

Emotional Analysis and Evaluation of Positive and Negative Aspects Based on Comprehensive Quality Evaluation in Tsinghua University High School

Tsinghua University High School

Zhaoyang Tian

Advisor: Yuyao Zhang

February, 2019

Abstract

Sponsored by Beijing Education Committee(北京市教育委员会), Tsinghua University High School has developed Comprehensive Quality Evaluation Platform for Beijing General High School Students(北京市普通高中学生综合素质评价平台), which aims that students can voluntarily upload the photos and text description and reveal their special activities and strength that cannot be reflected in class or by scores. The CQE system has been implemented for over a year, while no research towards the data in CQE has been made.

This paper demonstrates the data set in a brand new manner, evaluates the positive and negative feelings behind the data judges the most popular and welcomed activities among school students, and evaluates the positive and negative aspect of the popular activities.

First, this paper utilizes JIEBA module to segment the words and count the word frequency. Then this paper employs WORDCLOUD module to show the result in a straightforward way. Next this paper poses an original method to evaluate the positive and negative feelings of the data, bolstered by SnowNLP module to show that this method is valid and effective. This paper uses Supporting Vector Machine(SVM) to find the critical words of the activities that are most popular in school and use the word frequency to back up the key words. This paper compares the result between SVM and Bayes Distinction, finding that the result of SVM is better than that of Bayes Distinction, Finally, this paper finds the positive and negative aspects of the activities and reaches the conclusion.

This paper discovers some unexpected conclusions, such as “Nervous”(紧张) is a word that is often mentioned in pieces of data that are more likely to be treated as negative in sports activities, which includes sports meeting, soccer match, and basketball match. The most influential creative point is that this paper proposes a brand new method based on dictionary to judge whether a paragraph is positive and negative, and the idea is turned into algorithms. With comprehensive use of various kinds of techniques including SVM, Bayes Distinction this paper ultimately reaches out a conclusion towards a question that has long been asked, which is what kind of activities are favored by students, with the aid of the data on the newly-developed Comprehensive Quality Evaluation Platform.

Keywords:

JIEBA, WORDCLOUD, SVM, SnowNLP, Bayes Distinction, Original Emotional Evaluation Method, Comprehensive Quality Evaluation, Emotional Analysis

Contents

1	Introduction	5
1.1	Research Background	5
1.2	Current Research Status	5
1.3	Research Purpose and Significance	6
1.4	Research Method and General Process	6
2	Data Extraction	7
3	JIEBA Word Segmentation	8
3.1	JIEBA introduction	8
3.2	JIEBA Word Frequency Analysis	9
4	WORDCLOUD	14
4.1	WORDCLOUD Introduction	14
4.2	WORDCLOUD Production	14
5	Emotional Evaluation	14
5.1	SnowNLP	19
5.2	Original Emotional Evaluation Method	20
5.2.1	Enumeration of word lists and Preparation of the data	20
5.2.2	Definition of scores of emotional words	22
5.2.3	Definition of scores of degree words	22
5.2.4	Definition of scores of exclamation marks	22
5.2.5	Definition of scores of reverse words	23
5.2.6	Definition of scores of a whole paragraph	23
5.3	Cross-test between SnowNLP and Original Emotion Evaluation Method	23
6	Supporting Vector Machine	24
6.1	Analysis of SVM classification	24
6.2	Analysis of Timing Function	25
6.3	Analysis of SVM classification	25
6.4	Analysis of Bayes Distinction	26
6.5	Final Results	27
6.6	Cross-test between SVM and Word Frequency	28
7	Evaluation of Positive and Negative Aspects	28

8	Conclusion	28
8.1	Strength and Weakness	28
8.2	Conclusion	30
9	Acknowledgement	32
10	Declaration	32
11	Bibliography	33

1 Introduction

1.1 Research Background

With the reformation of College/University Entrance Examination(高等学校招生考试) in China, abbreviated as *Gaokao*(高考), Chinese government and authority is impulsively transforming the enrollment system of colleges and universities from test-oriented to comprehensive-quality-oriented, which requires high schools and universities to become more concentrated on the all-round developments of students rather than their test scores. Chinese President Xi Jinping has proposed that CCP must take fostering integrity and promoting rounded development of people as the fundamental task of education. During the time, sponsored by Beijing Education Committee(北京市教育委员会), Tsinghua University High School has developed Comprehensive Quality Evaluation Platform for Beijing General High School Students(北京市普通高中学生综合素质评价平台)(CQE)[1], which aims that students can voluntarily upload the photos and text description and reveal their special activities and strength that cannot be reflected in class or by scores. It has a large number of dimensions, including Ideology and Morality(思想品德), Scholastic Achievement(学业成就), Physical and Mental Health(身心健康), Artistic Accomplishment(艺术成就), Artistic Practice(艺术实践), and Social Practice(社会实践). From the point of view of schools, they are able to utilize and analyze the data to provide a more welcomed school for students. However, the information behind the data remains unknown.

1.2 Current Research Status

It has been a long time since education has been invented. [2] Nevertheless, the past research mostly focus on methodology and unique example conducted by a few researchers, with the limitation of small amount of data. With the presence of the new information age, it is possible for individual researchers to gather large amounts of data and and analyze the law behind the data. In the study conducted by Bingbing Xu [3], she advocates broadening research perspectives and introduced Key Competencies Assessment into the Student Assessment system in high school, to realize the transformation from “Comprehensive Quality Assessment” to “Key Competencies Assessment” in high school. In the study conducted by Zhaohui Chen [4], He selects a high school to apply the “3R” model to implement the comprehensive quality evaluation as a case study to examine the effectiveness of current evaluation system. The previous two researcher-

s fully and wholly conclude the origin, development, and the prospect of CQE. There are also many other researchers who provide insight to the understanding[5], concept[6], dilemma[7], solution[8], and value[9] of CQE. Nonetheless, none of the article mentions how to utilize the data in a proper way.

One of the significant researches is done by Dianjun Wang, Hui Ju, Weidong Meng[10], who are the initiator of CQE and the headmaster of the school. They elucidate the concept, design, purpose, as well as significance of CQE in detail. They also point out the fact that school can utilize the data in CQE to analyze students' interest, but fail to give out a elaborate method to do so.

1.3 Research Purpose and Significance

Since schools ought to go their great length offer colorful and meaningful activities to meet the need of students, this paper conducts research in the hope of offering a practical solution in a statistical way, This paper demonstrates the data set in a brand new manner, evaluates the positive and negative feelings behind the data, and recommends the activities for the school to refer to. The research purpose can be shown as below:

1. To demonstrate the data set in a brand new manner.
2. To evaluate the positive and negative feelings behind the data.
3. To judge the most popular and welcomed activities among school students.
4. To evaluate the positive and negative aspect of the popular activities.

The results of this paper will be of great reference and help to schools by offering a clear and vivid explanation of what activities are more welcomed as well as the positive and negative aspects. Thus they can hold more of the activities, develop the positive aspects, and avoid the negative aspects. The leader of the school can consequently adjust their activities in schools according to this paper to achieve a higher degree of satisfaction.

1.4 Research Method and General Process

Figure 1 above presents the whole modeling process. This paper extracts the data at the very beginning. In order to solve the problem illustrated above, this paper considers to divide the whole process into several parts. First this paper needs to segments the words, since the program cannot analyze the sentences as a whole. Then this paper

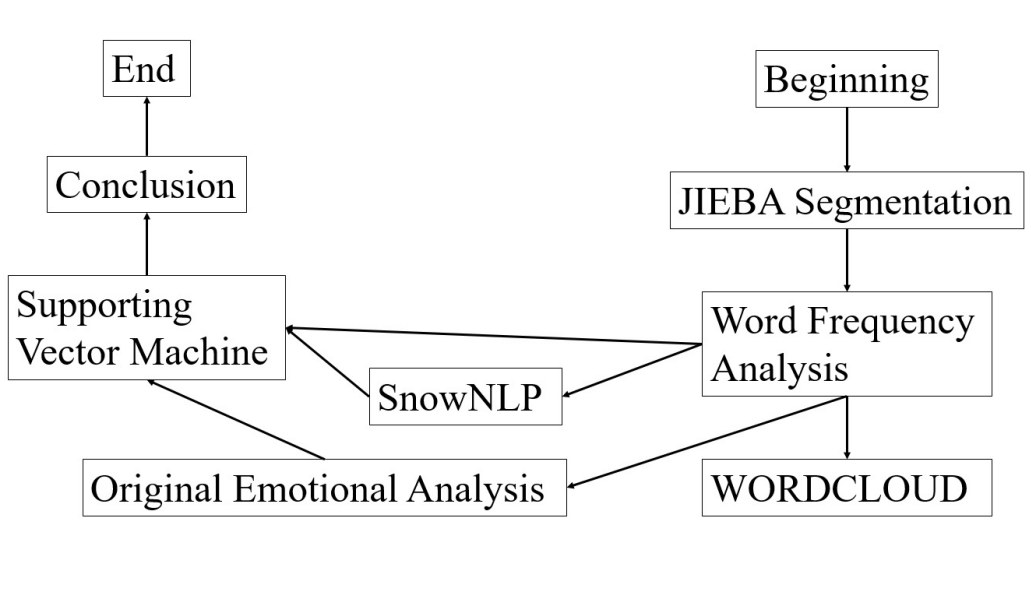


Figure 1: 工作流程图

counts the frequency of the words as the cornerstone of the whole process. Next, this paper employs two distinct techniques to define and determine how positive or negative pieces of data are and examines the consistency of both techniques, which ensures that the degree of positive and negative are solid. This paper applies Supporting Vector Machine to the data to obtain the words that are critical to judge whether each data is positive or negative in order to filter out the most popular activities, or the activities that leave the deepest impression on students. The result of SVM is also verified by word frequency analysis. Finally this paper analyzes the positive and negative aspects of the popular activities and reaches the conclusion and offers some recommendation to the supervisor of the school on the welcomed activities.

