Project3　实验报告

学号：22336289　　　姓名：袁鹏湘

实验内容

设计一个程序以分析文本，输出每个词的出现次数，以及出现位置所在的行号。具体包括：以txt格式存储的英文短文，统计词频值Top-K的单词；C/C++语言编写的源代码，以保留字符集作为待统计的词汇集；常见的中文文本，自定义关键词统计。

功能说明

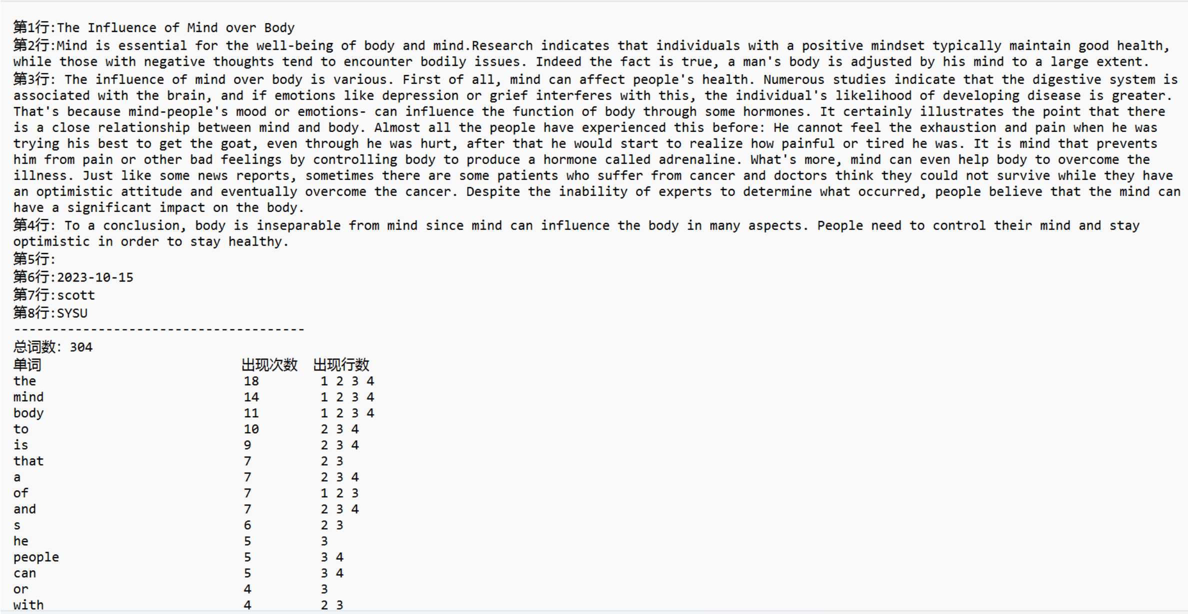
根据界面的指示输入操作代码，选择合适的分析方式，并输入正确的源文件名字（不在当前目录下时，需要使用绝对路径或相对路径指明）， 有自定义关键词的，也需要输入正确的关键词文件名字， 即可分析文本（文本名字可以自定义，实验报告中为了方便统一用test和key起名，实际上可以用任意名字）。结果会保存在当前目录下的Running Log.txt文件中。如果含中文字符，请确保文件格式以ANSI保存，用UTF-8的编码方式会导致不兼容GBK而乱码！

示例

- 示例1　(英文短文 test1.txt)

