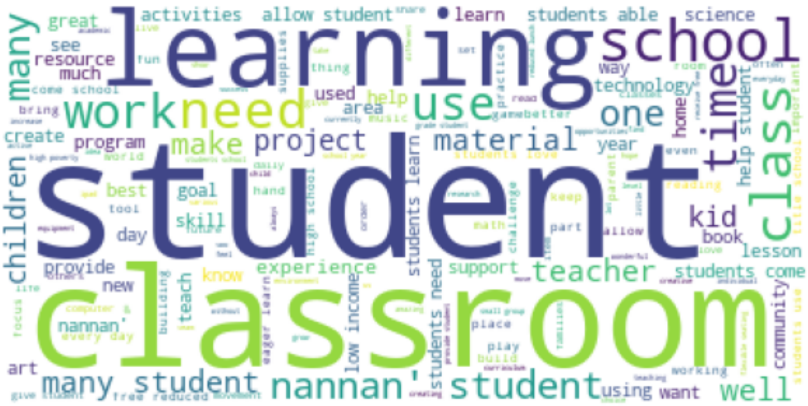
In [134]:
```
cluster_dataset = {i:[] for i in range(aggl_cluster_2.n_clusters)}

for index in tqdm(range(aggl_cluster_2.labels_.shape[0])):
    cluster_dataset[aggl_cluster_2.labels_[index]].append(X_train["preprocessed_essays"].iloc[index])

print("length of each cluster:")
for i in cluster_dataset:
    print(len(cluster_dataset[i]))
```

```
100%|██████████| 5000/5000 [00:00<00:00, 75602.20it/s]

length of each cluster:
3080
1920
```