2 Data Extraction

With the help of my advisor and my school, this paper obtains the data in the CQE system of grade G17(senior 2 in 2018-2019 academic year) in Tsinghua University High School from February, 2018 to June, 2018. Each pieces of data is an activity uploaded by the students. The data can be searched in the appendix. This paper has 6 dimensions to study, which are PE Practices(体育类实践活动), Innovative Achievement(创新成果), Excursion Experiences(游学经历), Club Events(社团活动), Artistic Achievements(艺术成果展示), and Artistic Practices(艺术类实践活动). The dataset covers activities

ranging from hiking, skating, field research, excursions, trips, public benefit activities, photographing, musical instruments, so and so forth.

3 JIEBA Word Segmentation

3.1 JIEBA introduction

JIEBA module plays a fundamental role in the research. [11] Since the original data is in Chinese rather than English, this paper need to divide sentences into separate words. In English, spaces mark the separation between words; in Chinese, this paper need to use other technique.

Mechanical word segmentation algorithm based on string matching. Match the string to be split with the entries in a sufficiently large machine dictionary. It can be divided into forward matching and reverse matching, maximum length matching and minimum length matching, and an integrated method combining simple word segmentation and word segmentation with tagging process. So the commonly used methods are: forward maximum matching, reverse maximum matching and least segmentation. In practical application, mechanical word segmentation is used as the initial segmentation method, and language information is used to improve the accuracy of segmentation. Priority is given to the recognition of words with obvious features. With these words as breakpoints, the original string is divided into smaller strings and then matched mechanically to reduce the matching error rate or to combine word segmentation with part-of-speech tagging.

The more adjacent words appear at the same time, the more likely they are to form a word. The frequency of words in the text is counted. The word segmentation method based on the frequency statistics of words is a full segmentation method. JIEBA is a statistical word segmentation method which uses dynamic programming to find the maximum probabilistic path and find the maximum word frequency-based grouping combination.

The basic principle of JIEBA module is as such. It has a word library which contains all the word in. The module make a match from the target sentence to the word library. As long as a word exist in the sentence, the word becomes an output of the word segmentation result, which is called *whole mode*(全模式). It also have another method which is *accurate mode*(精确模式), which aims to cut the sentence as accurate as possible.

Whole Mode: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学.

Accurate Mode: 我/ 来到/ 北京/ 清华大学.

力学竞赛
创新英语
圣陶
培文
数理化
应用数学
高一数学
东方少年
钱学森
钱班
高研

Table 1: Stopwords in JIEBA

JIEBA can also remove the stopwords. Removing words means to have a file to store articles to be changed, a file to store the deactivated table, and then compare with the words in the deactivated table, the same words are deleted, and finally the results are stored in a file. Stopwords can effectively solve the problem of getting invalid keywords.

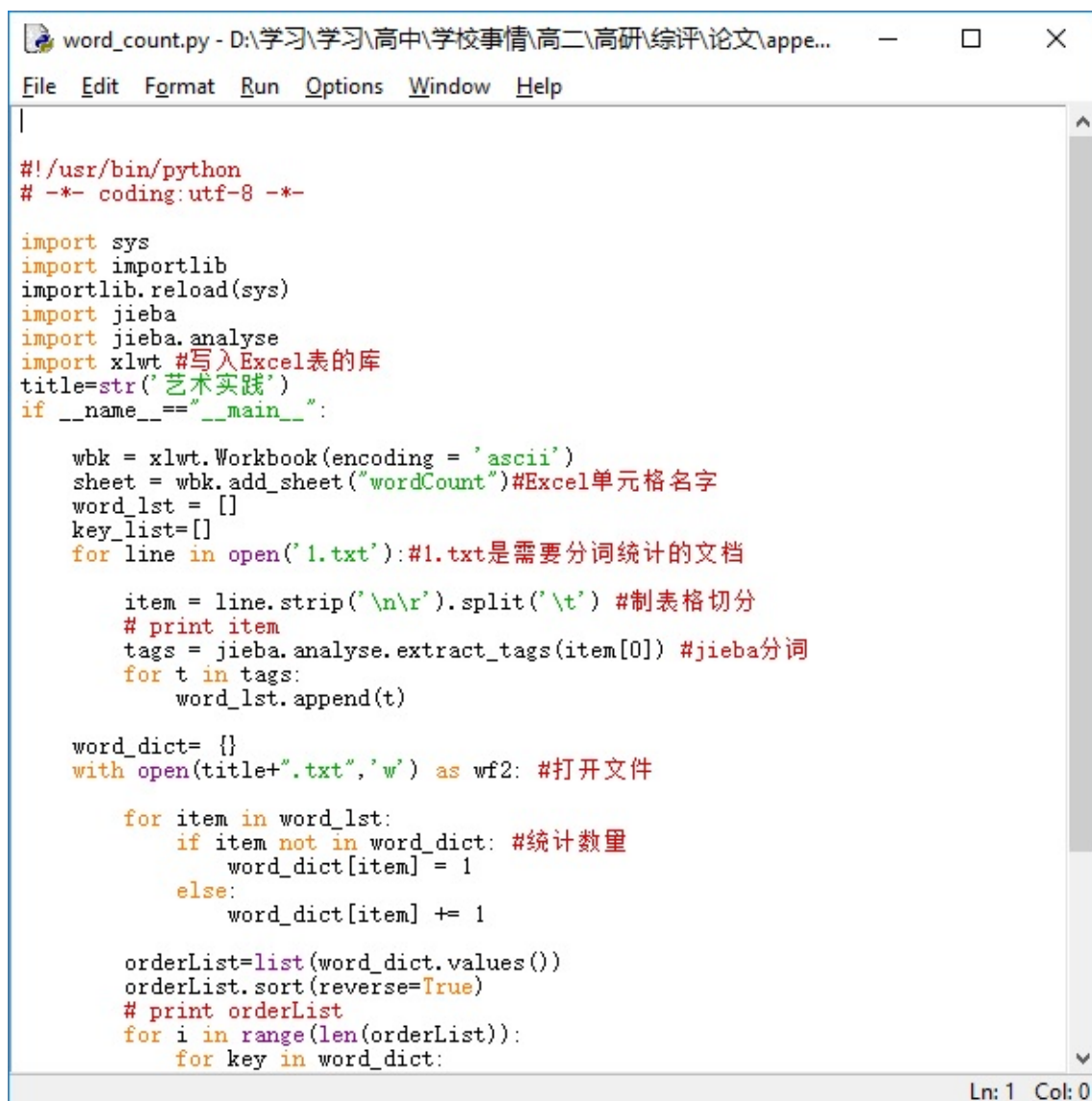
3.2 JIEBA Word Frequency Analysis

This paper apply JIEBA segmentation and word frequency analysis to the data and find out the most welcomed activities. The stopwords are shown as follows:

With the stopwords, part of the coding is shown as following figure 2.

Part of the results are shown as following figure 3-8. The full results are shown in the appendix. Meanwhile, this paper also made bar charts to show the results directly.

In the figures shown above, we can see that in the academic activities, the competitions take a major proportion, including the composition, mathematics, and English. Among all of the clubs, the chorus and the debate club are the most commonly seen clubs. The sports games, basketball match, and the soccer match are the most frequent activities. Art festival offer students an ideal platform for the students to show their talent. Photography and painting are extremely popular.



```
word_count.py - D:\学习\学习\高中\学校事情\高二\高研\综评\论文\appe...
File Edit Format Run Options Window Help

#!/usr/bin/python
# -*- coding:utf-8 -*-

import sys
import importlib
importlib.reload(sys)
import jieba
import jieba.analyse
import xlwt #写入Excel表的库
title=str('艺术实践')
if __name__=="__main__":

    wbk = xlwt.Workbook(encoding = 'ascii')
    sheet = wbk.add_sheet("wordCount")#Excel单元格名字
    word_lst = []
    key_list=[]
    for line in open('1.txt'):#1.txt是需要分词统计的文档

        item = line.strip('\n\r').split('\t') #制表格切分
        # print item
        tags = jieba.analyse.extract_tags(item[0]) #jieba分词
        for t in tags:
            word_lst.append(t)

    word_dict= {}
    with open(title+".txt",'w') as wf2: #打开文件

        for item in word_lst:
            if item not in word_dict: #统计数量
                word_dict[item] = 1
            else:
                word_dict[item] += 1

        orderList=list(word_dict.values())
        orderList.sort(reverse=True)
        # print orderList
        for i in range(len(orderList)):
            for key in word_dict:
```