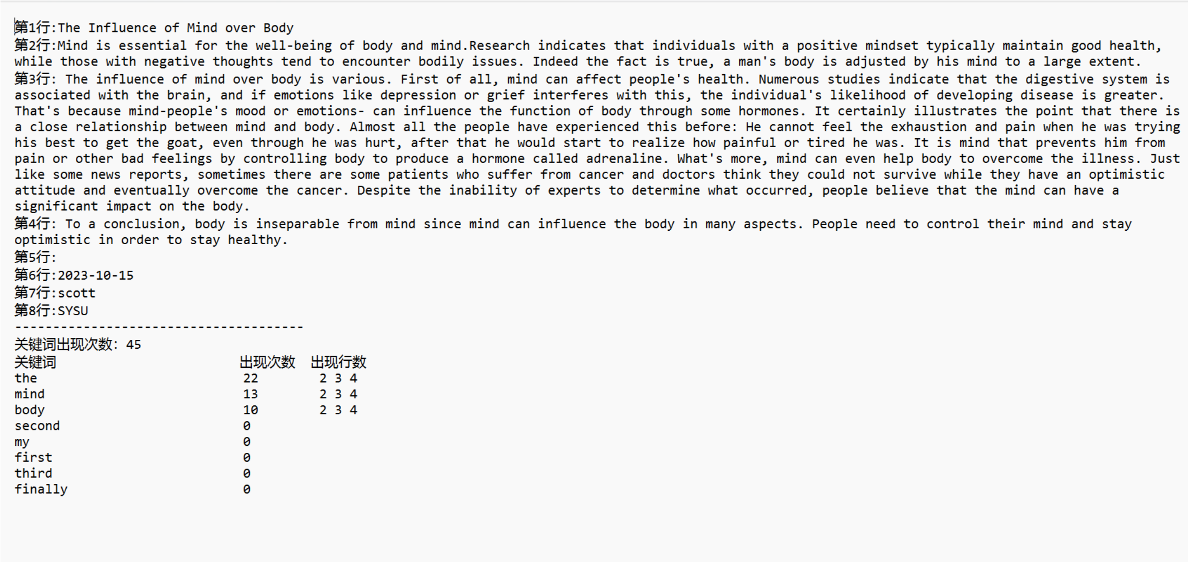
> The Influence of Mind over Body
> Mind is essential for the well-being of body and mind. Research indicates that individuals with a positive mindset typically maintain good health, while those with negative thoughts tend to encounter bodily issues. Indeed the fact is true, a man's body is adjusted by his mind to a large extent.
> The influence of mind over body is various. First of all, mind can affect people's health. Numerous studies indicate that the digestive system is associated with the brain, and if emotions like depression or grief interferes with this, the individual's likelihood of developing disease is greater. That's because mind-people's mood or emotions- can influence the function of body through some hormones. It certainly illustrates the point that there is a close relationship between mind and body. Almost all the people have experienced this before: He cannot feel the exhaustion and pain when he was trying his best to get the goat, even through he was hurt, after that he would start to realize how painful or tired he was. It is mind that prevents him from pain or other bad feelings by controlling body to produce a hormone called adrenaline. What's more, mind can even help body to overcome the illness. Just like some news reports, sometimes there are some patients who suffer from cancer and doctors think they could not survive while they have an optimistic attitude and eventually overcome the cancer. Despite the inability of experts to determine what occurred, people believe that the mind can have a significant impact on the body.
> To a conclusion, body is inseparable from mind since mind can influence the body in many aspects. People need to control their mind and stay optimistic in order to stay healthy.
>
> 2023-10-15
> scott
> SYSU

1、使用默认分析的结果：



```
第1行:The Influence of Mind over Body
第2行:Mind is essential for the well-being of body and mind.Research indicates that individuals with a positive mindset typically maintain good health,
while those with negative thoughts tend to encounter bodily issues. Indeed the fact is true, a man's body is adjusted by his mind to a large extent.
第3行: The influence of mind over body is various. First of all, mind can affect people's health. Numerous studies indicate that the digestive system is
associated with the brain, and if emotions like depression or grief interferes with this, the individual's likelihood of developing disease is greater.
That's because mind-people's mood or emotions- can influence the function of body through some hormones. It certainly illustrates the point that there
is a close relationship between mind and body. Almost all the people have experienced this before: He cannot feel the exhaustion and pain when he was
trying his best to get the goat, even through he was hurt, after that he would start to realize how painful or tired he was. It is mind that prevents
him from pain or other bad feelings by controlling body to produce a hormone called adrenaline. What's more, mind can even help body to overcome the
illness. Just like some news reports, sometimes there are some patients who suffer from cancer and doctors think they could not survive while they have
an optimistic attitude and eventually overcome the cancer. Despite the inability of experts to determine what occurred, people believe that the mind can
have a significant impact on the body.
第4行: To a conclusion, body is inseparable from mind since mind can influence the body in many aspects. People need to control their mind and stay
optimistic in order to stay healthy.
第5行:
第6行:2023-10-15
第7行:scott
第8行:SYSU
---------------------------------------
总词数: 304
单词                          出现次数  出现行数
the                           18         1 2 3 4
mind                          14         1 2 3 4
body                          11         1 2 3 4
to                            10         2 3 4
is                            9          2 3 4
that                          7          2 3
a                             7          2 3 4
of                            7          1 2 3
and                           7          2 3 4
s                             6          2 3
he                            5          3
people                        5          3 4
can                           5          3 4
or                            4          3
with                          4          2 3
```