Examining each cluster and its wordcloud

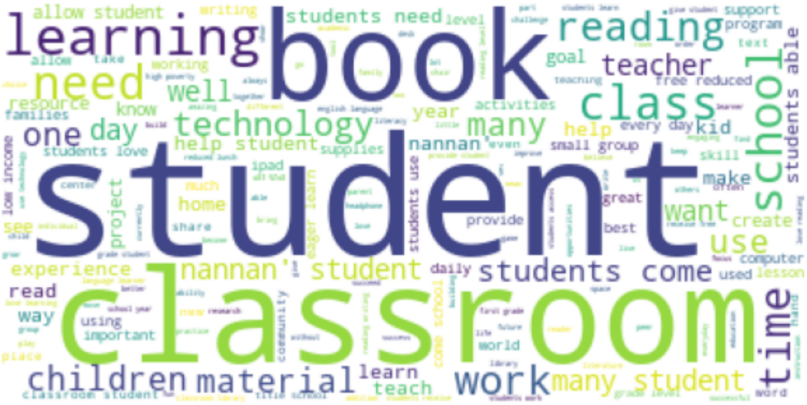
In [135]:
```
examine_cluster_with_wordcloud(cluster_dataset[0])
```

```
my 1st grade students attend small school strong community our elementary school provides learning experience meet
s needs students instruction differentiated unique learner they like move love read love lots positive attention f
lexible seating choice provided students allows work around room comfortably focused it provides students environm
ent need best every year students best get wiggles students best standing kneeling number positions little bodies
find comfy the choices students feel invested responsible learning when i give students choices i see greater enga
gement excitement higher desire learn their effort increases certain amount pride comes work some immediate benefi
ts flexible seating include burning calories using excess energy increased motivation engagement improving core st
rength overall posture the materials students need already home need help getting door the large carpet placed fro
nt classroom used gathering place mini lessons share time the bean bags carry around cushions used around classroo
m flexible seating choice reading writing the wobbly stools exercise balls used students may need get wiggles work
ing by donating project not help improve increase student attention focus ultimately help increase academic achiev
ement nannan
--------------------------------------------------------------------------------
i work amazing students springfield holyoke chicopee ages 14 many students struggled traditional public school set
ting luckily college prep school takes unique needs account pca works tfa americorps local community colleges help
students grow fully prepared college my study skills students diving subjects across spectrum including everything
geometry current affairs my students wide variety learning styles lot energy many students come community struggle
s limited resources they share dreams creating new businesses growing communities graduating college they passiona
te making world better place all students benefit regular opportunities move whether need move around room tap des
ks movement helps think with expo markers students freedom grow ideas large dry erase friendly desks without anxie
ty permanence paper this project help students grown minds executive functioning skills adding organization fidget
s expo markers help students learn organize parts school lives prepare organizing materials college nannan
--------------------------------------------------------------------------------
as teacher low income high poverty school district students faced many challenges classroom despite many challenge
s face i looking help engage learning exploring creativity us history they learn best active participation movemen
t around classroom hands activities many challenges face may prevent getting ahead early life necessary supplies f
rom minute walk classroom i focus potential growth i may not able control home lives however i control experience
school day help ignite love learning especially american history by creative positive way i hopeful inspire even e
arliest learners continue path academic excellence my students materials needed participate active engaged learnin
g activities the students use paper creating manipulatives help learn us history the paper also used differentiate
instruction according student needs the markers glue scissors enable develop projects enable explore creativity br
ing history life the pens pencils provide students may not necessary materials class ability fully participate as
malala yousafzai says one child one teacher one book one pen change world but first need book pens nannan
--------------------------------------------------------------------------------
```



```
================================================================================
In [136]: examine_cluster_with_wordcloud(cluster_dataset[1])
```

```
have ever made plans try new recipe create something you envisioned serving eating dish using new creation you may
even started creating dish project found pause project not ingredients materials needed do remember disappointment
frustration felt could not follow plan that sometimes happens students resources need succeed it amazing watch mid
dle school students grow see evidence growth collaborative conversations writing student products data they learni
ng set academic goals create plans utilize strategies meet goals unfortunately sometimes start moving full speed a
head toward achieving goals find not supplies resources need sadly lose momentum get back track success seventh ei
ghth grade students use interactive notebooks learn constitution prepare federal constitution test the notebooks u
sed take notes create study guides house foldable student products organizers student creations work social studie
s stations not interactive notebooks serve students well study constitution also serve portfolios the notebooks sh
owcase provide evidence students learning similarly sixth grade students use notebooks learning stations social st
udies learn writing traits process nannan
--------------------------------------------------------------------------------
my students kids when walk door reading language arts classroom i take ownership education well part lives they co
me variety levels finished harry potter transitioning picture books chapter books even reluctant readers avoid boo
ks costs my students best best or least i tell first day school i tell i secret share not tell anyone else school
sitting stool front room seeing attentive eyes eagerly waiting secret i tell hushed voice principal put best third
graders classroom i get privilege teaching best classes year it never fails live expectation any one one small gro
up instruction plus students but teacher 16 kids leading small group the answer engaging activity stations engaged
students busy focused learning become distraction small group students individual learners the reading writing act
ivity stations provide clear instructions easy student needed materials easy teacher after concept station introdu
ced class station added small group rotation time thus keeping 16 kids happily learning without teacher i instruct
small group 6 nannan
--------------------------------------------------------------------------------
in first grade classroom students come variety backgrounds experiences two thirds students come families speak ano
ther language home we first graders fluent six different languages born asia africa latin america united states ma
ny students also come families experiencing poverty they love coming school love my first graders come school exci
ted learn new things grow make friends play they active learners need movement throughout day maximize learning he
althy jumping running place dancing stretching helps grow fun ready learning i engage families first graders home
visits academic parent teacher teams the families students excited helping first graders become readers reading bo
oks home helping develop deeper understanding book talking books read home in class i students new united states l
earning speak english well learning read english their reading skills kindergarten pre kindergarten level they nee
d interesting books beginning reading level take home read families while learning read also need introduced conte
nt first grade science standards they need non fiction books help learn read simple repetitive text beautiful pict
ures match words first graders interested science especially animals my first graders beginning believe readers ne
ed continue reading home well school these books help successful make want keep reading your support project help
students become readers thinkers students families enjoy reading books together talking nannan
--------------------------------------------------------------------------------
```



```
=================================================================================
```

k=5 (clusters)

```
In [74]: from sklearn.cluster import AgglomerativeClustering
         X_tr_TFIDF_aggl = X_tr_5k.toarray()
         aggl_cluster_5 = AgglomerativeClustering(n_clusters=5)
         aggl_cluster_5.fit(X_tr_TFIDF_aggl)
```

```
Out[74]: AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                                 connectivity=None, distance_threshold=None,
                                 linkage='ward', memory=None, n_clusters=5)
```

```
In [75]: aggl_cluster_5.n_clusters
```

```
Out[75]: 5
```

```
In [76]: aggl_cluster_5.labels_
```

```
Out[76]: array([4, 4, 2, ..., 1, 4, 3])
```

In [137]:

```
cluster_dataset = {i:[] for i in range(aggl_cluster_5.n_clusters)}

for index in tqdm(range(aggl_cluster_5.labels_.shape[0])):
    cluster_dataset[aggl_cluster_5.labels_[index]].append(X_train["preprocessed_essays"].iloc[index])

print("length of each cluster:")
for i in cluster_dataset:
    print(len(cluster_dataset[i]))
```

