2020 秋季《计算机科学与技术导论》期末大作业

二、比较两个 DOC 文档的相似性

（1）技术要求：

1. 两个文档均为.DOC 文件格式;

2. 统计这两个文档中有多少字符相同，有多少个字符不同，统计出前 10 个高频字或词;

（2）程序思想：

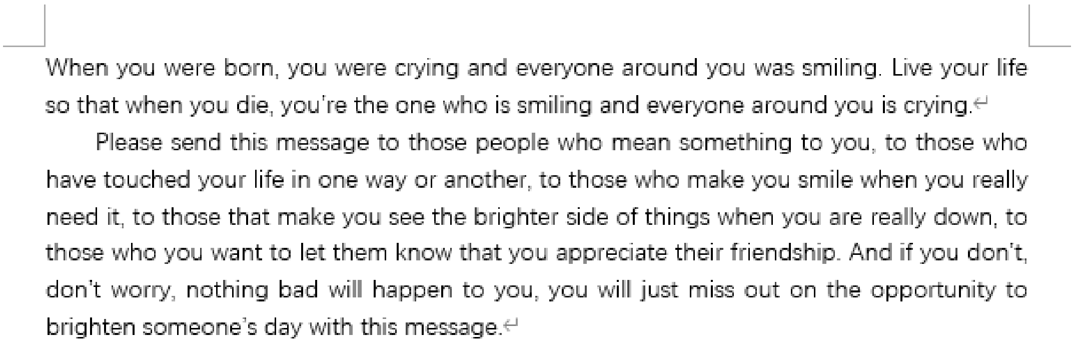
1. 打开 doc 文件：通过 poi 库读入 doc 文件，将 doc 文件中的文本存储到一个 String 中。

2. 统计单词频率：对文本遍历，只取出其中为字母的项，并将其全部转换为小写。若 map 中未存在该单词的键，则创建值为 1 的键值对，否则该单词的值加一。

3. 统计相同单词数与不同单词数：将两篇文章的单词和其个数存入 map1 与 map2 中。对其中的 map1 遍历，若单词在另一篇文章中出现则相同单词数量加一，用 map1 和 map2 的大小减去两倍相同单词数量即为不同单词的数量

(3) 输入输出数据:

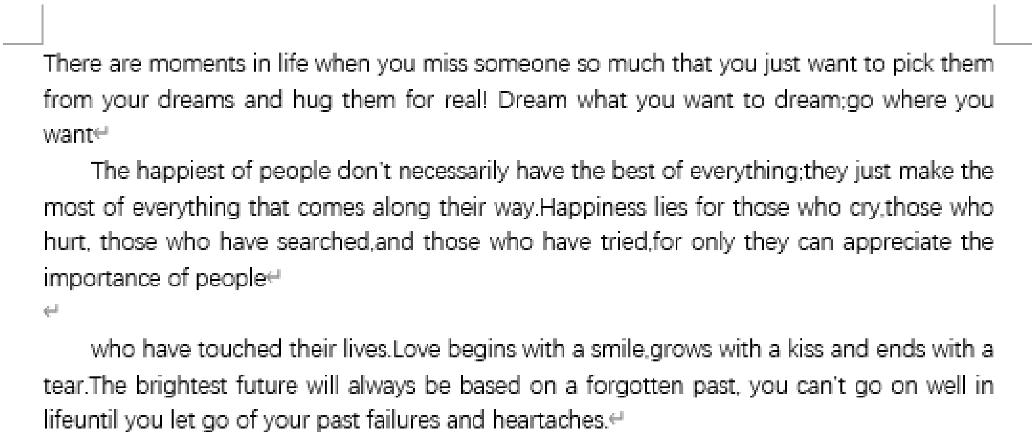
1. doc



When you were born, you were crying and everyone around you was smiling. Live your life so that when you die, you're the one who is smiling and everyone around you is crying.

Please send this message to those people who mean something to you, to those who have touched your life in one way or another, to those who make you smile when you really need it, to those that make you see the brighter side of things when you are really down, to those who you want to let them know that you appreciate their friendship. And if you don't, don't worry, nothing bad will happen to you, you will just miss out on the opportunity to brighten someone's day with this message.

2. doc

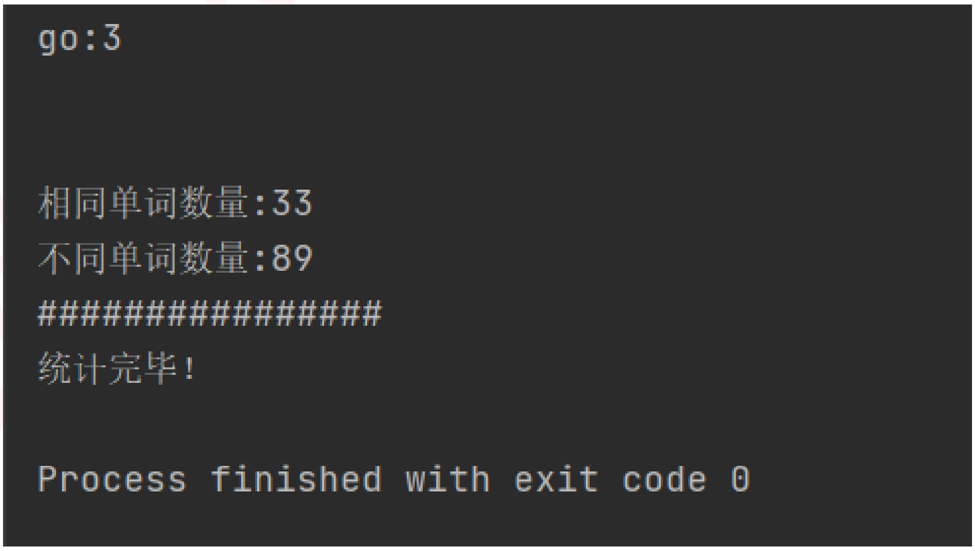


There are moments in life when you miss someone so much that you just want to pick them from your dreams and hug them for real! Dream what you want to dream;go where you want