Ln: 1 Col: 0

Figure 2: Coding of wordcount screenshot

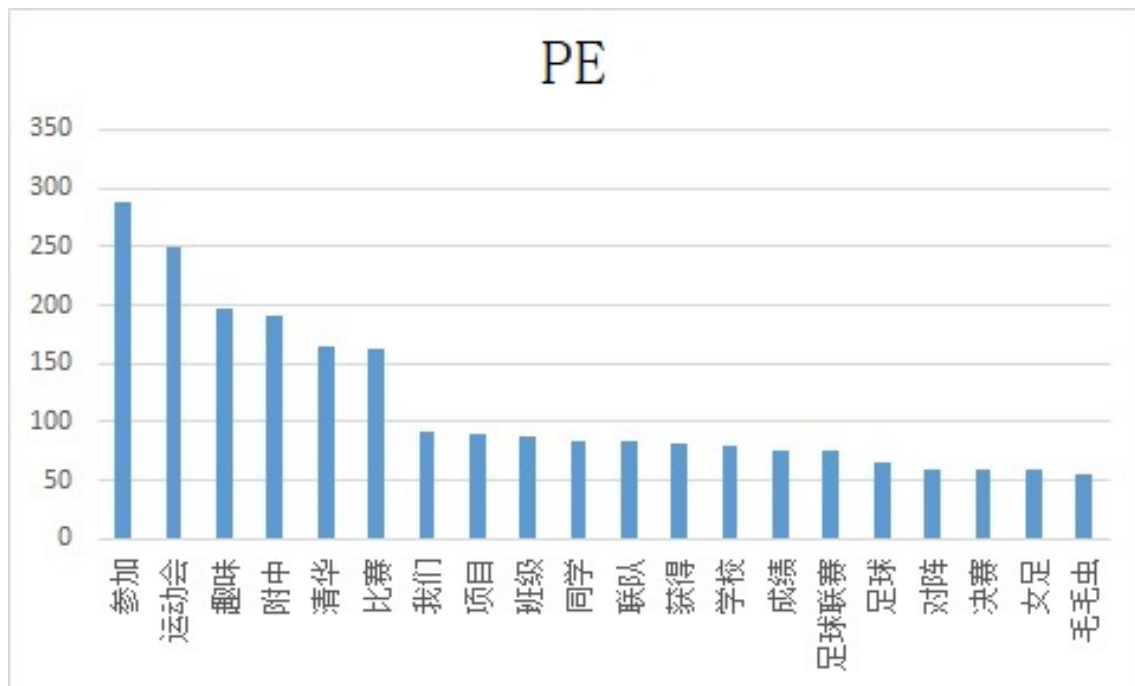


Figure 3: Bar chart of PE

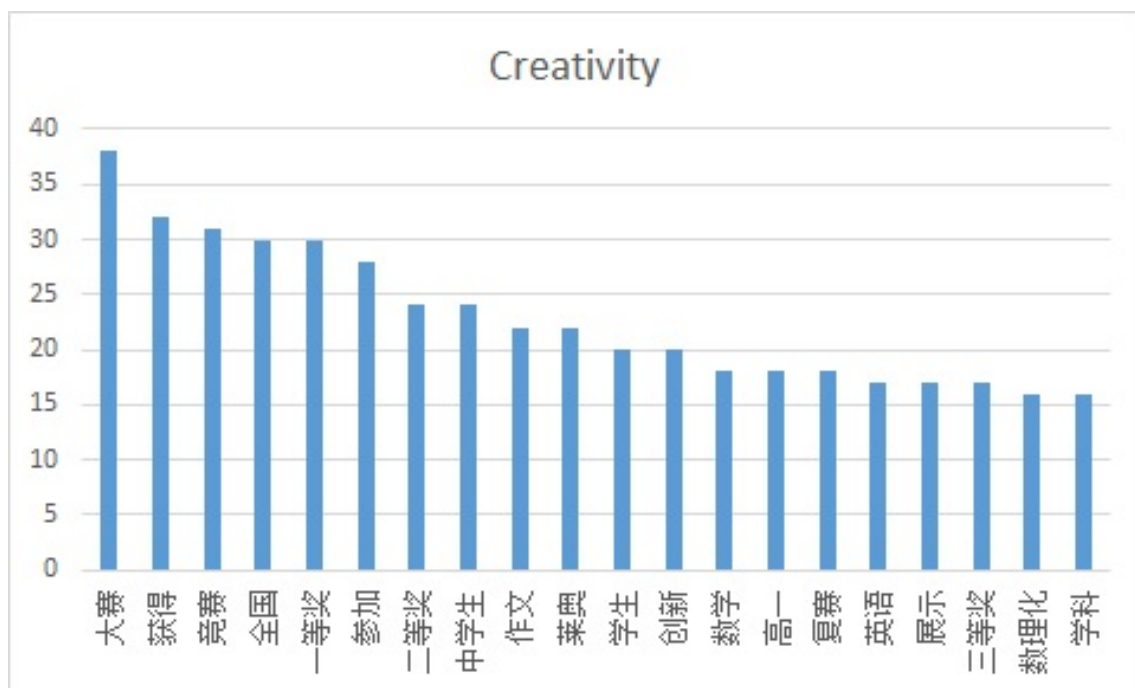


Figure 4: Bar chart of creativity

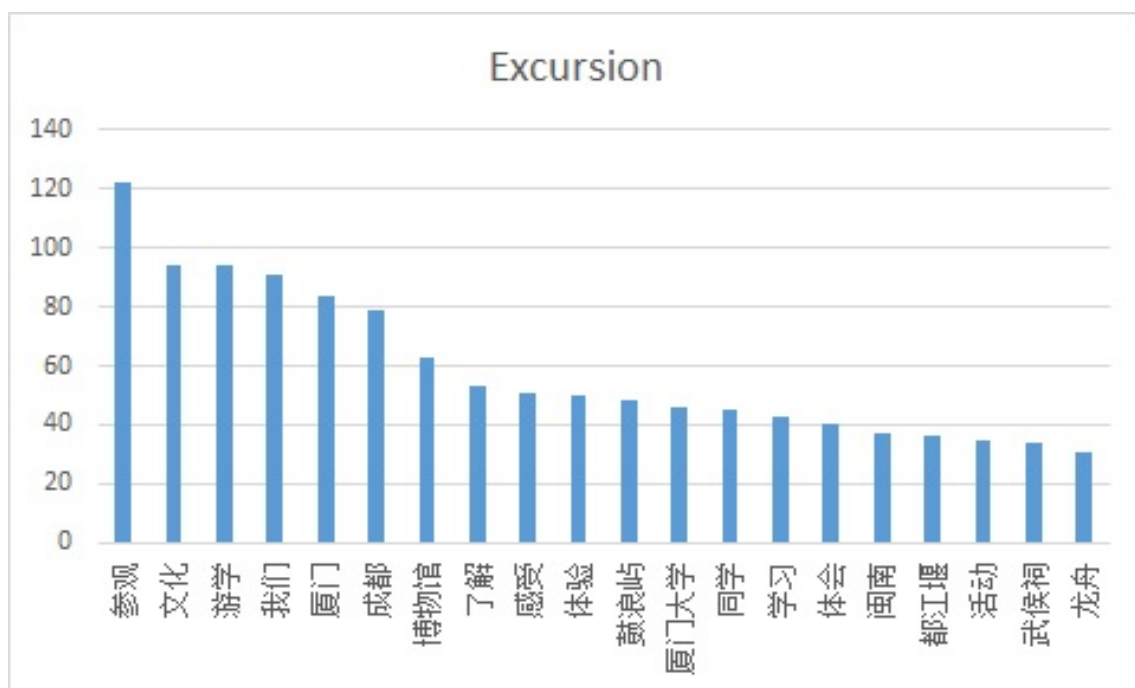


Figure 5: Bar chart of excursion



Figure 6: Bar chart of club

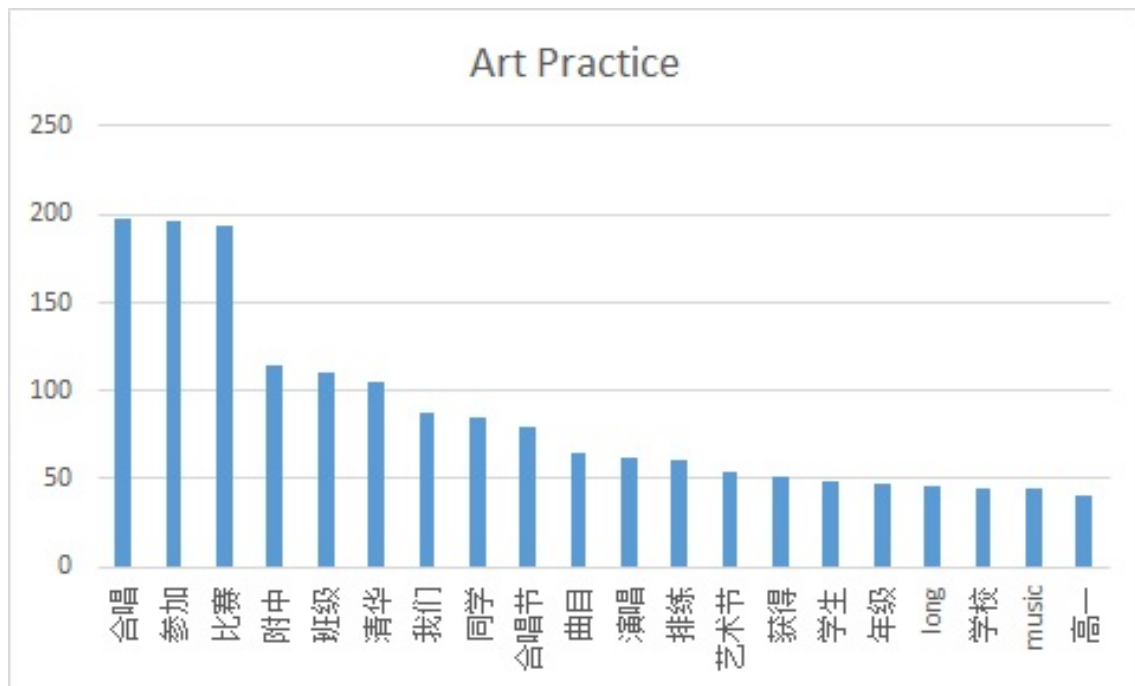


Figure 7: Bar chart of artistic practices

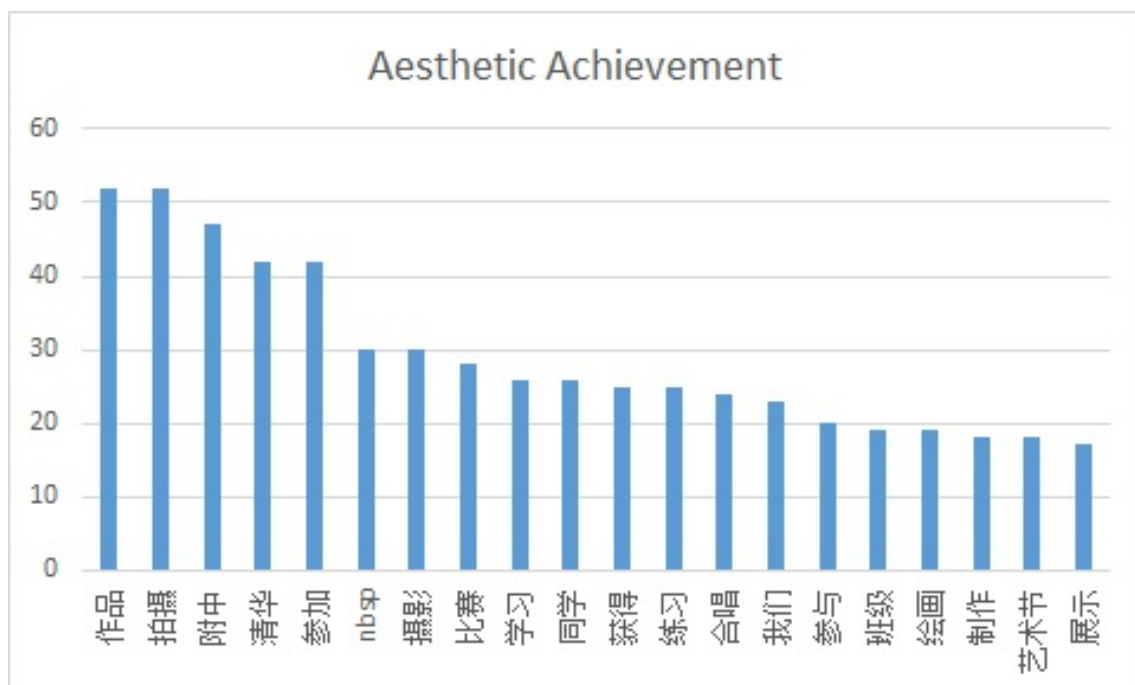


Figure 8: Bar chart of aesthetic achievements

4 WORDCLOUD

4.1 WORDCLOUD Introduction

WORDCLOUD module in PYTHON is mainly used to serve as a tool for rapid reading. [12] It can not only extract keywords, but also obtain sentences. Based on the keywords extracted by JIEBA module, it generates wordclouds which vividly shows the frequency of the words. The major advantage of the module is that the user is able to set the shape and the color of the background according to the interest of the users. In this way, the users no longer need to read the words, but the pictures instead, which is likely to be favored.

With the aid of WORDCLOUD module, this paper generates wordclouds through the set of parameters and load the data. The coding part of WORDCLOUD module is shown as following figure 9

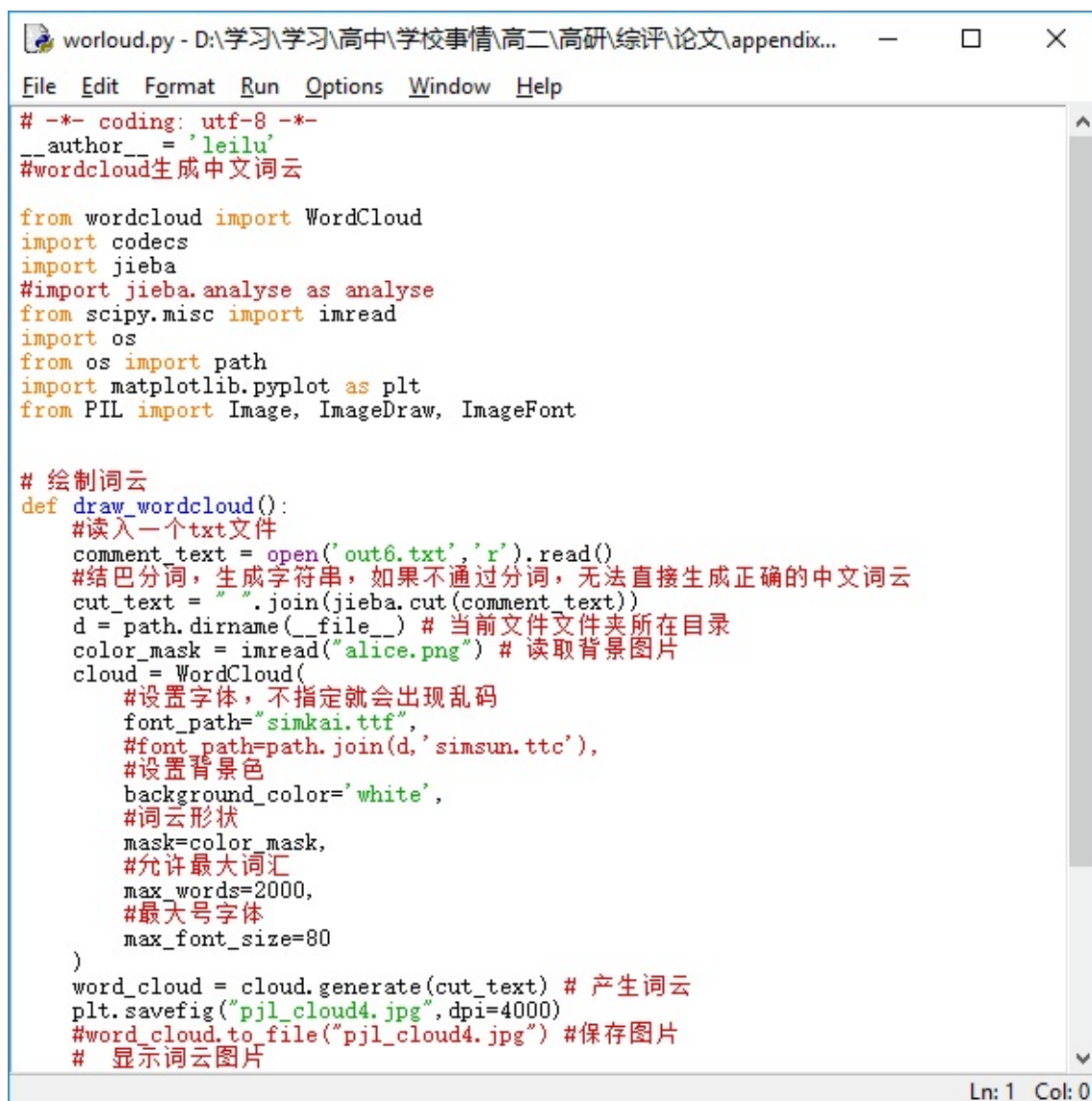
4.2 WORDCLOUD Production

This paper first convert the format of the source file from UTF-8 into ANSI. Then this paper obtain current path of the files, read the text *alice.txt* under the category *example*, and set background picture, color, maximum word number and maximum character size. The maximum number of words is 2000. Finally, this paper input the ensemble of the text, generate the RGB value from the background, and save the wordcloud.