2、使用自定义关键词文件（key1.txt）的结果：
(PS: 自定义关键词会严格按照关键词**区分大小写**来精确地分析，默认分析则不会)



```
第1行:The Influence of Mind over Body
第2行:Mind is essential for the well-being of body and mind.Research indicates that individuals with a positive mindset typically maintain good health,
while those with negative thoughts tend to encounter bodily issues. Indeed the fact is true, a man's body is adjusted by his mind to a large extent.
第3行: The influence of mind over body is various. First of all, mind can affect people's health. Numerous studies indicate that the digestive system is
associated with the brain, and if emotions like depression or grief interferes with this, the individual's likelihood of developing disease is greater.
That's because mind-people's mood or emotions- can influence the function of body through some hormones. It certainly illustrates the point that there is
a close relationship between mind and body. Almost all the people have experienced this before: He cannot feel the exhaustion and pain when he was trying
his best to get the goat, even through he was hurt, after that he would start to realize how painful or tired he was. It is mind that prevents him from
pain or other bad feelings by controlling body to produce a hormone called adrenaline. What's more, mind can even help body to overcome the illness. Just
like some news reports, sometimes there are some patients who suffer from cancer and doctors think they could not survive while they have an optimistic
attitude and eventually overcome the cancer. Despite the inability of experts to determine what occurred, people believe that the mind can have a
significant impact on the body.
第4行: To a conclusion, body is inseparable from mind since mind can influence the body in many aspects. People need to control their mind and stay
optimistic in order to stay healthy.
第5行:
第6行:2023-10-15
第7行:scott
第8行:SYSU
---------------------------------------
关键词出现次数: 45
关键词                        出现次数  出现行数
the                           22         2 3 4
mind                          13         2 3 4
body                          10         2 3 4
second                        0
my                            0
first                         0
third                         0
finally                       0
```

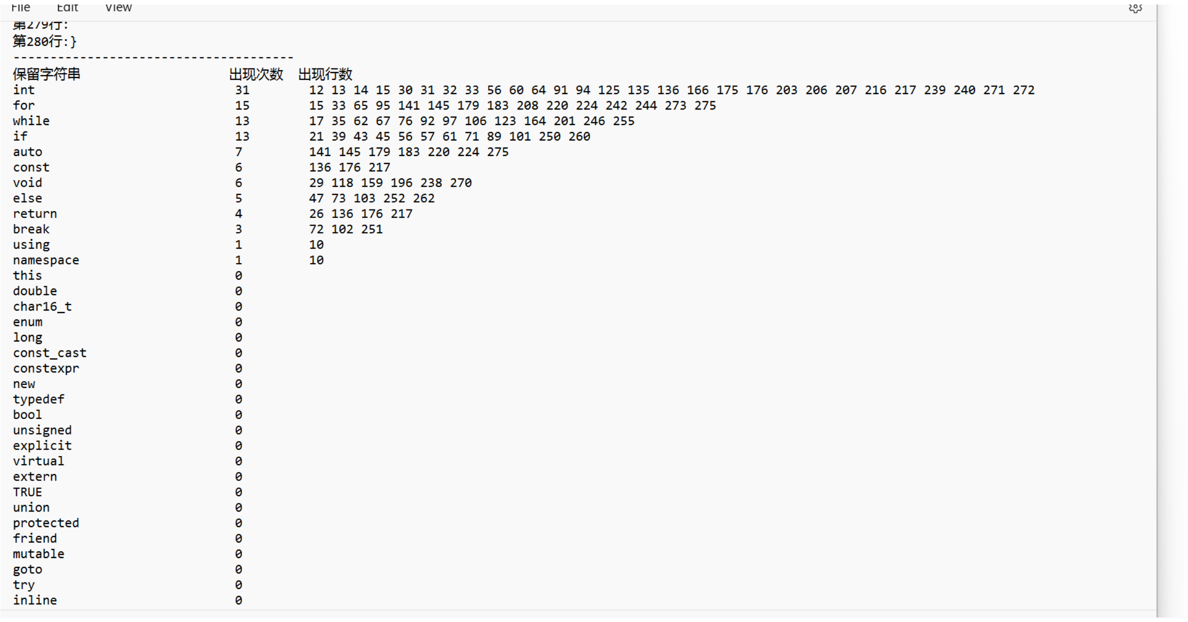
- 示例2 （代码片段 test2.txt截取部分）

```
#include "analy.h"
#include
#include
#include
#include <unordered_map>
#include
#include
#include
#include
using namespace std;

vector Text::get_next(string& t){
 int len=t.size();
 vector next(len,0);
 for(int i=1,j=0;i<len;i++)
   {
    while(j>0&&t[i]!=t[j])
```

