Experiment Report

Big three

姚宇航、李大松、李洋

1. Prototype System Introduction
1.1 Functions :
Chinese parsing algorithms
1.2 Running Environment:
Windows 8 / 10
1.3 Developing Environment:
PyScripter 3.0

2 .Task Allocation：

Yao Yuhang: User interface component

Li Dasong:　　[1]. Kernel algorithm of dealing with Chinese Parsing algorithms.

[2]. The algorithm of dealing with the segmentation of sequence.

[3]. I think we should take in functions of consulting the meanings of Chinese words and calculate the frequency of new words by enough many essays.

[4]. I find the way of using python to browse the "www.baidu.com"

Li Yang:　　[1]construct lexicon

[2]combine the parts of system, unify the style of explanatory note and variables

[3]construct the lexicon of the new function--explain the meaning of a word

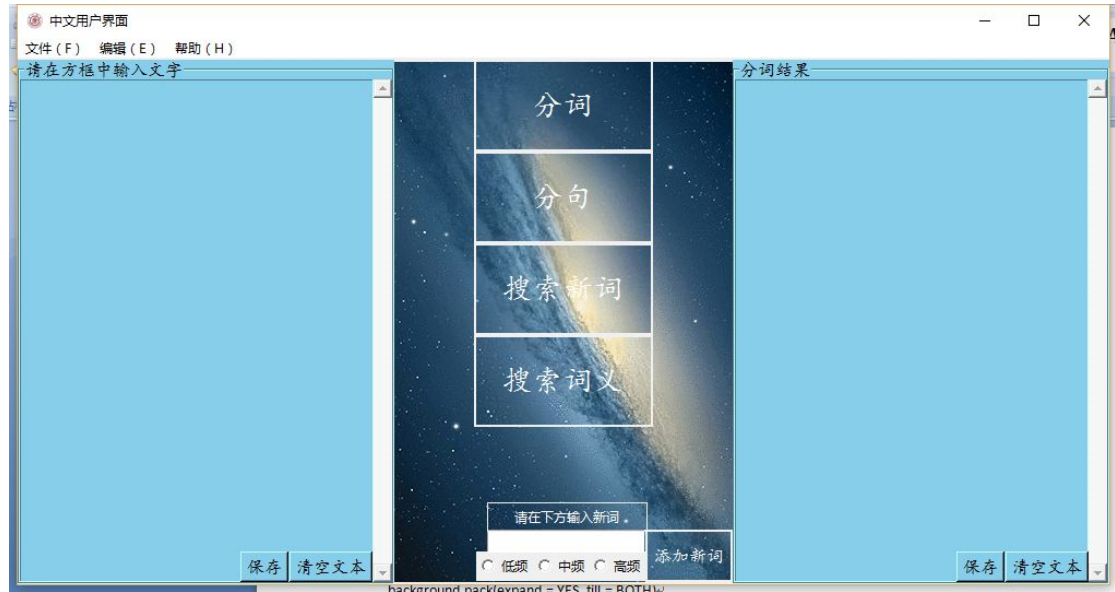[4]design the program which can find the new word from a text

[5]design the program which can delete the word of dictionary

[6]integration of system, system debugging and part of testing.

3.System Architecture and Algorithm Description
3.1User Interface Component
Main Window:

from tkinter import *

root = Tk()

root.mainloop()

Background

We can use an image as the background which covers the whole window.

Use canvas as a container of the image and use create_image to show the image and put it in a suitable place.

background=Canvas(root,height=1000,width=500)

image1=PhotoImage(file='100008742769036.gif')

background.create_image(300,-100,image =image1,anchor=NW)

background.pack(expand = YES, fill = BOTH)

LabelFrame

The LabelFrame can be used when you want to group a number of related widgets

Text

The Text widget provides formatted text display. It allows you to display and edit text with various styles and attributes. The widget also supports embedded images and windows.