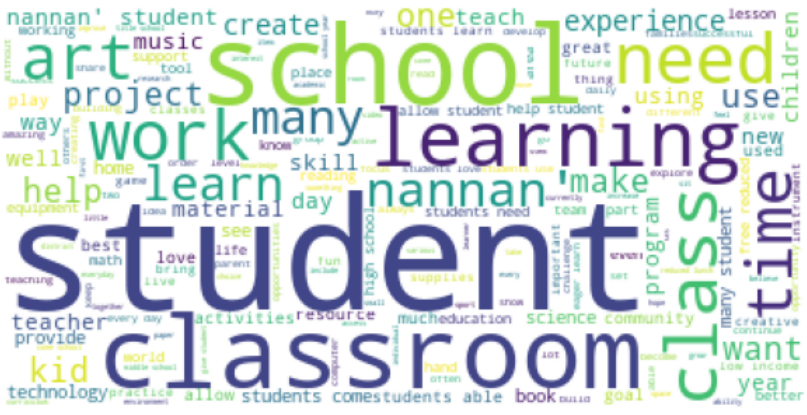
```
100%|██████████| 5000/5000 [00:00<00:00, 71686.97it/s]

length of each cluster:
1413
1086
581
790
1130
```

Examining each cluster and its wordcloud

In [138]:

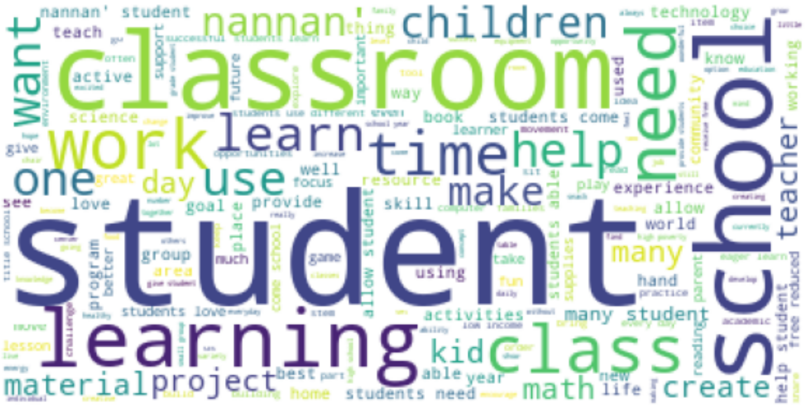
```
examine_cluster_with_wordcloud(cluster_dataset[0])
```

```
my students eclectic mix movers shakers they full energy energy not always reserved playground this active group 3
4 students insists learning easier entire body involved they like freedom wiggle work my students come variety soc
ioeconomic backgrounds ethnicities many school one constant lives no matter home life like one deserves opportunit
y learn environment welcoming encouraging as students become focused comfortable become productive reach highest p
otential potential sometimes surprises even my students told hard sit still work some described human bouncy balls
wanted know important sit still i told 34 kids classroom not much sitting still part important seat seat respect s
pace i thought discussion days later class meeting group students asked could get wiggle seats like seen classroom
s i replied great idea asked one going pay chairs because great items classroom generously donated donorschoose su
ggested submit project asking chairs share brilliant idea i said we spent time looking amazing chairs allow people
move work decided buoy chairs sturdy enough nine year old move like wind it proven active seating improves posture
strengthens core abdominal trunk muscles students constantly motion this also beneficial increasing strength muscl
e tone promotes motor control use pencil scissors classroom tools it wonderful give student opportunity get bodies
healthy minds work nannan
--------------------------------------------------------------------------------
my students diverse group i many kids would benefit movement class special education students adhd students medica
l needs students often students cannot stay focused movement keep going i often students not speak english movemen
t way connect some students cannot afford join sports project help day day reading writing often kids sitting long
periods time my class came idea get equipment room would help move healthy reading writing musical mats anyone we
saw mats thought fun would musical mat activity get moving thinking we could play music music stops whatever mat t
ells it might jog place hop my kids would never sitting long my kids bike pedals would love try elliptical ones th
is awesome way keep legs moving imagine kids room instead sitting chairs those students need hard space work added
trays work i cannot wait see working mats though please consider helping us add items room nannan
--------------------------------------------------------------------------------
my students kids i like call come everyday chip shoulder daring teach learning today important life so everyday i
motivate inspire kids all i want kids say done best my students majority native hawaiians pacific islanders the co
mmunities students come considered low socio economic status i feel students mentally tough come school ready work
they little rough around edges show teacher they anything many come rough tough home i teaching school since 2005
2006 school year our school operation since 2002 2003 school year still growing we consistently trying improve bet
ter school can i drink water normal question students i receive daily my students walk around hydro flasks filled
water carry around powerade drinks when students classroom ask drink water options limited they either drink water
sink classroom sink bathroom water fountain couple minutes away class all three options not viable i student i wou
ld not drink sink bathroom for project i requesting bottom load water cooler this water cooler located classroom s
tudents allowed fill containers water everyday i hoping water cooler encourage students drink water daily help cut
back sugary drinks nannan
--------------------------------------------------------------------------------
```