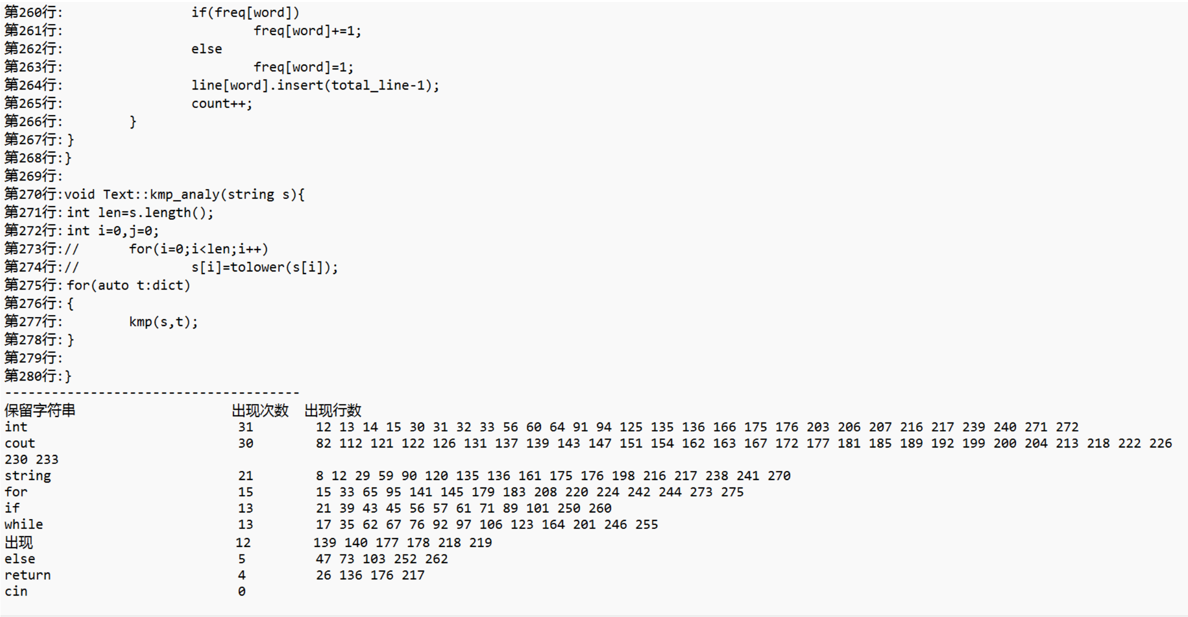
```
          {
           j=next[j-1];
        }
        if(t[i]==t[j]) {
           j++;
        }
        next[i]=j;
    }
    return next;
  }
  ……    #这里展示的并不完整！
```

1、使用默认分析的结果：



```
第279行:
第280行:}
--------------------------------------
保留字符串                     出现次数  出现行数
int                             31        12 13 14 15 30 31 32 33 56 60 64 91 94 125 135 136 166 175 176 203 206 207 216 217 239 240 271 272
for                             15        15 33 65 95 141 145 179 183 208 220 224 242 244 273 275
while                           13        17 35 62 67 76 92 97 106 123 164 201 246 255
if                              13        21 39 43 45 56 57 61 71 89 101 250 260
auto                            7         141 145 179 183 220 224 275
const                           6         136 176 217
void                            6         29 118 159 196 238 270
else                            5         47 73 103 252 262
return                          4         26 136 176 217
break                           3         72 102 251
using                           1         10
namespace                       1         10
this                            0
double                          0
char16_t                        0
enum                            0
long                            0
const_cast                      0
constexpr                       0
new                             0
typedef                         0
bool                            0
unsigned                        0
explicit                        0
virtual                         0
extern                          0
TRUE                            0
union                           0
protected                       0
friend                          0
mutable                         0
goto                            0
try                             0
inline                          0
```