Scrollbar

This widget is used to implement scrolled listboxes, canvases, and text fields.
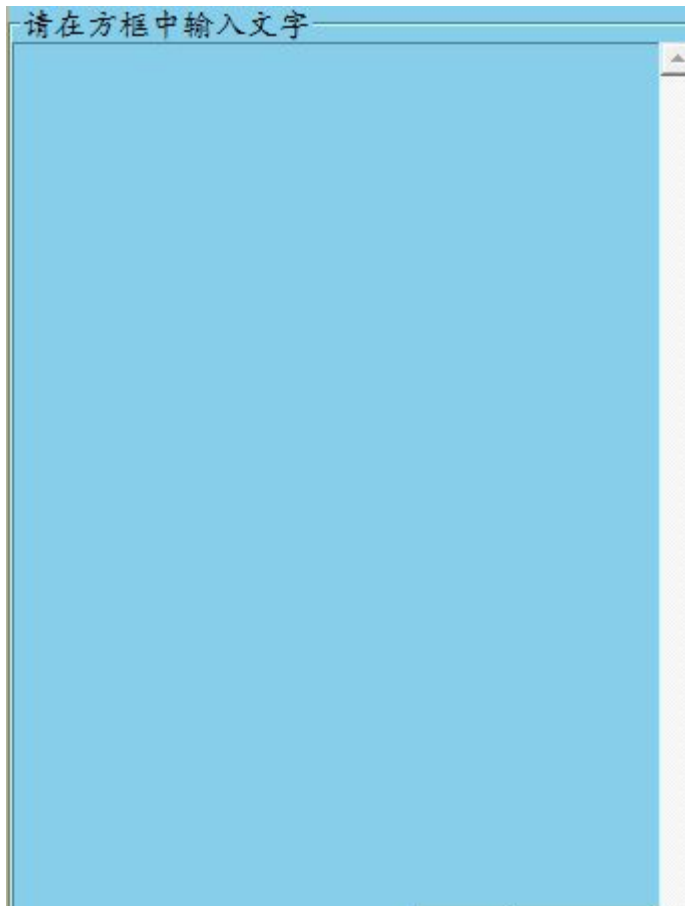
frm_L = LabelFrame(background,bg='SkyBlue',text="请在方框中输入文字", font=("楷体", 12), width=20, height=3)

scrollbar_LV=Scrollbar(frm_L, orient=VERTICAL)

scrollbar_LV.pack(side=RIGHT,fill=Y)

text1=Text(frm_L,width=40,height=30,font=("楷体"),yscrollcommand =

scrollbar_LV.set,bg='SkyBlue')

scrollbar_LV.config(command=text1.yview)

text1.pack()

Menu

The Menu widget is used to implement toplevel, pulldown, and popup menus.

This widget is used to display all kinds of menus used by an application.
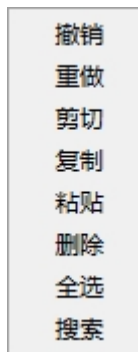
toplevel and pulldown menus
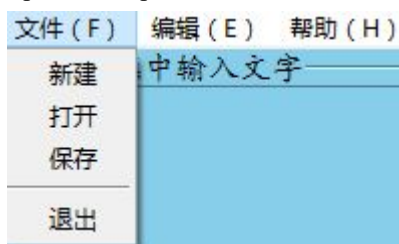
menubar = Menu(root)

filemenu = Menu(menubar, tearoff=False)

filemenu.add_command(label="新建", command=new1)

```python
filemenu.add_command(label="打开", command=open_file)
filemenu.add_command(label="保存", command=save2)

filemenu.add_separator()

filemenu.add_command(label="退出", command=Exit)
menubar.add_cascade(label="文件（F）", menu=filemenu)
```
popup menus


```python
# 创建一个弹出菜单

menu = Menu(text1,tearoff=False)
menu.add_command(label="撤销", command=undo)
menu.add_command(label="重做", command=redo)
menu.add_command(label="剪切", command=Cut)
menu.add_command(label="复制", command=Copy)
menu.add_command(label="粘贴", command=Paste)
menu.add_command(label="删除", command=Delete)
menu.add_command(label="全选", command=Select_All)
menu.add_command(label="搜索", command=findwordsmean)
def popup(event):
    menu.post(event.x_root, event.y_root)

# 绑定鼠标右键
text1.bind("<Button-3>", popup)
```

Button

The Button widget is a standard Tkinter widget used to implement various kinds of buttons. Buttons can contain text or images, and you can associate a Python function or method with each button. When the button is pressed, Tkinter automatically calls that function or method.


```python
Button(background,text='分词',image=image2,font=("楷体", 20),compound=CENTER,command =
fenci,bd=0).pack(side=TOP)
```

Entry

The entry widget is used to enter text strings. This widget allows the user to enter one line of text, in a single font.