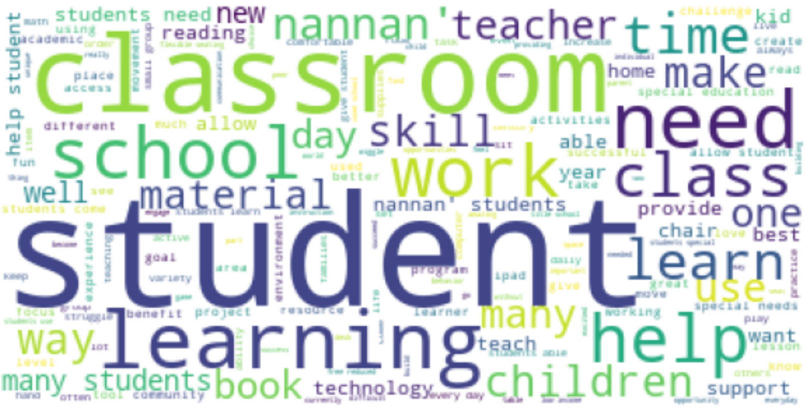
```
================================================================================
In [139]: examine_cluster_with_wordcloud(cluster_dataset[1])
```

```
as teacher low income high poverty school district students faced many challenges classroom despite many challenge
s face i looking help engage learning exploring creativity us history they learn best active participation movemen
t around classroom hands activities many challenges face may prevent getting ahead early life necessary supplies f
rom minute walk classroom i focus potential growth i may not able control home lives however i control experience
school day help ignite love learning especially american history by creative positive way i hopeful inspire even e
arliest learners continue path academic excellence my students materials needed participate active engaged learnin
g activities the students use paper creating manipulatives help learn us history the paper also used differentiate
instruction according student needs the markers glue scissors enable develop projects enable explore creativity br
ing history life the pens pencils provide students may not necessary materials class ability fully participate as
malala yousafzai says one child one teacher one book one pen change world but first need book pens nannan
--------------------------------------------------------------------------------
21st century students need access 21st century technology we not resources need provide challenged risk students t
echnology individualize learning increase student engagement the majority students black latino english language l
earners mild severe learning disabilities 90 receive free reduced price lunch they live shelters parents incarcera
ted siblings gangs adversity aspects lives yet despite obstacles students attendance daily ready motivated learn s
ucceed desire provide opportunities change lives we use imac classroom write record short films math common core m
ath concepts in videos students act real world math problems they also create videos teach classmates major math c
oncepts emphasizing ways talk math problems strategies used find solutions students also use imac access digital g
oogle classroom students access complete assignments online watch video lessons extra support sign small group hel
p needed students able use wireless mouse presenting class time student engagement improve use technology sparks i
nterest emphasizes student voice student thinking student centered classroom students learn researching skills app
ropriate internet use proper technology use keyboarding development using technology educational purposes well int
ernet safety rules nannan
--------------------------------------------------------------------------------
my students come classroom eager energetic ready learn everyday i spend 180 days precious children they hold speci
al part heart never leaves after 8 years teaching i heart full many students many students come low income home go
al provide meaningful comfortable learning experience these students faced several challenges classroom through do
norschoose i hope able provide experiences our classroom flexible seating classroom in flexible seating classroom
students many different seating options rather typical desk chair we currently rocker chairs mats pillows lap desk
s coffee tables bilibo seats yoga mats using with implementation flexible seating students requested chairs wiggle
after researching seating options felt wobble stools would best support flexible seating plan place through use it
ems seen great success stamina completing classwork by donating project students would able wiggle wobble staying
task these stools would make difference allowing active even working this especially helpful struggle staying stil
l working we hope consider helping fund project nannan
--------------------------------------------------------------------------------
```



```
================================================================================
In [140]: examine_cluster_with_wordcloud(cluster_dataset[2])
```

