

相似度计算

时间限制：1.0 秒

空间限制：512 MiB

刷新 ↺

下载题目目录（样例文件） (/staticdata/down/CSP202403-2.zip)

题目背景

两个集合的 Jaccard 相似度定义为：

$$Sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

即交集的大小除以并集的大小。当集合 A 和 B 完全相同时， $Sim(A, B) = 1$ 取得最大值；当二者交集为空时， $Sim(A, B) = 0$ 取得最小值。

题目描述

除了进行简单的词频统计，小 P 还希望使用 Jaccard 相似度来评估两篇文章的相似性。具体来说，每篇文章均由若干个英文单词组成，且英文单词仅包含“大小写英文字母”。对于给定的两篇文章，小 P 首先需要提取出两者的单词集合 A 和 B ，即去掉各自重复的单词。然后计算出：

- $|A \cap B|$ ，即有多少个不同的单词同时出现在两篇文章中；
- $|A \cup B|$ ，即两篇文章一共包含了多少个不同的单词。

最后再将两者相除即可算出相似度。需要注意，在整个计算过程中应当忽略英文字母大小写的区别，比如 the、The 和 THE 三者都应被视作同一个单词。

试编写程序帮助小 P 完成前两步，计算出 $|A \cap B|$ 和 $|A \cup B|$ ；小 P 将亲自完成最后一步的除法运算。

输入格式

从标准输入读入数据。

输入共三行。

输入的第一行包含两个正整数 n 和 m ，分别表示两篇文章的单词个数。

第二行包含空格分隔的 n 个单词，表示第一篇文章；

第三行包含空格分隔的 m 个单词，表示第二篇文章。

输出格式

输出到标准输出。

输出共两行。

第一行输出一个整数 $|A \cap B|$ ，即有多少个不同的单词同时出现在两篇文章中；

第二行输出一个整数 $|A \cup B|$ ，即两篇文章一共包含了多少个不同的单词。

样例1输入

```
3 2
The tHe thE
the THE
```

样例1输出

```
1
1
```

样例1解释

$A = B = A \cap B = A \cup B = \{\text{the}\}$

样例2输入

```
9 7
Par les soirs bleus dete jirai dans les sentiers
PICOTE PAR LES BLES FOULER LHERBE MENUE
```

样例2输出

```
2
13
```

样例2解释

$A = \{\text{bleus, dans, dete, jirai, les, par, sentiers, soirs}\} \quad |A| = 8$

$B = \{\text{bles, fouler, les, lherbe, menue, par, picote}\} \quad |B| = 7$

$A \cap B = \{\text{les, par}\} \quad |A \cap B| = 2$

样例3输入

```
15 15
Thou that art now the worlds fresh ornament And only herald to the gaudy spring
Shall I compare thee to a summers day Thou art more lovely and more temperate
```

样例3输出

4
24

子任务

80% 的测试数据满足： $n, m \leq 100$ 且所有字母均为小写；
全部的测试数据满足： $n, m \leq 10^4$ 且每个单词最多包含 10 个字母。

语言和编译选项

#	名称	编译器	额外参数	代码长度限制
0	g++	g++	-O2 -DONLINE_JUDGE	65536 B
1	gcc	gcc	-O2 -DONLINE_JUDGE	65536 B
2	java	javac		65536 B
3	python3	python3		65536 B

递交历史

#	状态	时间
---	----	----

当前没有提交权限，请返回认证首页 (/contest/33/home)检查是否已开启模拟认证 或 可以进行自由练习。