Entry(background, textvariable = var2).pack(side=BOTTOM)

Filedialog

The tkFileDialog module can be used to get a filename from the user. The module provides two convenience functions, one to get an existing filename so you can open it, and one to get a new filename, to save things into.
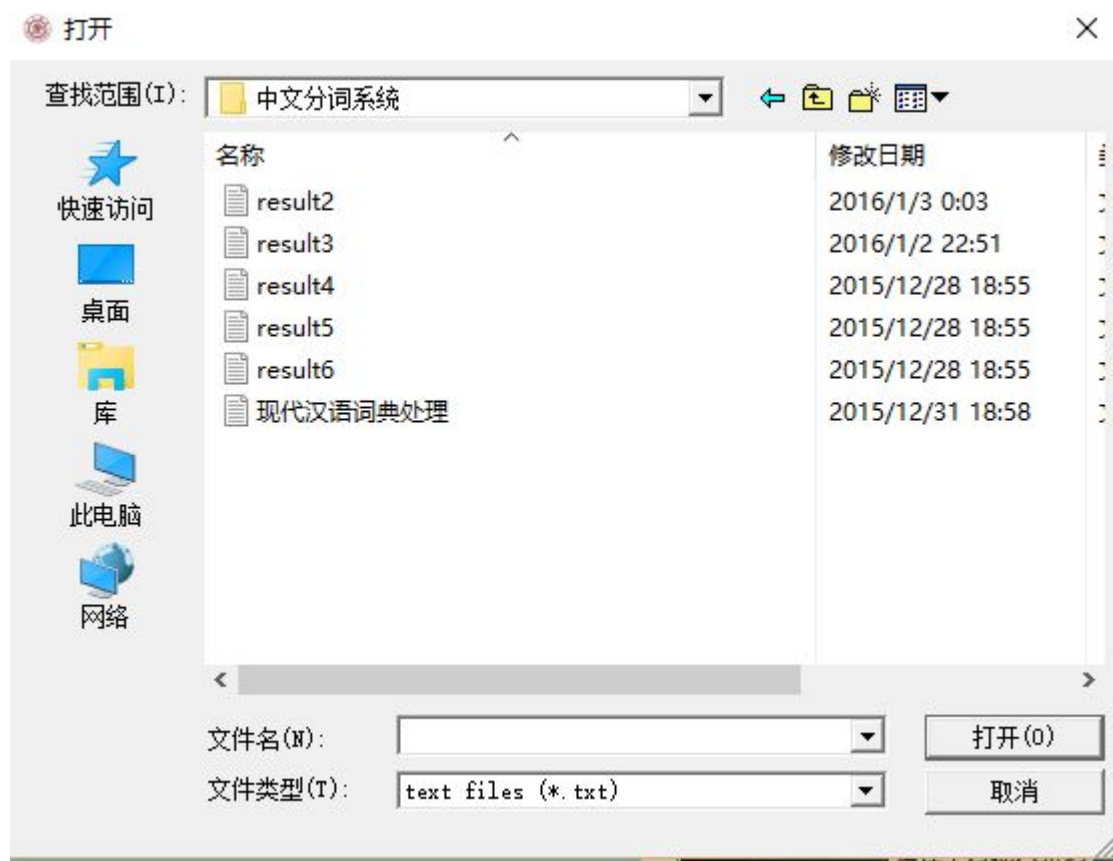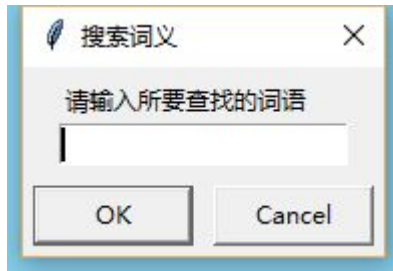

filename=askopenfilename(filetypes=[('text files', '.txt'),('all files', '.*')])
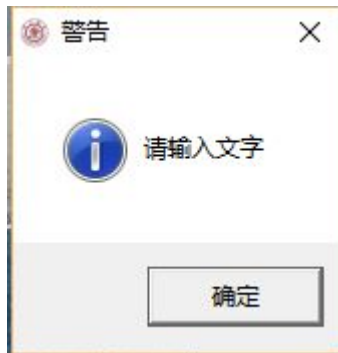
Messagebox

simpledialog

MessageBox shows a modal dialog box, which contains a system icon, a set of buttons and a brief specific to the application message, such as state or incorrect information.