In the figures 10 - 15 shown above, we can see that in the academic activities, the competitions take a major proportion, including the composition, mathematics, and English. Among all of the clubs, the chorus and the debate club are the most commonly seen clubs. The sports games, basketball match, and the soccer match are the most frequent activities. Art festival offer students an ideal platform for the students to show their talent. Photography and painting are extremely popular.

5 Emotional Evaluation

Emotional evaluation is to analyze whether a sentence is subjective or objective, and whether it is expressed in a positive way or negative way.



```
wordcloud.py - D:\学习\学习\高中\学校事情\高二\高研\综评\论文\appendix...
File Edit Format Run Options Window Help

# -*- coding: utf-8 -*-
__author__ = 'leilu'
#wordcloud生成中文词云

from wordcloud import WordCloud
import codecs
import jieba
#import jieba.analyse as analyse
from scipy.misc import imread
import os
from os import path
import matplotlib.pyplot as plt
from PIL import Image, ImageDraw, ImageFont

# 绘制词云
def draw_wordcloud():
    #读入一个txt文件
    comment_text = open('out6.txt', 'r').read()
    #结巴分词，生成字符串，如果不通过分词，无法直接生成正确的中文词云
    cut_text = " ".join(jieba.cut(comment_text))
    d = path.dirname(__file__) # 当前文件文件夹所在目录
    color_mask = imread("alice.png") # 读取背景图片
    cloud = WordCloud(
        #设置字体，不指定就会出现乱码
        font_path="simkai.ttf",
        #font_path=path.join(d, 'simsun.ttc'),
        #设置背景色
        background_color='white',
        #词云形状
        mask=color_mask,
        #允许最大词汇
        max_words=2000,
        #最大号字体
        max_font_size=80
    )
    word_cloud = cloud.generate(cut_text) # 产生词云
    plt.savefig("pjl_cloud4.jpg", dpi=4000)
    #word_cloud.to_file("pjl_cloud4.jpg") #保存图片
    # 显示词云图片
```