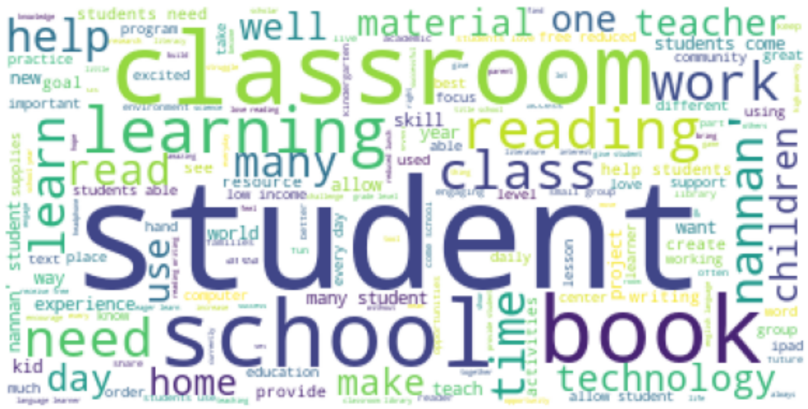
```
my 1st grade students attend small school strong community our elementary school provides learning experience meet
s needs students instruction differentiated unique learner they like move love read love lots positive attention f
lexible seating choice provided students allows work around room comfortably focused it provides students environm
ent need best every year students best get wiggles students best standing kneeling number positions little bodies
find comfy the choices students feel invested responsible learning when i give students choices i see greater enga
gement excitement higher desire learn their effort increases certain amount pride comes work some immediate benefi
ts flexible seating include burning calories using excess energy increased motivation engagement improving core st
rength overall posture the materials students need already home need help getting door the large carpet placed fro
nt classroom used gathering place mini lessons share time the bean bags carry around cushions used around classroo
m flexible seating choice reading writing the wobbly stools exercise balls used students may need get wiggles work
ing by donating project not help improve increase student attention focus ultimately help increase academic achiev
ement nannan
--------------------------------------------------------------------------------
i work amazing students springfield holyoke chicopee ages 14 many students struggled traditional public school set
ting luckily college prep school takes unique needs account pca works tfa americorps local community colleges help
students grow fully prepared college my study skills students diving subjects across spectrum including everything
geometry current affairs my students wide variety learning styles lot energy many students come community struggle
s limited resources they share dreams creating new businesses growing communities graduating college they passiona
te making world better place all students benefit regular opportunities move whether need move around room tap des
ks movement helps think with expo markers students freedom grow ideas large dry erase friendly desks without anxie
ty permanence paper this project help students grown minds executive functioning skills adding organization fidget
s expo markers help students learn organize parts school lives prepare organizing materials college nannan
--------------------------------------------------------------------------------
my students coming classroom eager begin educational journey one first experience my classroom one firsts journey
continue reach potential student person productive member society i offer first reading experiences friendships ex
posure technology true socialization peers i also try make meaningful positive possible young my students need leg
o table explore learning center time our students use materials explore ability building creating planning manipul
ating using fine gross motor skills these activities used learning center times throughout day read write learning
together they able use manipulatives whole group small group time your generous donation project improve pre k cla
ssroom building stronger environment learning growing this change students lives better they love school feel succ
ess early age fun learning your generous donation project improve pre k classroom building stronger readers writer
s this change students lives better love school feel success early age fun learning nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

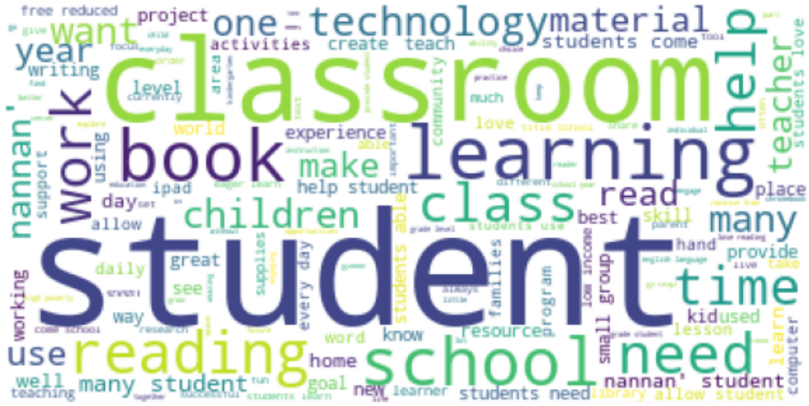
In [141]:
```
examine_cluster_with_wordcloud(cluster_dataset[3])
```

```
in first grade classroom students come variety backgrounds experiences two thirds students come families speak ano
ther language home we first graders fluent six different languages born asia africa latin america united states ma
ny students also come families experiencing poverty they love coming school love my first graders come school exci
ted learn new things grow make friends play they active learners need movement throughout day maximize learning he
althy jumping running place dancing stretching helps grow fun ready learning i engage families first graders home
visits academic parent teacher teams the families students excited helping first graders become readers reading bo
oks home helping develop deeper understanding book talking books read home in class i students new united states l
earning speak english well learning read english their reading skills kindergarten pre kindergarten level they nee
d interesting books beginning reading level take home read families while learning read also need introduced conte
nt first grade science standards they need non fiction books help learn read simple repetitive text beautiful pict
ures match words first graders interested science especially animals my first graders beginning believe readers ne
ed continue reading home well school these books help successful make want keep reading your support project help
students become readers thinkers students families enjoy reading books together talking nannan
--------------------------------------------------------------------------------
my students smartest greatest 3rd graders ever they set goals strive reach every day my kiddos challenge make laug
h teach valuable lessons they balance engaging learning environment technology could always use i want students us
e many forms media technology order enhance learning success please support wonderful students watch reach stars y
ear my students need two laptops classroom students use many forms media technology throughout day along given cur
riculum some computer programs used blended learning st math study island accelerated reader these programs allow
differentiation students learn pace however grade levels share computer cart students wait turn use laptops more l
aptops would enhance learning environment giving students opportunities complete blended learning programs researc
h projects educational activities classroom nannan
--------------------------------------------------------------------------------
my students eighth graders urban california public school they represent cultural socio economic diversity city my
middle school students arrive classroom wide range skills highly gifted second language learners some english stud
ents still read many lost never found love reading these wide ranging abilities interest levels makes challenge fi
nd engaging reading material draw learners literature help kids understand lives others reading real life worldly
situations helps students think circumstances new ways these great titles characters world worlds beyond help buil
d classroom library students book hand without school library students need access wide variety engaging titles an
d reading fiction shown help young people build empathy when students read critically relate characters situations
better able understand people work real world they interact people compassion empathy nannan
--------------------------------------------------------------------------------
```