file = askstring(title = '搜索词义',prompt = '请输入所要查找的词语')

messagebox

The MessageBox module provides an interface to the message dialogs.

showwarning('警告','请输入文字')

3.2 Algorithm Description

1.We choose to deal with the segmentation by dealing with the cleavage sites.

2. besides using the largest length of words, We combine the RMM and MM to find the different sites of segmentation. And then by comparing the frequency of the different words, I can choose the better segmentation and improve the accuracy.

3. the program can deal with the segmentation of sequences.

4.After that I find we should breadth our functions. I devices to use "baidu" .
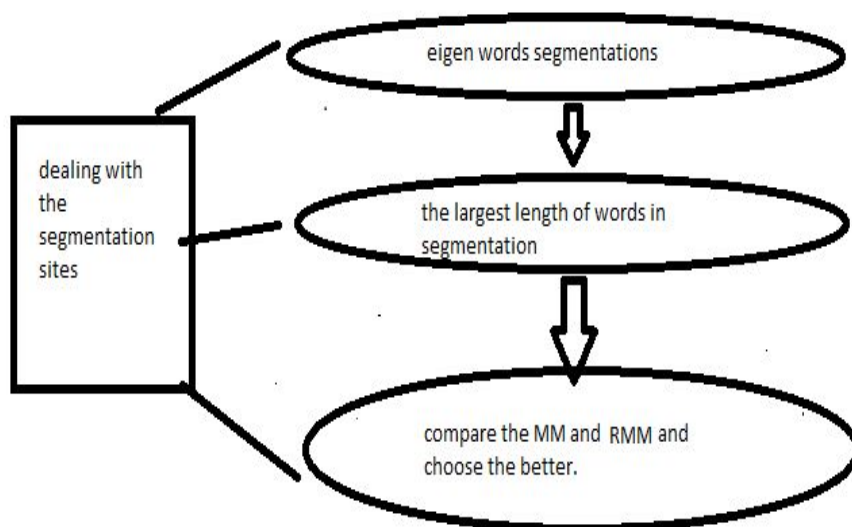
3.3 Lexicon:
3.3.1 Main lexicon:

The lexicon is divided into 5 parts, which contain Chinese vocabulary with different length. This can get less calculation in searching the words. The key of this lexicon is Chinese characters, the value is the frequency of the vocabulary, for our algorithm is based on comparing the frequency of vocabularies to determined the proper dividing.

```
                          ┌──────────┐
                          │   File   │
                          └──────────┘
                                │ read
                                │
                      ◇ classfy by the length ◇

            write in 5 txts │ length=?
    ┌────────┬──────────┬──────────┬──────────┬──────────┐
    │ 2      │ 3        │ 4        │ 5        │ >5       │
 ╭────────╮ ╭────────╮ ╭────────╮ ╭────────╮ ╭────────╮
 │ result2│ │ result3│ │ result4│ │ result5│ │ result6│
 ╰────────╯ ╰────────╯ ╰────────╯ ╰────────╯ ╰────────╯
    └──────────────────────── read ────────────────────┘
    ┌────────┬──────────┬──────────┬──────────┬──────────┐
 ┌──────┐  ┌──────┐  ┌──────┐  ┌──────┐  ┌──────┐
 │ dic2 │  │ dic3 │  │ dic4 │  │ dic5 │  │ dic6 │
 └──────┘  └──────┘  └──────┘  └──────┘  └──────┘
```
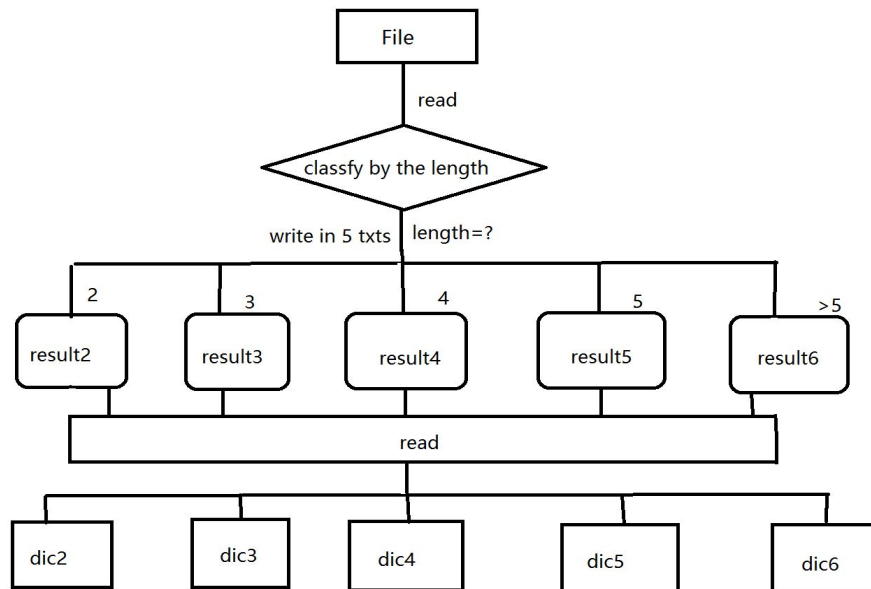
3.3.2 Lexicon of the function--explain the meaning of a word:
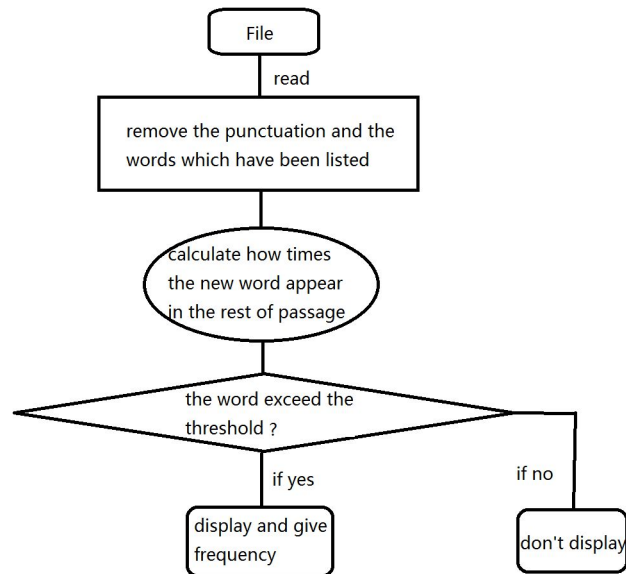
This lexicon is modified by the "Modern Chinese Dictionary". The key of it is the Chinese characters, the value is the meaning of it.

3.4 Special functions:

3.4.1 Find a new word of a text:

We can find words which are not collected in lexicon. Then we give the proper frequency of the new words.

Function: Find the unlisted words in a passage (And give a frequency of it)

```
                        ┌──────────┐
                        │   File   │
                        └──────────┘
                             │ read
                   ┌─────────────────────────┐
                   │ remove the punctuation and the │
                   │ words which have been listed   │
                   └─────────────────────────┘
                             │
                        ╭──────────────╮
                        │ calculate how times │
                        │ the new word appear │
                        │ in the rest of passage │
                        ╰──────────────╯
                             │
                    ◇ the word exceed the threshold ? ◇
                    │ if yes              │ if no
              ┌──────────────┐      ┌──────────────┐
              │ display and give │      │ don't display │
              │ frequency       │      └──────────────┘
              └──────────────┘
```

Stress:

    [1] threshold: The proper threshold is important for it can determined which words
        is what we need. According to the length of the passage, we have
        different threshold.

$$\text{threshold} = \frac{\text{The sum of numbers of the appearance of the 8 words which appear the most of times} + \text{the sum of numbers of the appearance of the 8 words which appear the least of times}}{16}$$

    [2] The calculation of the frequency:

        Because of the difference of the corpora, the times of a new word a
        new word appears can't be considered as a frequency, so we deal
        with it by proportion.We calculate all the listed word in the passage's
        proper proportion to get a average number. So we can give a
        reasonable frequency instead of giving a random number.