Ln: 1 Col: 0

Figure 9: Coding of WORDCLOUD module



Figure 10: Wordcloud of PE

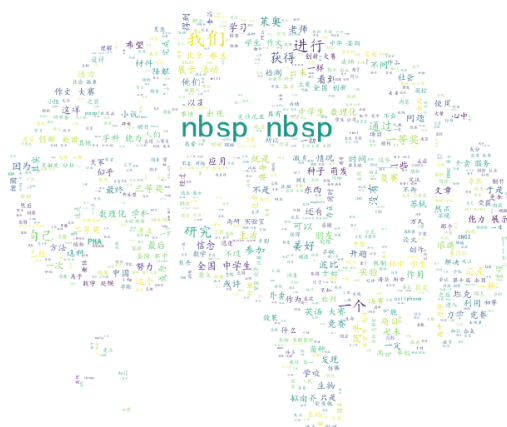


Figure 11: Wordcloud of creativity

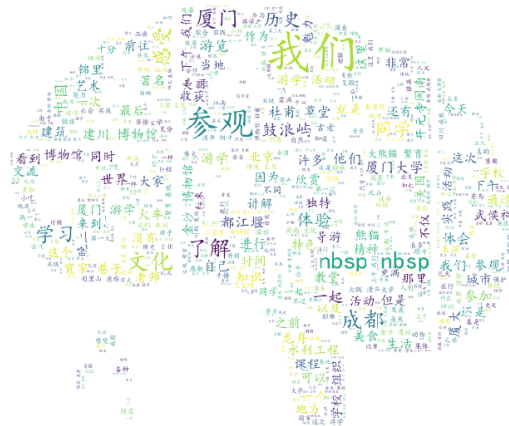
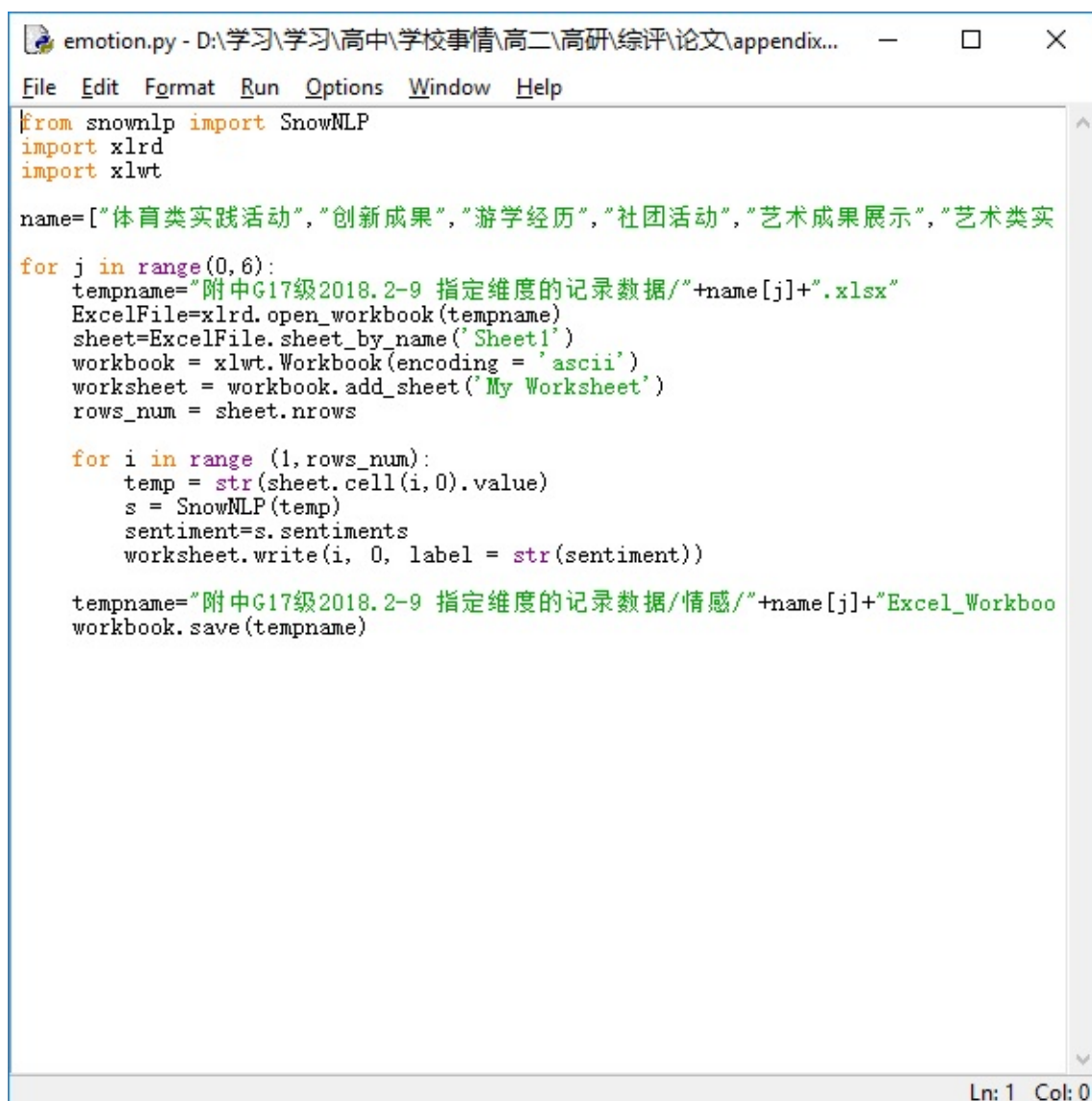




Figure 14: Wordcloud of artistic practices



Figure 15: Wordcloud of aesthetic achievements



```
emotion.py - D:\学习\学习\高中\学校事情\高二\高研\综评\论文\appendix...
File Edit Format Run Options Window Help

from snownlp import SnowNLP
import xlrd
import xlwt

name=["体育类实践活动","创新成果","游学经历","社团活动","艺术成果展示","艺术类实

for j in range(0,6):
    tempname="附中G17级2018.2-9 指定维度的记录数据/"+name[j]+".xlsx"
    ExcelFile=xlrd.open_workbook(tempname)
    sheet=ExcelFile.sheet_by_name('Sheet1')
    workbook = xlwt.Workbook(encoding = 'ascii')
    worksheet = workbook.add_sheet('My Worksheet')
    rows_num = sheet.nrows

    for i in range (1,rows_num):
        temp = str(sheet.cell(i,0).value)
        s = SnowNLP(temp)
        sentiment=s.sentiments
        worksheet.write(i, 0, label = str(sentiment))

    tempname="附中G17级2018.2-9 指定维度的记录数据/情感/"+name[j]+"Excel_Workboo
    workbook.save(tempname)
```

Ln: 1 Col: 0

Figure 16: Coding of SnowNLP

5.1 SnowNLP

SnowNLP is a module which supports Chinese Natural Language Processing. [13] It utilizes Bayes model to train the data that has been preset in the module. The data this paper inputs functions as a test set of the model. I read the file in the training set and utilizes the function `s.sentiments` to find the score of the emotional analysis. Part of the coding is shown in the following figure 16