```
================================================================================
In [142]: examine_cluster_with_wordcloud(cluster_dataset[4])
```

```
have ever made plans try new recipe create something you envisioned serving eating dish using new creation you may
even started creating dish project found pause project not ingredients materials needed do remember disappointment
frustration felt could not follow plan that sometimes happens students resources need succeed it amazing watch mid
dle school students grow see evidence growth collaborative conversations writing student products data they learni
ng set academic goals create plans utilize strategies meet goals unfortunately sometimes start moving full speed a
head toward achieving goals find not supplies resources need sadly lose momentum get back track success seventh ei
ghth grade students use interactive notebooks learn constitution prepare federal constitution test the notebooks u
sed take notes create study guides house foldable student products organizers student creations work social studie
s stations not interactive notebooks serve students well study constitution also serve portfolios the notebooks sh
owcase provide evidence students learning similarly sixth grade students use notebooks learning stations social st
udies learn writing traits process nannan
--------------------------------------------------------------------------------
my students kids when walk door reading language arts classroom i take ownership education well part lives they co
me variety levels finished harry potter transitioning picture books chapter books even reluctant readers avoid boo
ks costs my students best best or least i tell first day school i tell i secret share not tell anyone else school
sitting stool front room seeing attentive eyes eagerly waiting secret i tell hushed voice principal put best third
graders classroom i get privilege teaching best classes year it never fails live expectation any one one small gro
up instruction plus students but teacher 16 kids leading small group the answer engaging activity stations engaged
students busy focused learning become distraction small group students individual learners the reading writing act
ivity stations provide clear instructions easy student needed materials easy teacher after concept station introdu
ced class station added small group rotation time thus keeping 16 kids happily learning without teacher i instruct
small group 6 nannan
--------------------------------------------------------------------------------
my students special they hard workers eager learn every day our school located highway 80 rural alabama unfortunat
ely many students not working internet homes i would love get experience learning basic literacy mathematical skil
ls easy use handheld tablet small group table centers there many apps would benefit students my kindergarten class
benefit tablets classroom many ways i plan using tablets independent center classroom sometimes small group table
they access accelerated reading program students read book take test apps include starfall epic reading reads book
s students many i also plan installing app math called xtra math drills students adding subtraction facts the tabl
ets also provide practice kindergarten class ever changing technological world nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

2.7 Apply DBSCAN

Considering only 5k pts due to computational limts. Process followed to find best eps: 1) find the distance from every point to its nearest neighbour point [i.e, pairwise distance] for different 'k'. 2) sort the distances 3) plot the distances. Now, the distance where higher number of the pts are lying is our eps. Basically can be identified from the plot, where behavior of plots changes quickly.

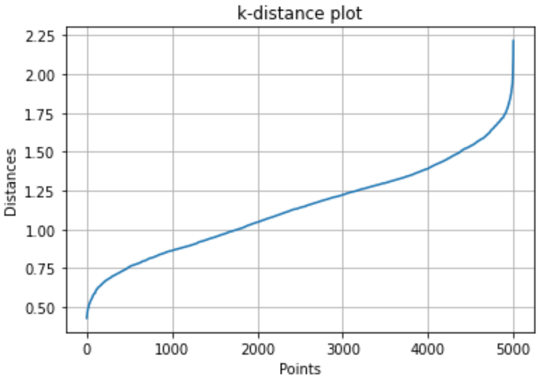
K-distance graph

```
In [98]: %%time
         from sklearn.neighbors import NearestNeighbors

         NN = NearestNeighbors(n_neighbors=4).fit(X_tr_5k)
         distances, indices = NN.kneighbors(X_tr_5k)
```

```
CPU times: user 1.59 s, sys: 62 ms, total: 1.65 s
Wall time: 1.66 s
```

```
In [0]: sorted_dist = sorted(distances[:,-1])
```

In [100]:
```
plt.plot(sorted_dist)
plt.grid(True)
plt.xlabel('Points')
plt.ylabel('Distances')
plt.title('k-distance plot')
plt.show()
```



DBSCAN Algorithm with best eps

In [102]:
```
%%time
from sklearn.cluster import DBSCAN

# choose eps=1.60
dbscan = DBSCAN(eps=1.60)
dbscan.fit(X_tr_5k)
```

```
CPU times: user 1.39 s, sys: 61 ms, total: 1.45 s
Wall time: 1.45 s
```