$$\text{Frequency} = \text{the times of new word appears} \times \text{times}$$

$$\text{Times} = \text{Average}\left(\frac{\text{listed word's frequency of our dictionary}}{\text{the number of the same word appears in the passage}}\right)$$

3.4.2 Explain the meaning of the word:

When users input a word, we can find its means from Modern Chinese Dictionary. If it isn't in the Modern Chinese Dictionary, we will search it in internet(www.baidu.com).
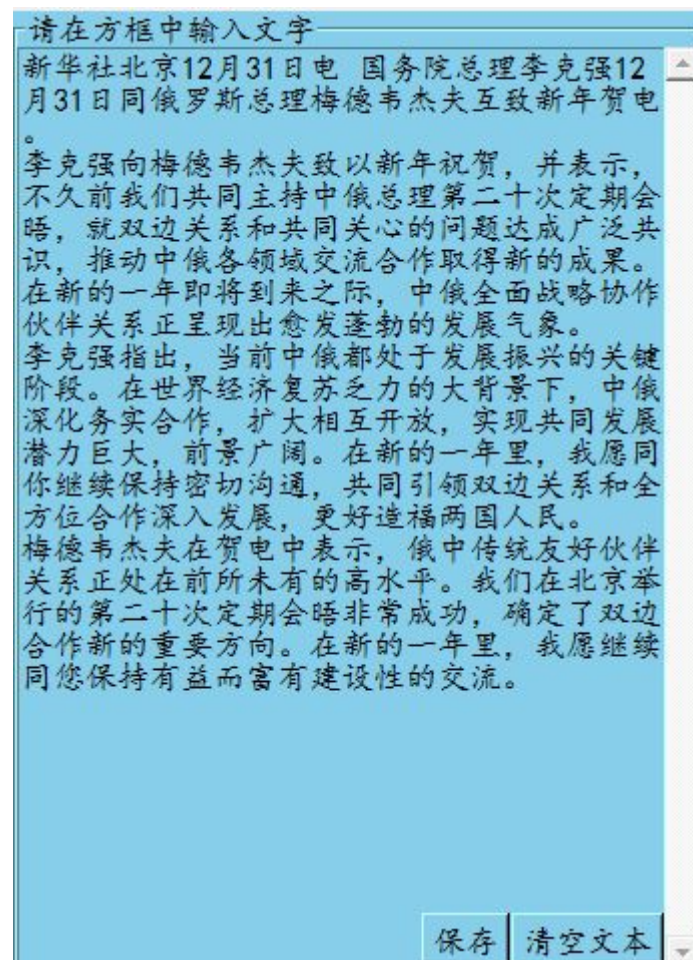
Function: Explain the meaning of the word
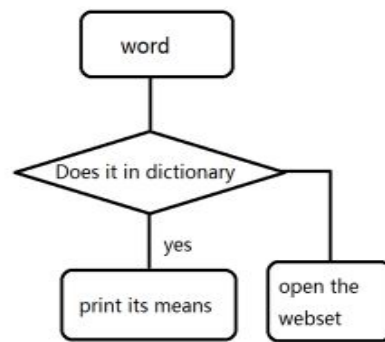
4.1 Screenshots and 4.2 Testing Procedure, Data and Result operation:

1.Input the text

open a file or input directly .

**2. Word or sentence segmentation**

分词

分句

3.Show result.