2、使用自定义关键词文件（key2.txt）的结果：

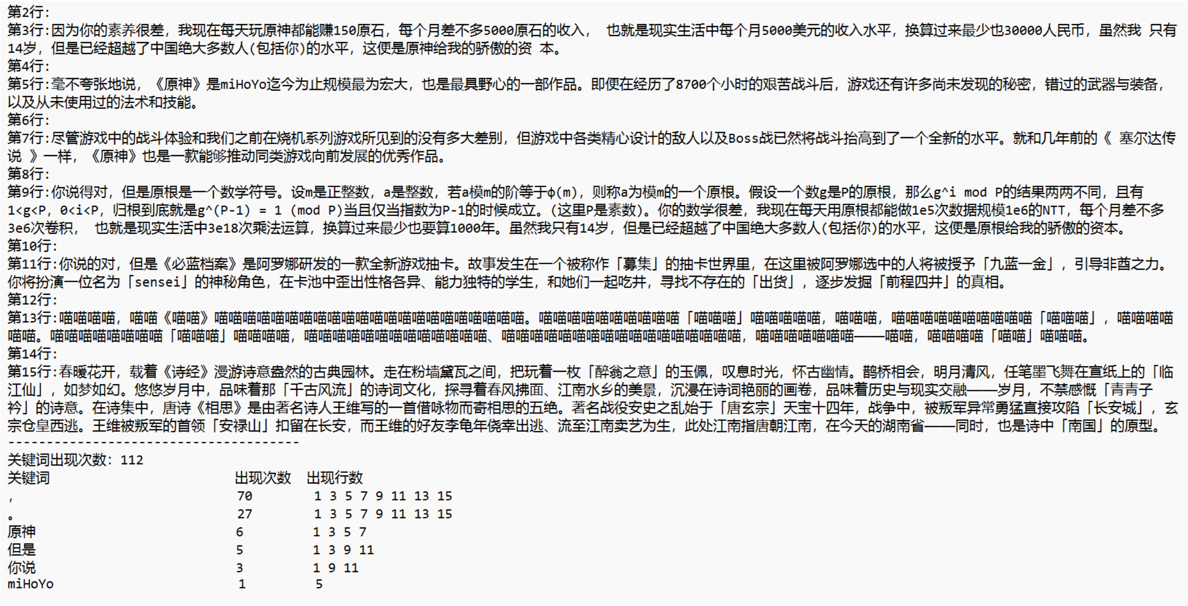


```
第260行:                if(freq[word])
第261行:                        freq[word]+=1;
第262行:                else
第263行:                        freq[word]=1;
第264行:                line[word].insert(total_line-1);
第265行:                count++;
第266行:        }
第267行:}
第268行:}
第269行:
第270行:void Text::kmp_analy(string s){
第271行:int len=s.length();
第272行:int i=0,j=0;
第273行://      for(i=0;i<len;i++)
第274行://              s[i]=tolower(s[i]);
第275行:for(auto t:dict)
第276行:{
第277行:        kmp(s,t);
第278行:}
第279行:
第280行:}
--------------------------------------
保留字符串                     出现次数  出现行数
int                             31        12 13 14 15 30 31 32 33 56 60 64 91 94 125 135 136 166 175 176 203 206 207 216 217 239 240 271 272
cout                            30        82 112 121 122 126 131 137 139 143 147 151 154 162 163 167 172 177 181 185 189 192 199 200 204 213 218 222 226
230 233
string                          21        8 12 29 59 90 120 135 136 161 175 176 198 216 217 238 241 270
for                             15        15 33 65 95 141 145 179 183 208 220 224 242 244 273 275
if                              13        21 39 43 45 56 57 61 71 89 101 250 260
while                           13        17 35 62 67 76 92 97 106 123 164 201 246 255
出现                            12        139 140 177 178 218 219
else                            5         47 73 103 252 262
return                          4         26 136 176 217
cin                             0
```