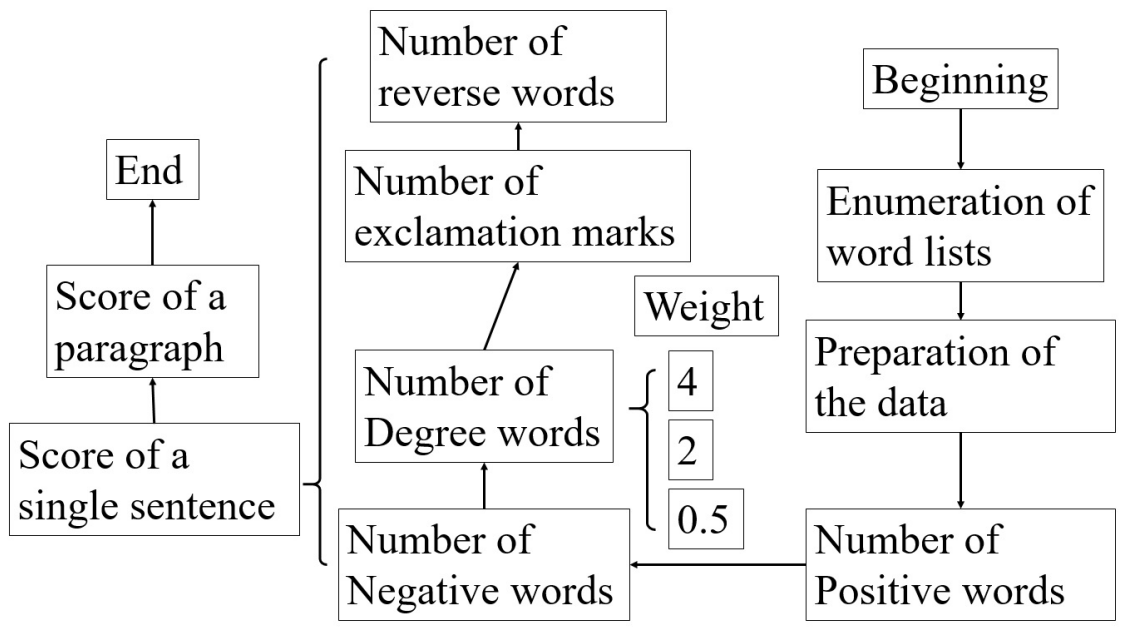


Figure 17: flow chart of original evaluation method

5.2 Original Emotional Evaluation Method

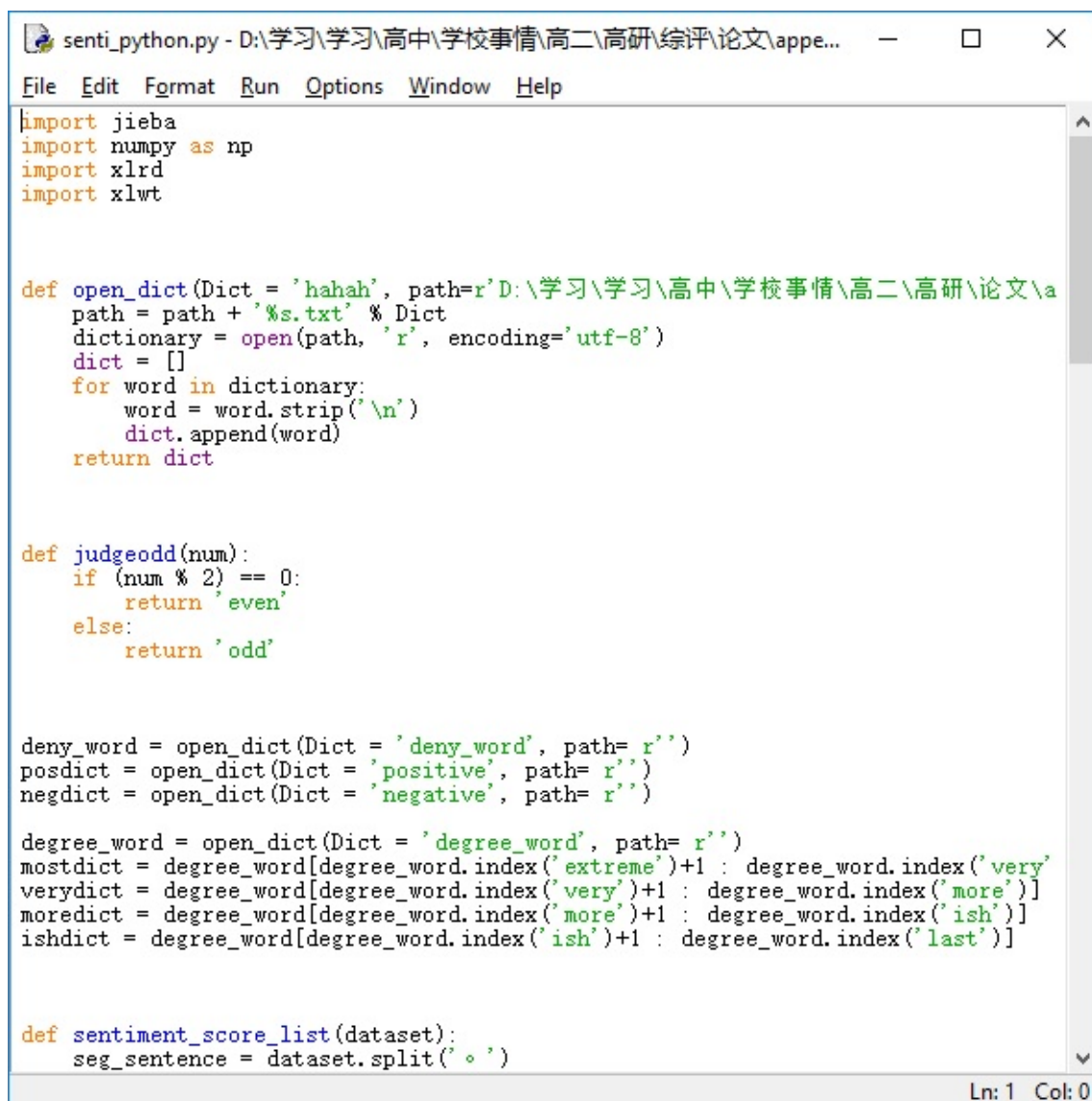
This paper has also created an original method to evaluate the emotional quantity of sentences. The flow chart is shown following figure 17:

Part of the coding is shown in the following figure 18

5.2.1 Enumeration of word lists and Preparation of the data

First, this paper enumerates a positive dictionary and a negative dictionary. The positive emotional words are such: fine, good, magnificent, gorgeous, so and so forth; the negative emotional words such as: bad, ill, wicked, so and so forth. The dataset is searched online. [14] Then this paper uses JIEBA module to segment the words in the sentences that this paper wants to evaluate. This paper also enumerates a set of degree words, which may enhance the meaning or weaken the meaning. Degree words contain extreme, very, much, so and so forth. This paper also contains a set of reversion words, which shows the reversion of the meanings such as however, nevertheless, so and so forth.

First, this paper reads the data in the dataset and divide them by sentences according to periods. In this way, each paragraphs is cut into sentences. This paper also loads the 4 word dictionaries into the program. This paper goes through all the pieces of data by turns.



```
senti_python.py - D:\学习\学习\高中\学校事情\高二\高研\综评\论文\appe...
File Edit Format Run Options Window Help

import jieba
import numpy as np
import xlrd
import xlwt

def open_dict(Dict = 'hahah', path=r'D:\学习\学习\高中\学校事情\高二\高研\论文\ap...
    path = path + '%s.txt' % Dict
    dictionary = open(path, 'r', encoding='utf-8')
    dict = []
    for word in dictionary:
        word = word.strip('\n')
        dict.append(word)
    return dict

def judgeodd(num):
    if (num % 2) == 0:
        return 'even'
    else:
        return 'odd'

deny_word = open_dict(Dict = 'deny_word', path= r'')
posdict = open_dict(Dict = 'positive', path= r'')
negdict = open_dict(Dict = 'negative', path= r'')

degree_word = open_dict(Dict = 'degree_word', path= r'')
mostdict = degree_word[degree_word.index('extreme')+1 : degree_word.index('very')]
verydict = degree_word[degree_word.index('very')+1 : degree_word.index('more')]
moredict = degree_word[degree_word.index('more')+1 : degree_word.index('ish')]
ishdict = degree_word[degree_word.index('ish')+1 : degree_word.index('last')]

def sentiment_score_list(dataset):
    seg_sentence = dataset.split('。')
```

Ln: 1 Col: 0

Figure 18: Original Emotional Evaluation Method

5.2.2 Definition of scores of emotional words

This paper makes 4 categories of words and punctuations for detection: emotional words, degree words, exclamation marks, and reverse words.

This paper devises a set of methods to define positive and negative. This paper will finally give out two scores to demonstrate the positive and negative degree, representing positive score and negative score respectively.

If there is a positive word, the value of positive score will add 1. If there is a negative word, the value of negative score will add 1.

This means that this paper finds the emotional words of the clause and records the positive or negative words and their position, which means in which sentence they are located in. This paper defines a variable, *i*, to save the current place of the words, a variable, *a*, to save the place of the emotional words, and a variable, *poscount*, to save the score of the positive words. This paper creates an algorithm to match current word with the word in dictionary to determine whether it is an emotional word. This paper creates two individual functions to judge positive words and negative words respectively.

5.2.3 Definition of scores of degree words

“Excellent” is much stronger than “good”, and “too bad” is much stronger than “a little bad”. So this paper finds out whether there is a degree of modification and gives a weight to different degrees after we find the emotional words. For example, words like “extreme”, “inestimable”, “too” will have an emotional score multiplied by 4. “Comparative” has an emotional score of multiplied by 2. Words like “A little bit” have a score multiplied by 0.5.

This paper finds the degree words before the emotional words. Then this paper sets weights for degree words and multiplies them by emotional values. This paper utilizes another variable, *poscount2*, to save the emotional score of currents after multiplied by the weight of degree.

5.2.4 Definition of scores of exclamation marks

Exclamation marks mean strong emotions. Therefore, if an exclamation mark is found in a sentence, the positive value will be added by 2.

This paper judges whether there is an exclamation mark at the end of the clause. If there is an exclamation mark, the corresponding emotional value, which is the *poscount2*, is added by 2.

5.2.5 Definition of scores of reverse words

Sometimes the word “good” does not mean “good”, because there is a “no” in front of it. When this paper finds emotional words, it looks forward for reverse words, such as “nevertheless”. This paper counts the number of times that these reverse words appears. If they are singular, the emotional score of the word will be multiplied by - 1; But if they are even, the emotions will not be reversed, and the emotion score of the word remains to be 1.

This paper finds the reverse words before the emotional words. If the number of the reverse words is odd, the score of the word is multiplied by - 1. If it is even, the score of the word will remain unchanged. This paper defines poscount3 to save the emotional score of the words after reversion. And then this paper moves the current location of the scanning word to one word forward.

5.2.6 Definition of scores of a whole paragraph

If the positive emotional score is negative, it will be added to the negative emotional score and the positive emotional score will be 0 and vice versa.

Finally this paper finds the mean of the positive score and negative score of each words respectively and obtain the final two scores of a paragraph.

The reason why this paper does no use a single score to show the positive and negative score of the paragraph is that there are two aspects in a sentence which cannot be expressed by a single score. And the setting of this weight will also affect the final emotional score. Therefore, the final treatment of this sentence is to get a positive score and a negative score of the paragraph.

5.3 Cross-test between SnowNLP and Original Emotion Evaluation Method

This paper examines the correlation coefficients between the score of SnowNLP and the positive score of Original Emotion Evaluation Method, shown in the following table. This paper also examines the correlation coefficients between the score of SnowNLP and the negative score of Original Emotion Evaluation Method, shown in the following table 2.

To manifest that the parameters set in previous paragraphs are appropriate, this paper adjusts the weight set in each paragraph. When the weight of degree words is set by a times of 4, this paper obtain the following results in the table 3

PE Practices	0.05883	-0.09246
Innovative Achievements	0.084022	-0.02478
Excursion Experiences	0.031351	-0.09307
Club Events	0.053741	-0.07059
Artistic Achievements	0.093259	-0.02783
Artistic Practices	0.083377	-0.04228

Table 2: Correlation Coefficients

PE Practices	0.066125	-0.07219
Innovative Achievements	0.038332	0.00614
Excursion Experiences	0.020345	-0.09307
Club Events	0.054993	-0.05287
Artistic Achievements	-0.019932	-0.06625
Artistic Practices	0.036731	0.01120

Table 3: Correlation Coefficients at the time of 4

From the table, we can see that if this paper utilizes the parameters listed above, although the correlation coefficients are not relatively high, they can manifest a positive or negative relationship consistently. If this paper revises the parameters, the result cannot demonstrate a positive or negative tendency of the relationship. Therefore, the result of the two methods are coherent and this paper takes the advantage of the parameters set previously.