分词结果

新华社|北京|12|月|31|日电|　|国务院|总理|李克强|12|月|31|日|同|俄罗斯|总理|梅德韦杰夫|互|致|新年|贺电|。|
李克强	向	梅德韦杰夫	致以	新年	祝贺	，	并	表示	，	不久前	我们	共同	主持	中	俄	总理	第二十	次	定期	会晤	，	就	双边关系	和	共同	关心	的	问题	达成	广泛	共识	，	推动	中	俄	各	领域	交流合作	取得	新	的	成果	。	在	新	的	一	年	即将	到来之际	，	中	俄	全面	战略	协作	伙伴	关系	正	呈现	出	愈发	蓬勃	的	发展	气象	。
李克强	指出	，	当前	中	俄	都	处于	发展	振兴	的	关键	阶段	。	在	世界经济	复苏	乏力	的	大	背景	下	，	中	俄	深化	务实	合作	，	扩大	相互	开放	，	实现	共同发展	潜力	巨大	，	前景	广阔	。	在	新	的	一	年	里	，	我	愿	同	你	继续	保持	密切	沟通	，	共同	引领	双边关系	和	全方位	合作	深入	发展	，	更好	造福	两	国	人民	。
梅德韦杰夫	在	贺电	中	表示	，	俄	中	传统	友好	伙伴	关系	正	处在	前所未有	的	高	水平	。	我们	在	北京	举行	的	第二十	次	定期	会晤	非常	成功	，	确定	了	双边合作	新	的	重要	方向	。	在	新	的	一	年	里	，	我	愿	继续	同	您	保持	有益	而	富有	建设性	的	交流	。

保存　清空文本

4.save
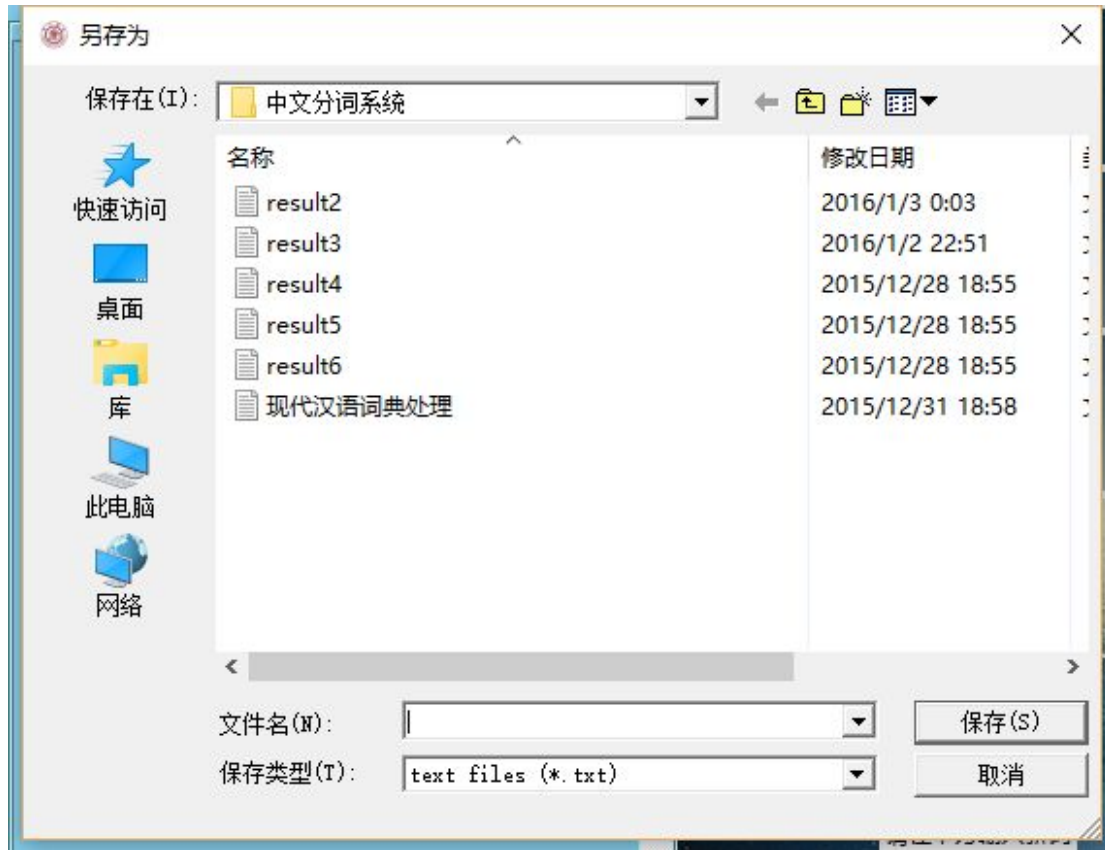
High level function:
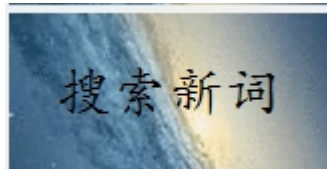**1.Used for developer to search for new words.**