- 示例3 （中文文本 test3.txt）

PS:由于文本量过大，不展示输入了

使用自定义关键词文件（key3.txt）的结果：



```
第2行:
第3行:因为你的素养很差，我现在每天玩原神都能赚150原石，每个月差不多5000原石的收入， 也就是现实生活中每个月5000美元的收入水平，换算过来最少也30000人民币，虽然我 只有
14岁，但是已经超越了中国绝大多数人(包括你)的水平，这便是原神给我的骄傲的资 本。
第4行:
第5行:毫不夸张地说，《原神》是miHoYo迄今为止规模最为宏大，也是最具野心的一部作品。即便在经历了8700个小时的艰苦战斗后，游戏还有许多尚未发现的秘密，错过的武器与装备，
以及从未使用过的法术和技能。
第6行:
第7行:尽管游戏中的战斗体验和我们之前在烧机系列游戏所见到的没有多大差别，但游戏中各类精心设计的敌人以及Boss战已然将战斗抬高到了一个全新的水平。就和几年前的《 塞尔达传
说 》一样，《原神》也是一款能够推动同类游戏向前发展的优秀作品。
第8行:
第9行:你说得对，但是原根是一个数学符号。设m是正整数，a是整数，若a模m的阶等于φ(m)，则称a为模m的一个原根。假设一个数g是P的原根，那么g^i mod P的结果两两不同，且有
1<g<P，0<i<P，归根到底就是g^(P-1) = 1 (mod P)当且仅当指数为P-1的时候成立。(这里P是素数)。你的数学很差，我现在每天用原根都能做1e5次数据规模1e6的NTT，每个月差不多
3e6次卷积， 也就是现实生活中3e18次乘法运算，换算过来最少也要算1000年。虽然我只有14岁，但是已经超越了中国绝大多数人(包括你)的水平，这便是原根给我的骄傲的资本。
第10行:
第11行:你说的对，但是《必蓝档案》是阿罗娜研发的一款全新游戏抽卡。故事发生在一个被称作「募集」的抽卡世界里，在这里被阿罗娜选中的人将被授予「九蓝一金」，引导非酋之力。
你将扮演一位名为「sensei」的神秘角色，在卡池中歪出性格各异、能力独特的学生，和她们一起吃井，寻找不存在的「出货」，逐步发掘「前程四井」的真相。
第12行:
第13行:喵喵喵喵，喵喵《喵喵》喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵。喵喵喵喵喵喵喵喵喵「喵喵喵」喵喵喵喵喵，喵喵喵，喵喵喵喵喵喵喵喵喵喵「喵喵喵」，喵喵喵喵
喵喵。喵喵喵喵喵喵喵喵「喵喵喵」喵喵喵喵，喵喵喵喵喵喵喵喵喵喵喵喵喵喵、喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵喵，喵喵喵喵喵喵喵——喵喵，喵喵喵喵「喵喵」喵喵喵。
第14行:
第15行:春暖花开，载着《诗经》漫游诗意盎然的古典园林。走在粉墙黛瓦之间，把玩着一枚「醉翁之意」的玉佩，叹息时光，怀古幽情。鹊桥相会，明月清风，任笔墨飞舞在宣纸上的「临
江仙」，如梦如幻。悠悠岁月中，品味着那「千古风流」的诗词文化，探寻着春风拂面、江南水乡的美景，沉浸在诗词艳丽的画卷，品味着历史与现实交融——岁月，不禁感慨「青青子
衿」的诗意。在诗集中，唐诗《相思》是由著名诗人王维写的一首借咏物而寄相思的五绝。著名战役安史之乱始于「唐玄宗」天宝十四年，战争中，被叛军异常勇猛直接攻陷「长安城」，玄
宗仓皇西逃。王维被叛军的首领「安禄山」扣留在长安，而王维的好友李龟年侥幸出逃、流至江南卖艺为生，此处江南指唐朝江南，在今天的湖南省——同时，也是诗中「南国」的原型。
-------------------------------------
关键词出现次数: 112
关键词                          出现次数  出现行数
，                               70        1 3 5 7 9 11 13 15
。                               27        1 3 5 7 9 11 13 15
原神                             6         1 3 5 7
但是                             5         1 3 9 11
你说                             3         1 9 11
miHoYo                           1         5
```

关键代码展示

- **analy.h头文件展示**

```cpp
#ifndef _ANALY_H_
#define _ANALY_H_
#include <fstream>
#include <unordered_map>
#include <set>
#include <vector>
#include <algorithm>
#include <string>
#include <cctype>
using namespace std;

class Text{
    public:
        Text(ifstream& inFile,int choose);//无自定义关键词
        Text(ifstream& inFile,ifstream& keyFile);//自定义关键词文件

        void read1();//英语短文单词
        void read2();//c语言程序保留字符串
        void read3();//自定义关键词

        void analy(string s);//分析英语源文本
        void kmp_analy(string s);//KMP算法匹配

    private:
        set<string> dict;//分词表
        unordered_map<string,set<int>> line;//记录出现的行数
        unordered_map<string,int> freq;//记录出现的次数
        int total_line;//总行数
        int total_word;//总词数
        int key_count;//关键词数
        int count;//总词数
        ifstream* inFile;//暂存输入流
        vector<int> get_next(string& pattern);
        void kmp(string& s,string& t);
};

#endif //_ANALY_H_
```