6 Supporting Vector Machine

6.1 Analysis of SVM classification

Supporting Vector Machine (SVM) is an algorithm which has great advantages in solving non-linear problems. It maps low-dimensional data into high-dimensional space and finds the optimal hyperplane as the criterion for classifying given data. Its basic principle is to use a hyperplane to separate some existing sample points as a binary classification method. [15]

In this paper, the positive and negative data that have been graded are used as the training set of support vector machine, all of which are used as the test set. The positive and negative data are labeled in the previous part. This paper utilizes the mean of the

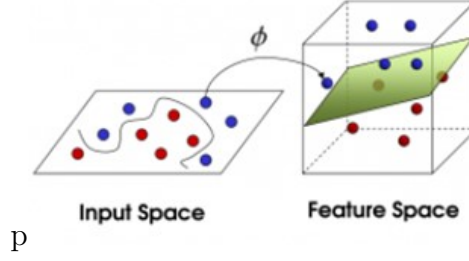


Figure 19: Diagram of SVM

two methods. To obtain a single score in the latter method, this paper subtracts the negative score from the positive score. Then this paper standardizes the score of both methods, of which the mean is 0 and the variation is 1 respectively. If the score of a piece of data is negative, then this paper defines it as negative; otherwise this paper defines it as positive. Thus, this paper obtains the keywords.

In this paper, PYTHON is used to calculate the solution of SVM. The diagram of SVM is shown in the figure below.

6.2 Analysis of Timing Function

This paper creates a function for time the total runtime of the algorithm. It uses two variables to save the beginning time of the algorithm and the finishing time of the algorithm, finding the difference to obtain the total runtime.

6.3 Analysis of SVM classification

This paper utilizes svm model in sklearn module to accomplish the goal. This paper chooses RBF as the kernel function for the number of sample is much larger than the number of traits. This paper also prints the recall rate, precision rate, and the f1-score of the model. Recall rate refers to how many real positive samples are there in positive samples selected; precision rate refers to how many positive samples are selected in all the real samples; f1-score can be calculated by the following formula.

$$f1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

6.4 Analysis of Bayes Distinction

This paper also considers using Bayes Distinction to replace SVM, of which the principles are as follows.

Bayes Distinction ideally satisfies the requirements of such issue that each individual of the ensemble exists at different frequencies, which indicates that this paper needs to take into consideration that the different possibilities that each individual exists. This paper utilizes GaussianNB in `sklearn.naive_bayes` to achieve the goal. Distance distinction does not take into account the frequency of each sample in the whole and does not take into account the loss caused by the wrong distinction. The Bayes distinction method modifies on the basis of distance distinction, and the formula is defined as in formula. [16]

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum P(A|B_j)P(B_j)}$$

Among which $P(B_i|A)$ represents a posteriori probability; $P(A|B_i)$ represents a prior probability; $P(B_i)$ represents the frequency at which the sample appears; Σ represents the total covariance matrixes. The distinction rule is that the posterior probability is the highest and the average wrong distinction loss is the lowest, which brings out the rule is as follows: If the condition meets the following formula :

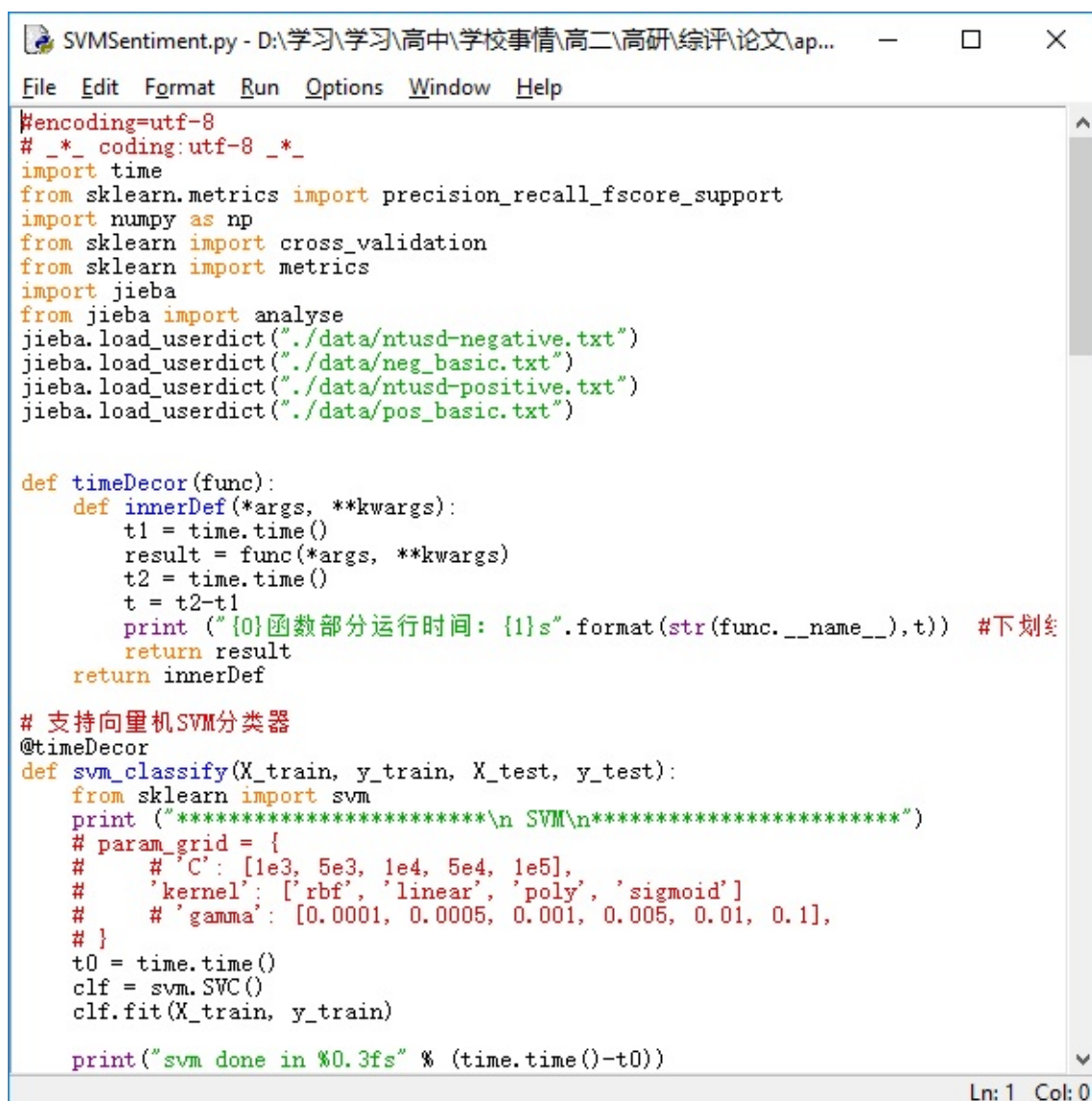
$$P(G_l|x_0) = \frac{p_l f_l(x_0)}{\sum p_j f_j(x_0)} = \max_{1 \leq i \leq k} \frac{p_i f_i(x_0)}{\sum p_j f_j(x_0)}$$

Then we categorize x_0 into G_l , among which G_i is the ensemble, $f(x)$ is the probability density function of G_{i2} , p_i is prior probability of G_i , which is the probability that it belongs a certain category when sample x_0 occurs, and k is the number of G_i . The solution formula for distinction analysis is as the following formulas.

$$ECM = \sum_{i=1}^k p_i \sum_{j \neq i} C(\frac{j}{i}) P(\frac{j}{i})$$

$$p(\frac{j}{i}) = P(X \in D_j/G_i) = \int_{D_j} f_i(x) dx \quad i \neq j$$

In this case, $P(\frac{j}{i})$ represents the conditional probability of wrongly categorizing the sample of G_i to the ensemble $G_j(\frac{j}{i})$ is the loss caused by this categorization. D_k is a division of a set of distinction samples. ECM is the average wrong distinction loss. The solution to a Bayes distinction analysis is to make the smallest set of solutions.



```
#encoding=utf-8
# *_ coding:utf-8 *_
import time
from sklearn.metrics import precision_recall_fscore_support
import numpy as np
from sklearn import cross_validation
from sklearn import metrics
import jieba
from jieba import analyse
jieba.load_userdict("./data/ntusd-negative.txt")
jieba.load_userdict("./data/neg_basic.txt")
jieba.load_userdict("./data/ntusd-positive.txt")
jieba.load_userdict("./data/pos_basic.txt")

def timeDecor(func):
    def innerDef(*args, **kwargs):
        t1 = time.time()
        result = func(*args, **kwargs)
        t2 = time.time()
        t = t2-t1
        print (" {0}函数部分运行时间: {1}s".format(str(func.__name__),t)) #下划线
        return result
    return innerDef

# 支持向里机SVM分类器
@timeDecor
def svm_classify(X_train, y_train, X_test, y_test):
    from sklearn import svm
    print ("*****\n SVM\n*****")
    # param_grid = {
    #     # 'C': [1e3, 5e3, 1e4, 5e4, 1e5],
    #     # 'kernel': ['rbf', 'linear', 'poly', 'sigmoid']
    #     # 'gamma': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1],
    # }
    t0 = time.time()
    clf = svm.SVC()
    clf.fit(X_train, y_train)

    print("svm done in %0.3fs" % (time.time()-t0))
```

Figure 20: Coding of SVM

From the result, it can be seen that the accuracy of Bayes distinction is worse than that of SVM, which is the reason why this paper does not consider to use Bayes distinction.