请在方框中输入文字　　　保存　清空文本

秋天的雨

雨，像银灰色黏湿的蛛丝，织成一片轻柔的网，网住了整个秋的世界．天也是暗沉沉的，像古老的住宅里缠满着蛛丝网的屋顶．那堆在天上的灰白色的云片，就像屋顶上剥落的白粉．在这古旧的屋顶的笼罩下，一切都是异常的沉闷．园子里绿羃羃的古榴桑树、葡萄藤．都不过代表着过去盛夏的繁荣，现在已成了古罗马建筑的遗迹一样，在萧萧的雨声中瑟缩不宁，回忆着光荣的过去．草色已经转入忧郁的苍黄，地下找不出一点新鲜的花朵；宿舍墙外一带种的娇嫩的洋水仙，垂了头，含着满眼的泪珠，在那里叹息它们的薄命，才过了两天的晴美的好日子又遇到这样喜气薰蒸的雨天．只有墙角的桂花，枝头已经缀着几个黄金一样宝贵的嫩蕊，小心地隐藏在绿油油椭圆形的叶瓣下，透露出一点新生命萌芽的希望．

雨静悄悄地下着，只有一点细细的渐沥沥的声音．桔红色的房屋，像披着袈裟鲜艳的老僧，垂头合目，受着雨底的洗礼．那潮湿的红砖，发出有刺激的猪血的颜色和墙下绿油油的桂叶成为强烈的对照．灰色的癞蛤蟆，在湿烂发霉的泥地里跳跃着；在秋雨的沉闷的网底，只有它是唯一的充满愉快的生气的东西．它背上灰黄斑驳的花纹，跟沉闷的天空遥遥相应，造成和谐的色调．它噗秃噗秃地跳着，从草窠里，跳到泥里，测出深绿的水花．

雨，像银灰色黏濡的蛛丝，织成一片轻柔的网，网住了整个秋的世界

搜索新词

Text is too small or there are few new words appearing, so it is not sufficient to identify new words.
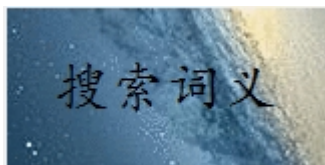
分词结果　　　　　　保存　清空文本

文本太小或新词出现频率太少，不足以认定有新词

分词结果　保存　清空文本

我了	205
一种	176
听雨	205
在的	323
一片	176
的下	117
它的	205
的的	411
了的	176
着的	323
的一	117
有一	117
很雪	176
种的	117
你的	763
的有	117
出的	146
是的	264
会有	117
一夜	117
我的	235
有的	117

2.

搜索词义

搜索词义

请输入所要查找的词语

OK　Cancel

or

分词结果　保存　清空文本

我了	205
一种	176
听雨	205
在的	323
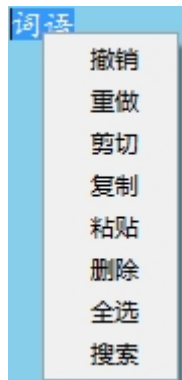
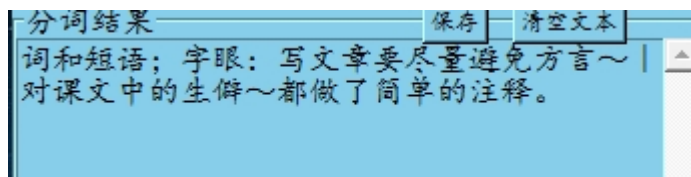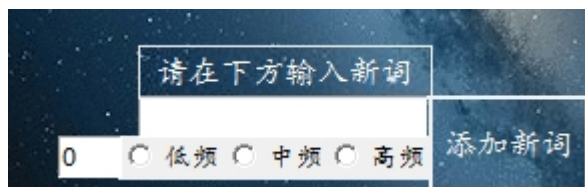**result:**

3.add new word

5.Conclusion

we can nearly solve this problem. there are some places where we can make a progress.

We learn to cooperate with others and we enjoy working together with each other and communicating with each other. Although we meet many difficult problems, we work together and solve almost of them. I believe we learn much dignities from this experience.

We can make great progress in learning python. We get inspiration and   innovation from this process, which , we believe, is the most important.