In [144]:
```
dbscan.labels_
```

Out[144]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

In [146]:
```
dbscan_clusters=[[],[]]
for index in tqdm(range(dbscan.labels_.shape[0])):
    dbscan_clusters[dbscan.labels_[index]].append(X_train["preprocessed_essays"].iloc[index])
```

```
100%|██████████| 5000/5000 [00:00<00:00, 65791.35it/s]
```

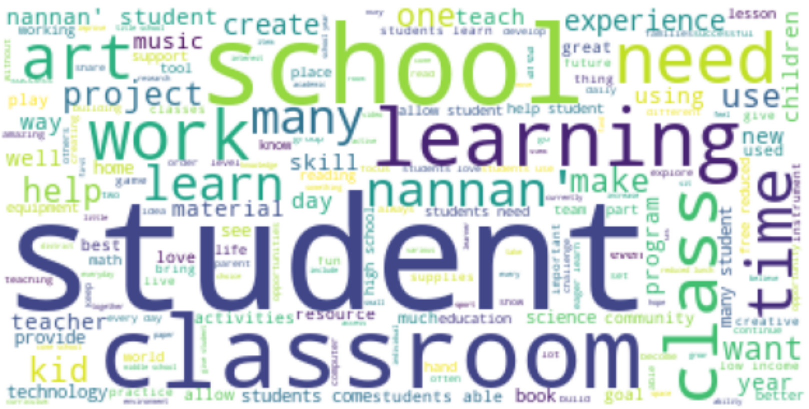
In [147]:
```
for i in dbscan_clusters:
  print(len(i))
```

```
4765
235
```