6.5 Final Results

Part of the coding is shown as the following figure 20

The extracted keywords in this paper are in the appendix. This paper selects the keywords related to activities and get the results as follows:

PE Practices: 运动会 足球 篮球

Innovative Achievements: 创新英语 高研 力学

Excursion Experiences: 成都 巷子 龙舟 宽窄 四川 大熊猫 华侨大学 三星堆 青城山
川剧 变脸 火锅 草堂 熊猫

Club Events: 社团 辩论赛 艺术节 明哲 模联

Artistic Achievements: 李宁杯 电影节 电视台

Artistic Practices: 合唱节 艺术节 电声乐

6.6 Cross-test between SVM and Word Frequency

This paper does a verification of the words filtered out above and the frequency of the words in previous parts. The results are as following table 4:

It can be seen that the critical words above are of high word frequency. Moreover, this paper also adds some words with high word frequency into the critical words list. Therefore, the critical words, which represent the most welcomed activities among students, are really the words that have a relatively bigger impact on students.

7 Evaluation of Positive and Negative Aspects

After this paper has picked out the hot activities with the aid of SVM, this paper search the pieces of data that contains the fields of the names of hot activities. Since there are not a large number of it, which is about 10-20 pieces of data each, this paper decides to find the positive and negative aspects of the activities manually. If there are too many pieces of data, this paper can count the word frequency again under the condition that the data contains a certain activity in order to find the aspects that ought to be developed or refrained.

8 Conclusion

8.1 Strength and Weakness

The method we propose in the paper has effectively made up the vacancy and deficiency of the previous evaluating process regarding the analysis of data in CQE. The CQE system has been implemented for over a year, while no research towards the data in CQE has been made, and several main advantages are as the following. For a start, it

运动会	249
足球	66
篮球	31
数理化	16
创新英语	17
高研	13
力学	13
厦门	84
成都	79
都江堰	36
武侯祠	34
龙舟	31
草堂	27
宽窄	23
华侨大学	15
大熊猫	25
四川	20
川剧	15
变脸	11
辩论	20
模联	34
艺术节	12
明哲	7
李宁杯	7
电影	13
电视台	6
合唱节	16
艺术节	18
合唱	198

Table 4: Word Frequency of Selected Words

analyzes the emotion of the data in CQE, the results of which are seldom considered by schools but actually of great significance. The leader of the school can take the emotion index from the ensemble of the students into consideration, deciding which aspects of the activities can bring positive emotion towards students and fit the need of the students. Furthermore, as the latter part of the paper indicates, the process we propose can further the understanding of the school towards the activities they hold. By using the method in this paper, they know the effect of the activities they hold based on quantitative data rather than on their ambiguous natural instinct. Besides the application of the evaluating process in real life, the original method of emotional evaluation is also more advanced and comprehensive than that in the previous thesis. For the method of cross test in two sections of the paper, they not only meet the exact needs of the data being processed and the expected outcome, but they are also more precise and reliable, ensuring the credibility of the research as a whole.

Admittedly, there are several shortcomings concerning the whole paper. For instance, limited by the volume of the data, the rating of the emotion is not accurate enough. Also, after the data related to certain key words has been extracted, this paper ought to find the word frequency separately for the positive data and the negative data. Since there are too less pieces of data, this paper merely find the characteristics of the data of low emotional value and the data of high emotional value manually, which means the results are not very desirable. However, considering the techniques being applied as a whole, the advantages outweigh the deficiencies, thus making the research reliable for reference and have high practical value.

8.2 Conclusion

This paper has discovered some intriguing and unexpected conclusion throughout the whole process.

- “Nervous”(紧张) is a word that is often mentioned in pieces of data that are more likely to be treated as negative in sports activities, which includes sports meeting, soccer match, and basketball match.
- If there are too less words in a piece of data, it will tend to be viewed as a negative one, because it contains too little information which means the hosts of the data may be less involved in the activity. For instance, if a person simply participates in the admission of a sports meeting or merely works as a staff, it is probable that the data is viewed as negative.

- If a person describe an experience as something like “visit”(参观), “observe”(观察), it will be more likely to be viewed as negative than that described as something like “participate”(参加), “involvement”(体验).
- Data related to innovative achievements, excursion experience, club events are very likely to be viewed as positive.
- A piece of descriptive data is more likely to be positive than a piece of narrative data.

This paper also discovered some conclusion that can easily be predicted.

- If someone loses a match, the piece of data will almost definitely viewed as negative.
- When describing a experience with words such as “fortunately”(有幸), “hilarious”(愉快), the experience is very probable to be positive.

In addition to the specific conclusion and some reasonable explanation, this paper also yields significant results in the following aspects. First, This paper proposes a brand new method based on dictionary to judge whether a paragraph is positive and negative, and the idea is turned into algorithms. When comparing that result with the result of a existing module, SnowNLP, this paper finds that the new method is effective. It can be used into further study and judge whether a paragraph is positive and negative. The new technique is an easily accessible method and produces a relatively reliable result.

Second, the research can be used to analyze the data in CQE systems and decide which aspects of a certain activity are the positive ones that the school should develop and which aspects of a certain activity are the negative ones that the school should hamper. The results reflect the tendency of students towards different types of activities, and their preference is carefully studied using the methods mentioned previously. What this paper develops is not only a result towards the CQE data in Tsinghua High School, but a set of algorithms that can be introduced into data from other school and even other data under new situations.

Third, the different aspects of activities give the administrators of school a broader view of which ones are the keys to promoting the quality of school activities and lay the stepping stone for the further optimization relating to the school activities. These methods enable school administrators to examine the activities they organize, and according to the cross data testing, the algorithm and results are reliable and can be applied for other practical uses.

9 Acknowledgement

I would like to express my gratitude to all the ones who helped me during the writing of this thesis. I acknowledge the help of my supervisor, Yuyao Zhang, who has offered me many suggestions in academic studies. In the preparation of the thesis, she has spent much time reading through each draft and provided me with advice. I also like to express my gratitude to my parents and peer students who have always been helping me out of difficulties.

10 Declaration

The paper submitted by author is the research and achievement under the guidance of the advisor. To the best of the knowledge of the author, the paper does not contain any research results that have been published or written by others except those listed in the bibliography. All the work are original unless listed. If there is any dishonest, the author willing to undertake all relevant responsibilities.

11 Bibliography

- [1] <https://gzzp.bjedu.cn>, Comprehensive Quality Evaluation Platform for Beijing General High School Students [EB/OL].
- [2] Ze L., Reflections on the Evaluation of Comprehensive Quality in the Reform of College Entrance Examination [J], Contemporary Educational Science, 2017(4): 32-36, 45.
- [3] Bingbing X., From Comprehensive Quality Assessment to Key Competencies Assessment—Research on the Transition of Student Assessment in High School [D]. Shanghai, China: East China Normal University, 2016.
- [4] Zhaohui C., Study on Implementation of Senior High School students' Comprehensive Quality Assessment [D]. Zhengzhou, China: Henan University, 2016.
- [5] Huoyun C. Zheng K., Research on Comprehensive Quality Evaluation of Senior High School Students [J], Global Education, 2010(9): 3-8,12.
- [6] Jing W. Xi J. Dong L., Evaluation of Students' Comprehensive Quality with the Aim of Promoting Development—The Concept and Practice of the Construction of the Second Class Transcript [J], Learning Resource and Technology, 2018(9): 132-137.
- [7] Xiaoming W. Nianjin D., History and Evolution of Ten Years Reform of the Comprehensive Quality Evaluation of Students in General Senior High School [J], Modern Education and Management, 2015(11): 74-79.
- [8] Zhijun L. Hongxia Z., Evaluation on Senior High School Students Comprehensive Quality: Reality, Problem and Prospect [J], Curriculum, Teaching Material and Method, 33(1): 18-23.
- [9] Long C., The Difficulty and Breakthrough of "Hard Link" between Comprehensive Quality Evaluation and College Enrollment [J], Journal of The Chinese Society of Education, 2017(7): 19-23.
- [10] Dianjun W. Hui J. Weidong Meng. The High School Attached to Tsinghua University, Development and Application of the Comprehensive Student Quality Evaluation System Based on Big Data: Innovative Practice of the High School Attached to Tsinghua University [J], Chinese Examinations, 2018(1): 46-52, 66.
- [11] https://blog.csdn.net/daniel_ustc/article/details/48195287, Source Code Analysis of Chinese Word Segmentation in JIEBA Module [EB/OL].
- [12] <https://www.cnblogs.com/bymo/p/9334981.html>, Keyword extraction based on WORDCLOUD Module: The use of wordcloud, source code analysis, generation of Chinese word cloud, and code rewriting [EB/OL].

[13] <https://blog.csdn.net/google19890102/article/details/80091502>, Emotional Analysis—Deepening the Principle and Practice of SnowNLP [EB/OL].

[14] <https://blog.csdn.net/jzy3711/article/details/84760981>, Positive and Negative Emotional Words [EB/OL].

[15] https://blog.csdn.net/qq_16953611/article/details/82414129, Sklearn—SVM [EB/OL].

[16] Haiwei W. Yu X. Yalin W. A bivariate hierarchical Bayesian approach to predicting customer purchase behavior [J], Journal of Harbin Engineering University, 2007, 28(8): 949-954.