- **analy.cpp展示**

```
#include "analy.h"
#include <fstream>
#include <iostream>
#include <iomanip>
#include <unordered_map>
#include <set>
#include <algorithm>
#include <string>
#include <cctype>
using namespace std;

/*kmp算法的next数组获取和kmp匹配分析*/
vector<int> Text::get_next(string& t){
    int len=t.size();
    vector<int> next(len,0);
    for(int i=1,j=0;i<len;i++)
    {
        while(j>0&&t[i]!=t[j])
        {
            j=next[j-1];
        }
        if(t[i]==t[j]) {
            j++;
        }
        next[i]=j;
    }
    return next;
}

void Text::kmp(string& s,string& t){
    vector<int> next=get_next(t);
    int n=s.size();
    int m=t.size();
    for(int i=0,j=0;i<n;i++)
    {
        while(j>0&&s[i]!=t[j])
        {
            j=next[j-1];
        }
        if(s[i]==t[j])
        {
            j++;
        }
        if(j==m)
        {
            if(!freq[t])
            freq[t]=1;
            else
            freq[t]+=1;
            key_count+=1;
            line[t].insert(total_line-1);
            j=next[j-1];
        }
    }
}
/*构造函数，分两个：一个用于默认分析，一个用于有自定义关键词文件的分析*/
Text::Text(ifstream& inFile,int
choose):inFile(&inFile),total_line(1),total_word(0),count(0),key_count(0){
    if(choose==1)
    {
        string s,word;
        int i=0,j=0;
        ifstream inFile("cpp_reserved.txt");
        while(getline(inFile,s))
        {
            int len=s.length();
            for(i=0;i<len;)
            {
                while(i<len&&(s[i]==' '||s[i]=='\n'))
                {
                    i++;
                }
                if(i==len)
                break;
                else
                {
                    j=i;
                    while(i<len&&s[i]!=' '&&s[i]!='\n')
                    {
                        i++;
                    }
                    word=s.substr(j,i-j);
                    dict.insert(word);
//                  cout<<word<<" ";
                }
            }
        }
    }
}

Text::Text(ifstream& inFile,ifstream&
keyFile):inFile(&inFile),total_line(1),total_word(0),count(0),key_count(0){
        string s,word;
        int i=0,j=0;
        while(getline(keyFile,s))
        {
            int len=s.length();
            for(i=0;i<len;)
            {
                while(i<len&&(s[i]==' '||s[i]=='\n'))
                {
                    i++;
                }
                if(i==len)
                break;
                else
                {
                    j=i;
                    while(i<len&&s[i]!=' '&&s[i]!='\n')
                    {
                        i++;
                    }
                    word=s.substr(j,i-j);
                    dict.insert(word);
//                  cout<<word<<" ";
                }
            }
        }
}
/*三个read（）函数其实大同小异，只是采用的分析方式不同，输出格式有不同*/
void Text::read1(){
    ofstream outFile("Running Log.txt");
    string s;//每次读取一行
    cout<<endl<<"成功读取以下内容："<<endl;
    cout<<"--------------------------------------"<<endl;
    if(dict.size())
        for(auto cur:dict)
            freq[cur]=0;
    while(getline(*inFile,s,'\n'))
    {
        int temp=total_line++;
        cout<<"第"<<temp<<"行:"<<s<<endl;
        outFile<<"第"<<temp<<"行:"<<s<<endl;
        analy(s);
    }
    total_line--;
    cout<<"--------------------------------------"<<endl;
    outFile<<"--------------------------------------"<<endl;

    /*排序输出Top-K的单词*/
    vector<pair<string,int>> temp(freq.begin(),freq.end());
    sort(temp.begin(),temp.end(),[](const pair<string,int>& a, const pair<string,int>& b){return a.second>b.second;});
    cout<<"总词数："<<count<<endl;
    outFile<<"总词数："<<count<<endl;
    cout<<left<<setw(30)<<"单词"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    outFile<<left<<setw(30)<<"单词"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    for(auto cur:temp)
    {
        cout<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        outFile<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        for(auto cur2:line[cur.first])
        {
            cout<<cur2<<" ";
            outFile<<cur2<<" ";
        }

        cout<<endl;
        outFile<<endl;
    }
    cout<<"--------------------------------------"<<endl;
    outFile.close();
```

```cpp
}


void Text::read2(){
    ofstream outFile("Running Log.txt");
    string s;//每次读取一行
    cout<<endl<<"成功读取以下内容："<<endl;
    cout<<"--------------------------------------"<<endl;
    if(dict.size())
        for(auto cur:dict)
            freq[cur]=0;
    while(getline(*inFile,s,'\n'))
    {
        int temp=total_line++;
        cout<<"第"<<temp<<"行:"<<s<<endl;
        outFile<<"第"<<temp<<"行:"<<s<<endl;
        kmp_analy(s);
    }
    total_line--;
    cout<<"--------------------------------------"<<endl;
    outFile<<"--------------------------------------"<<endl;

    vector<pair<string,int>> temp(freq.begin(),freq.end());
    sort(temp.begin(),temp.end(),[](const pair<string,int>& a, const pair<string,int>& b){return a.second>b.second;});
    cout<<left<<setw(30)<<"保留字符串"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    outFile<<left<<setw(30)<<"保留字符串"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    for(auto cur:temp)
    {
        cout<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        outFile<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        for(auto cur2:line[cur.first])
        {
            cout<<cur2<<" ";
            outFile<<cur2<<" ";
        }

        cout<<endl;
        outFile<<endl;
    }
    cout<<"--------------------------------------"<<endl;
    outFile.close();
}

void Text::read3(){
    ofstream outFile("Running Log.txt");
    string s;//每次读取一行
    cout<<endl<<"成功读取以下内容："<<endl;
    cout<<"--------------------------------------"<<endl;
    if(dict.size())
        for(auto cur:dict)
            freq[cur]=0;
    while(getline(*inFile,s,'\n'))
    {
        int temp=total_line++;
        cout<<"第"<<temp<<"行:"<<s<<endl;
        outFile<<"第"<<temp<<"行:"<<s<<endl;
        int len=s.length();
        int i=0;
        kmp_analy(s);
    }
    total_line--;
    cout<<"--------------------------------------"<<endl;
    outFile<<"--------------------------------------"<<endl;

    cout<<"关键词出现次数："<<key_count<<endl;
    outFile<<"关键词出现次数："<<key_count<<endl;

    vector<pair<string,int>> temp(freq.begin(),freq.end());
    sort(temp.begin(),temp.end(),[](const pair<string,int>& a, const pair<string,int>& b){return a.second>b.second;});
    cout<<left<<setw(30)<<"关键词"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    outFile<<left<<setw(30)<<"关键词"<<left<<setw(10)<<"出现次数"<<left<<setw(10)<<"出现行数"<<endl;
    for(auto cur:temp)
    {
        cout<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        outFile<<left<<setw(30)<<cur.first<<left<<setw(10)<<cur.second;
        for(auto cur2:line[cur.first])
        {
            cout<<cur2<<" ";
            outFile<<cur2<<" ";
        }

        cout<<endl;
        outFile<<endl;
    }
    cout<<"--------------------------------------"<<endl;
    outFile.close();
}

/*默认分析方法，一次读取一行后，对字符串按空格分词*/
void Text::analy(string s){
    int len=s.length();
    int i=0,j=0;
    string word;
    for(i=0;i<len;i++)
        s[i]=tolower(s[i]);
    for(i=0;i<len;)
    {
        while(i<len&&!isalnum(s[i]))
        {
            i++;
        }
        if(i==len)
        break;
        else
        {
            j=i;
            while(i<len&&isalnum(s[i]))
            {
                i++;
            }
            word=s.substr(j,i-j);
            if(freq[word])
                freq[word]+=1;
            else
                freq[word]=1;
            line[word].insert(total_line-1);
            count++;
        }
    }
}

/*采用了kmp算法的分析方法，按照kmp算法进行分词而不是空格*/
void Text::kmp_analy(string s){
    int len=s.length();
    int i=0,j=0;
//  for(i=0;i<len;i++)
//      s[i]=tolower(s[i]);
    for(auto t:dict)
    {
        kmp(s,t);
    }

}
```