The happiest of people don't necessarily have the best of everything;they just make the most of everything that comes along their way.Happiness lies for those who cry,those who hurt, those who have searched,and those who have tried,for only they can appreciate the importance of people

who have touched their lives.Love begins with a smile,grows with a kiss and ends with a tear.The brightest future will always be based on a forgotten past, you can't go on well in lifeuntil you let go of your past failures and heartaches.

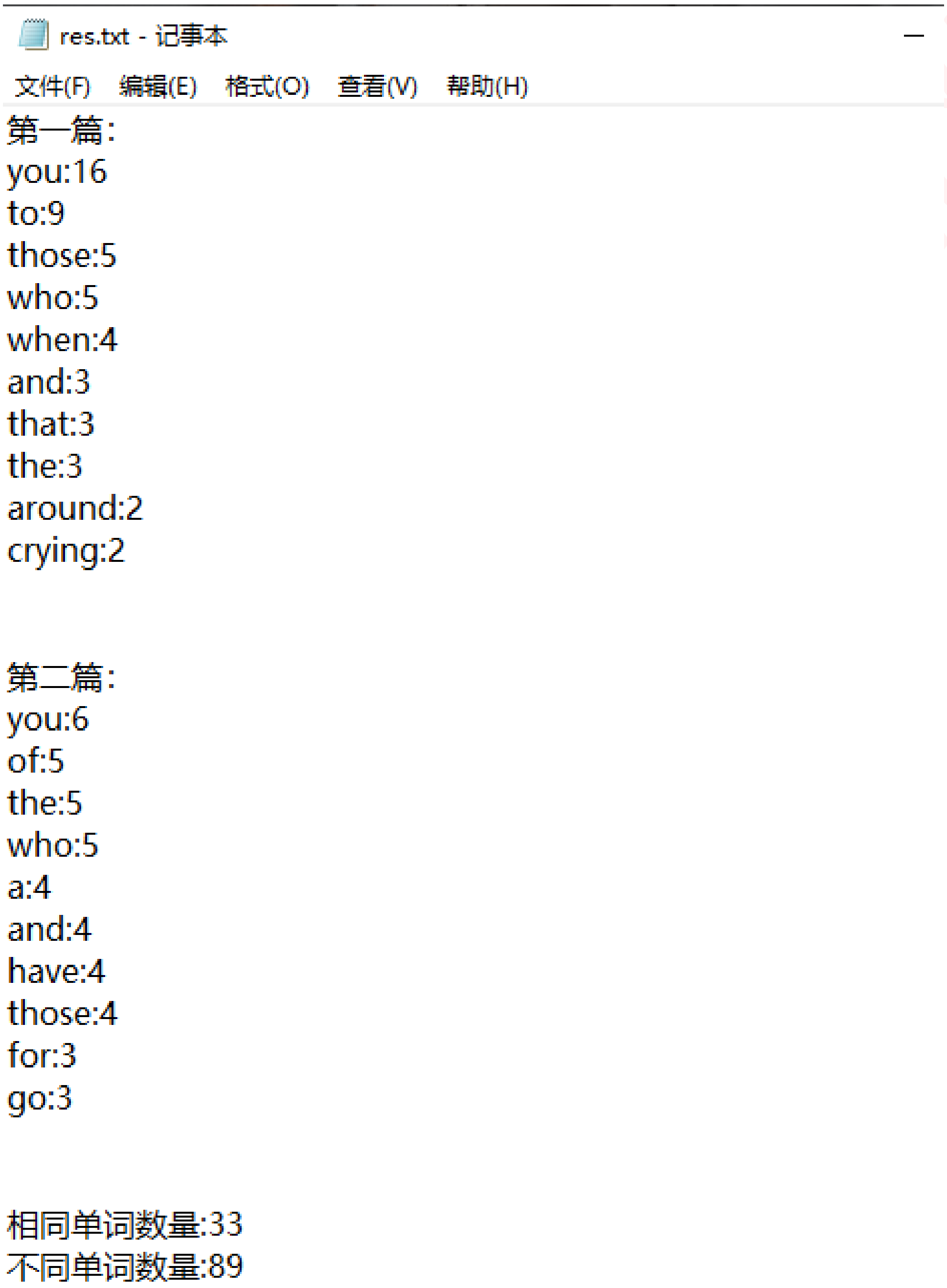
输出结果:



```
go:3


相同单词数量:33
不同单词数量:89
################
统计完毕!

Process finished with exit code 0
```



```
第一篇:
you:16
to:9
those:5
who:5
when:4
and:3
that:3
the:3
around:2
crying:2


第二篇:
you:6
of:5
the:5
who:5
a:4
and:4
have:4
those:4
for:3
go:3


相同单词数量:33
不同单词数量:89
```