Examining each cluster and its wordcloud

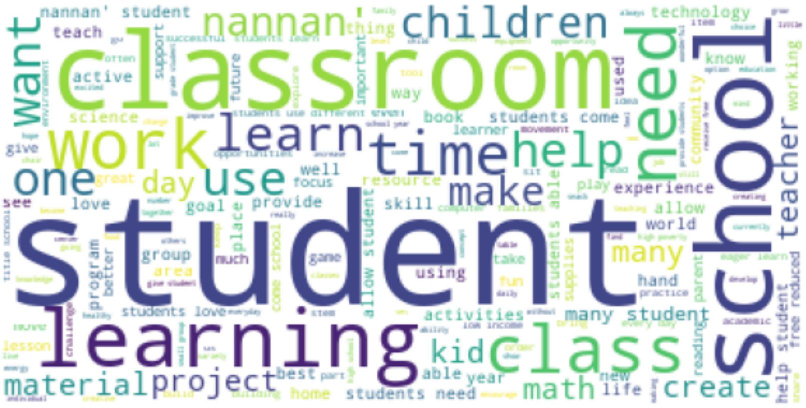
```
In [148]: examine_cluster_with_wordcloud(cluster_dataset[0])
```

```
my students eclectic mix movers shakers they full energy energy not always reserved playground this active group 3
4 students insists learning easier entire body involved they like freedom wiggle work my students come variety soc
ioeconomic backgrounds ethnicities many school one constant lives no matter home life like one deserves opportunit
y learn environment welcoming encouraging as students become focused comfortable become productive reach highest p
otential potential sometimes surprises even my students told hard sit still work some described human bouncy balls
wanted know important sit still i told 34 kids classroom not much sitting still part important seat seat respect s
pace i thought discussion days later class meeting group students asked could get wiggle seats like seen classroom
s i replied great idea asked one going pay chairs because great items classroom generously donated donorschoose su
ggested submit project asking chairs share brilliant idea i said we spent time looking amazing chairs allow people
move work decided buoy chairs sturdy enough nine year old move like wind it proven active seating improves posture
strengthens core abdominal trunk muscles students constantly motion this also beneficial increasing strength muscl
e tone promotes motor control use pencil scissors classroom tools it wonderful give student opportunity get bodies
healthy minds work nannan
--------------------------------------------------------------------------------
my students diverse group i many kids would benefit movement class special education students adhd students medica
l needs students often students cannot stay focused movement keep going i often students not speak english movemen
t way connect some students cannot afford join sports project help day day reading writing often kids sitting long
periods time my class came idea get equipment room would help move healthy reading writing musical mats anyone we
saw mats thought fun would musical mat activity get moving thinking we could play music music stops whatever mat t
ells it might jog place hop my kids would never sitting long my kids bike pedals would love try elliptical ones th
is awesome way keep legs moving imagine kids room instead sitting chairs those students need hard space work added
trays work i cannot wait see working mats though please consider helping us add items room nannan
--------------------------------------------------------------------------------
my students kids i like call come everyday chip shoulder daring teach learning today important life so everyday i
motivate inspire kids all i want kids say done best my students majority native hawaiians pacific islanders the co
mmunities students come considered low socio economic status i feel students mentally tough come school ready work
they little rough around edges show teacher they anything many come rough tough home i teaching school since 2005
2006 school year our school operation since 2002 2003 school year still growing we consistently trying improve bet
ter school can i drink water normal question students i receive daily my students walk around hydro flasks filled
water carry around powerade drinks when students classroom ask drink water options limited they either drink water
sink classroom sink bathroom water fountain couple minutes away class all three options not viable i student i wou
ld not drink sink bathroom for project i requesting bottom load water cooler this water cooler located classroom s
tudents allowed fill containers water everyday i hoping water cooler encourage students drink water daily help cut
back sugary drinks nannan
--------------------------------------------------------------------------------
```



```
================================================================================
In [149]: examine_cluster_with_wordcloud(cluster_dataset[1])
```

```
as teacher low income high poverty school district students faced many challenges classroom despite many challenge
s face i looking help engage learning exploring creativity us history they learn best active participation movemen
t around classroom hands activities many challenges face may prevent getting ahead early life necessary supplies f
rom minute walk classroom i focus potential growth i may not able control home lives however i control experience
school day help ignite love learning especially american history by creative positive way i hopeful inspire even e
arliest learners continue path academic excellence my students materials needed participate active engaged learnin
g activities the students use paper creating manipulatives help learn us history the paper also used differentiate
instruction according student needs the markers glue scissors enable develop projects enable explore creativity br
ing history life the pens pencils provide students may not necessary materials class ability fully participate as
malala yousafzai says one child one teacher one book one pen change world but first need book pens nannan
--------------------------------------------------------------------------------
21st century students need access 21st century technology we not resources need provide challenged risk students t
echnology individualize learning increase student engagement the majority students black latino english language l
earners mild severe learning disabilities 90 receive free reduced price lunch they live shelters parents incarcera
ted siblings gangs adversity aspects lives yet despite obstacles students attendance daily ready motivated learn s
ucceed desire provide opportunities change lives we use imac classroom write record short films math common core m
ath concepts in videos students act real world math problems they also create videos teach classmates major math c
oncepts emphasizing ways talk math problems strategies used find solutions students also use imac access digital g
oogle classroom students access complete assignments online watch video lessons extra support sign small group hel
p needed students able use wireless mouse presenting class time student engagement improve use technology sparks i
nterest emphasizes student voice student thinking student centered classroom students learn researching skills app
ropriate internet use proper technology use keyboarding development using technology educational purposes well int
ernet safety rules nannan
--------------------------------------------------------------------------------
my students come classroom eager energetic ready learn everyday i spend 180 days precious children they hold speci
al part heart never leaves after 8 years teaching i heart full many students many students come low income home go
al provide meaningful comfortable learning experience these students faced several challenges classroom through do
norschoose i hope able provide experiences our classroom flexible seating classroom in flexible seating classroom
students many different seating options rather typical desk chair we currently rocker chairs mats pillows lap desk
s coffee tables bilibo seats yoga mats using with implementation flexible seating students requested chairs wiggle
after researching seating options felt wobble stools would best support flexible seating plan place through use it
ems seen great success stamina completing classwork by donating project students would able wiggle wobble staying
task these stools would make difference allowing active even working this especially helpful struggle staying stil
l working we hope consider helping fund project nannan
--------------------------------------------------------------------------------
```



```
================================================================================
```

3. Conclusions

K-Means

1. Firstly we ran KMeans on k=[2, 3, 4, 5, 6, 7, 10, 15, 20].
2. Then we plotted K vs inertia graph and we observed that optimal k value is 7 using elbow knee method.
3. We plotted a word cloud for that cluster's preprocessed essay data.

Agglomerative Clustering

1. Agglomerative clustering works for dense matrices, so we converted into dense but considered only 5k pts due to computational limts.
2. I ran for n_clusters=[2,5]
3. We plotted a word cloud for that cluster's preprocessed essay data.

DBSCAN

1. There are two parameters eps, minpts.
2. So we plotted k-distance graph on distances(sorted) against no. of points.
3. Then we ran DBSCAN with optimal eps
4. We plotted a word cloud for that cluster's preprocessed essay data.

In [6]:

```
from prettytable import PrettyTable
pt = PrettyTable(["ALgorithm","Vectorizer","parameters"])
pt.add_row(['KMeans','TFIDF','n_clusters=7'])
pt.add_row(['Agglomerative','TFIDF','n_clusters=2&5'])
pt.add_row(['DBSCAN','TFIDF','eps=1.60'])
print(pt)
```

```
+---------------+------------+----------------+
|   ALgorithm   | Vectorizer |   parameters   |
+---------------+------------+----------------+
|     KMeans    |    TFIDF   |  n_clusters=7  |
| Agglomerative |    TFIDF   | n_clusters=2&5 |
|     DBSCAN    |    TFIDF   |    eps=1.60    |
+---------------+------------+----------------+
```

In [0]: