中国科学技术大学

本科实验报告

课程名称　　计算机程序设计

实验名称　　作家书写习惯分析（英文）

姓　名　　赵国华

学　院　　信息科学技术学院

系　别　　电子信息（大类）

专　业　　电子信息（大类）

年　级　　大一

学　号　　PB20081589

任课教师　　张四海老师

2024 年 12 月 21 日

一、 实验目的

编程实现若干著作的数据统计，分析作家书写习惯，得出结论。

二、实验要求

1.基本功能：实现书籍内容的统计程序编写，设计统计指标，分析英文作家的书写习惯，如句子长短、单词长短、最喜欢用词等。

2.扩展分析：自选某个（或某些）角度分析小说。如分析男女作家写作风格的差异、写作语言、写作风格随着年代的变化等

三、实验内容

1.数据收集：（1）有作品文件供下载分析（2）也可自行寻找。

2.代码编写：编写读取、分析文章的程序。

3.书写报告：内容详实，图文并茂。

四、算法分析报告:

1.文件读取。

1）单词获取——简易状态机模式:

运用 fgetc 函数逐个读取文档中的字符，出现单词分割符则截断保存为字符串，从而使得单词能够进行后续处理。此处没有直接运用 isalpha 判断是因为单词内部会出现连字符‘-’和引号‘\’‘。此函数处理了不同条件，包括跳过非字母字符、记录单词长度等。

2）句子获取。

利用 fgets 函数和循环结构逐行读取文件，通过指针操作逐个分析字符。遇到句子分隔符句号、问号、感叹号则截止，添加参数 len 动态统计句子的长度。此处没有选择运用单词多少定义句子长度是因为不同单词的长度各不相同，用作统计参数不够准确。

3）获取时没有用 strtok，是为了只遍历一次，降低时间复杂度。

2.数据统计与分析。

1）通用算法：

①利用结构体数组储存不同的单词或不同长度的句子及对应的出现次数。

②调用获取函数对全文边读取边统计，降低时间复杂度。

③运用选择排序法，按照指定数据（如：单词频数、句子频数）对结构体数组排序。

2）单词统计模块：

运用 status 参数对统计函数的代码做简易分类，使得该段代码能够统计不同长度限制的单词，更具移植性。此处没有选择运用 ifdef-endif 结构，是为了使代码面向的对象更容易操作。

此处没有选择用哈希表，是因为尝试之后发现哈希表无法直接进行排序，故最终选择直接用普通的结构体数组存储单词数据。

3）特殊单词的统计：

为了统计指定的单词，在原有的单词统计函数的基础上增加了指定结构体数组参数的传入，利用字符串的比较函数 strcmp 比较得结果。

3.辅助功能：

1）检查内存分配函数：检查内存分配是否成功。

2）文件操作函数：检查文件是否安全打开。

2）结构体数组初始化与释放函数：二维数组的释放。

五．文本分析报告

1.探究目的。

意图通过分析文本的数据，探究不同历史背景下不同作者的写作

风格的变化以及”爱情“这一议题随历史发展产生的变化。随着探究的深入，进一步探究了女性角色在这类小说中的地位随时间发生的变化。

2.文本选择。

以时间为轴，以“爱情“为主题，选取了英语文学中从十六世纪到二十一世纪的十篇不同经典作品，涵盖了戏剧、叙事小说、意识流小说等多种题材。

名称	修改日期	类型
1.Romeo and Juliet - William Shakespeare	2024/12/15 13:55	Text 源文件
2.pride and prejudice - Jane Austen	2024/12/15 17:11	Text 源文件
3.Jane Eyre - Charlotte Brontë	2022/12/27 20:52	Text 源文件
4.Wuthering Heights - Emily Brontë	2022/12/27 20:56	Text 源文件
5.Little Women - Louisa May Alcott	2024/12/15 13:45	Text 源文件
6.Tess of the d'Urbervilles - Thomas Hardy	2024/12/15 15:36	Text 源文件
7.Painted Veil - William Somerset Maugham	2024/12/15 15:30	Text 源文件
8.Mrs. Dalloway - Virginia Woolf	2024/12/15 15:17	Text 源文件
9.Love in the time of cholera - Gabriel Garcí...	2024/12/15 16:43	Text 源文件
10.One Day - David Nicholls	2024/12/15 15:48	Text 源文件

文本背景:

1. **文艺复兴时期**（约 1500 年-1660 年）：英国经历了文化和艺术的繁荣，莎士比亚和培根等文化巨匠活跃于这一时期。
2. **工业革命**（约 1760 年-1840 年）：英国成为世界上第一个工业化国家，经历了巨大的社会和经济变革。
3. **维多利亚时代**（1837 年-1901 年）：以维多利亚女王命名，这一时期英国经历了帝国扩张和工业繁荣。
4. **两次世界大战**（1914 年-1918 年，1939 年-1945 年）：英国在这两次世界大战中扮演了重要角色，战争对英国社会和经济产生了深远影响。
5. **战后时期**（1945 年至今）：英国经历了去殖民化、经济和社会改革，以及加入和退出欧洲联盟的过程。

3.数据分析结果

1）词汇基本数据

作品名称	作家名称	出版日期	单词总数	不同单词数	词汇丰富度	长单词总数	不同长单词数	长单词丰富度	长单词频率
Romeo and Juliet	William Shakespeare	1597	25553	4355	0.17	1577	820	0.52	0.0597
Pride and Prejudice	Jane Austen	1813	121980	7023	0.06	15335	3471	0.23	0.1257
Jane Eyre	Charlotte Brontë	1847	184791	15798	0.09	21164	8045	0.38	0.1145
Wuthering Heights	Emily Brontë	1847	115930	10849	0.09	14018	4987	0.36	0.1209
Little Women	Louisa May Alcott	1868	158373	11582	0.07	14369	4745	0.33	0.0907
Tess of the d'Urbervilles	Thomas Hardy	1891	153086	15250	0.1	18704	7835	0.42	0.1222
The Painted Veil	W. Somerset Maugham	1925	72024	6467	0.09	6028	2370	0.39	0.0837
Mrs. Dalloway	Virginia Woolf	1925	64264	8869	0.14	6877	3497	0.51	0.107
Love in the Time of Cholera	Gabriel García Márquez	1985	149715	11986	0.08	17539	5446	0.31	0.1171
One Day	David Nicholls	2009	130931	13642	0.1	13912	5651	0.41	0.1063

注：①长单词指字符数>=7 的单词。
②词汇丰富度值相异单词总数（即表格中的不同单词数）与单词总数的比值。
③长单词频率指长单词总数与单词总数的比值。

① 篇幅与词汇总量方面：

16 世纪的英国处于文艺复兴时期，人员流动较少，词汇发展不完善。因此这一时期的作品总篇幅短，词汇量较少 。

18-19 世纪的英国经过工业革命之后殖民主义扩张，且美国资本主义发展强劲，英语因此吸收了全世界的词汇而得以长足发展。这一时期的作品篇幅普遍很长，词汇总量相较之前大幅增长。

20-21 世纪，英语词汇量随着全球化的进一步扩充，但是由于作品的风格更加多元，篇幅与用词不再呈现统一的发展趋势。

② 词汇丰富度方面：

由数据看出，词汇丰富度不受制于时代，更取决于作家本身的风格。篇幅最短的 Romeo and Juliet 和 Mrs. Dalloay 的词汇丰富度反而最高，是因为两位作者分别以词汇创新和前景化语言闻名，也即词汇丰富度与作家的词汇创新强相关。

③ 数据不足之处：

尽管作者属于不同年代、不同国家、不同性别，作品的词汇数据也可能趋同，如 Little Women 和 Love in the Time of Cholera，这说明**仅通过词汇基本数据分析作者风格是不够全面的。**

2）单词使用频次数据：

①全体单词中使用最多的前二十个单词

Romeo and Juliet

1. the 出现的次数为 597
2. I 出现的次数为 576
3. and 出现的次数为 478
4. to 出现的次数为 436
5. a 出现的次数为 398
6. of 出现的次数为 355
7. is 出现的次数为 305
8. my 出现的次数为 304
9. in 出现的次数为 291
10. that 出现的次数为 272
11. you 出现的次数为 266
12. me 出现的次数为 264
13. And 出现的次数为 250
14. not 出现的次数为 244
15. thou 出现的次数为 235
16. with 出现的次数为 225
17. be 出现的次数为 198
18. it 出现的次数为 190
19. this 出现的次数为 180
20. for 出现的次数为 163

Price and Prejudice

1. to 出现的次数为 4081
2. the 出现的次数为 4057
3. of 出现的次数为 3607
4. and 出现的次数为 3395
5. her 出现的次数为 2129
6. I 出现的次数为 2055
7. a 出现的次数为 1876
8. was 出现的次数为 1843
9. in 出现的次数为 1783
10. that 出现的次数为 1527
11. not 出现的次数为 1396
12. she 出现的次数为 1381
13. it 出现的次数为 1277
14. be 出现的次数为 1228
15. his 出现的次数为 1191
16. had 出现的次数为 1150
17. you 出现的次数为 1140
18. as 出现的次数为 1131
19. he 出现的次数为 1100
20. for 出现的次数为 1032

Jane Eyre

1. the 出现的次数为 7301
2. I 出现的次数为 7019
3. and 出现的次数为 6247
4. to 出现的次数为 5000
5. of 出现的次数为 4283
6. a 出现的次数为 4220
7. in 出现的次数为 2648
8. you 出现的次数为 2582
9. was 出现的次数为 2490
10. my 出现的次数为 2050
11. it 出现的次数为 2011
12. me 出现的次数为 1998
13. her 出现的次数为 1664
14. that 出现的次数为 1530
15. as 出现的次数为 1495
16. he 出现的次数为 1486
17. had 出现的次数为 1471
18. not 出现的次数为 1455
19. with 出现的次数为 1350
20. is 出现的次数为 1263

Wuthering Heights

1. and 出现的次数为 4435
2. the 出现的次数为 4289
3. I 出现的次数为 3518
4. to 出现的次数为 3452
5. a 出现的次数为 2245
6. of 出现的次数为 2215
7. he 出现的次数为 1558
8. you 出现的次数为 1504
9. her 出现的次数为 1500
10. in 出现的次数为 1410
11. his 出现的次数为 1376
12. that 出现的次数为 1139
13. it 出现的次数为 1116
14. was 出现的次数为 1114
15. she 出现的次数为 1085
16. me 出现的次数为 1042
17. my 出现的次数为 1024
18. him 出现的次数为 907
19. not 出现的次数为 903
20. as 出现的次数为 894

Little Women

1. and 出现的次数为 6736
2. the 出现的次数为 6043
3. to 出现的次数为 4244
4. a 出现的次数为 3672
5. of 出现的次数为 2850
6. her 出现的次数为 2784
7. I 出现的次数为 2734
8. in 出现的次数为 2005
9. it 出现的次数为 1815
10. for 出现的次数为 1800
11. was 出现的次数为 1719
12. she 出现的次数为 1677
13. you 出现的次数为 1669
14. as 出现的次数为 1584
15. with 出现的次数为 1510
16. that 出现的次数为 1417
17. but 出现的次数为 1066
18. Jo 出现的次数为 1053
19. he 出现的次数为 1035
20. his 出现的次数为 886

Tess of the d'Urbervilles

1. the 出现的次数为 8257
2. of 出现的次数为 4491
3. and 出现的次数为 4206
4. to 出现的次数为 4096
5. a 出现的次数为 3128
6. her 出现的次数为 2695
7. in 出现的次数为 2406
8. was 出现的次数为 2118
9. that 出现的次数为 1785
10. had 出现的次数为 1769
11. she 出现的次数为 1752
12. I 出现的次数为 1693
13. as 出现的次数为 1512
14. he 出现的次数为 1406
15. it 出现的次数为 1365
16. you 出现的次数为 1327
17. not 出现的次数为 1191
18. his 出现的次数为 1187
19. for 出现的次数为 1037
20. with 出现的次数为 1025

The Painted Veil

1. the 出现的次数为 2518
2. to 出现的次数为 2188
3. and 出现的次数为 2023
4. a 出现的次数为 1872
5. her 出现的次数为 1582
6. was 出现的次数为 1543
7. of 出现的次数为 1393
8. she 出现的次数为 1335
9. that 出现的次数为 1185
10. I 出现的次数为 1005
11. in 出现的次数为 1002
12. you 出现的次数为 975
13. he 出现的次数为 893
14. had 出现的次数为 860
15. it 出现的次数为 809
16. She 出现的次数为 768
17. with 出现的次数为 651
18. not 出现的次数为 641
19. his 出现的次数为 627
20. him 出现的次数为 557

Mrs. Dalloay

1. the 出现的次数为 2989
2. and 出现的次数为 1621
3. of 出现的次数为 1528
4. to 出现的次数为 1446
5. a 出现的次数为 1317
6. was 出现的次数为 1230
7. her 出现的次数为 1204
8. she 出现的次数为 1139
9. in 出现的次数为 1100
10. had 出现的次数为 908
11. he 出现的次数为 881
12. it 出现的次数为 673
13. that 出现的次数为 595
14. with 出现的次数为 567
15. his 出现的次数为 483
16. on 出现的次数为 440
17. for 出现的次数为 439
18. at 出现的次数为 426
19. him 出现的次数为 408
20. said 出现的次数为 404

Love in the Time of Cholera

1. the 出现的次数为 10009
2. of 出现的次数为 4855
3. to 出现的次数为 3957
4. and 出现的次数为 3860
5. in 出现的次数为 3046
6. he 出现的次数为 2920
7. a 出现的次数为 2810
8. was 出现的次数为 2748
9. that 出现的次数为 2668
10. her 出现的次数为 2474
11. had 出现的次数为 2445
12. his 出现的次数为 2233
13. she 出现的次数为 2113
14. with 出现的次数为 1928
15. for 出现的次数为 1445
16. not 出现的次数为 1425
17. it 出现的次数为 1193
18. him 出现的次数为 1145
19. as 出现的次数为 1111
20. on 出现的次数为 1053

One Day

1. the 出现的次数为 5352
2. and 出现的次数为 4007
3. a 出现的次数为 3062
4. to 出现的次数为 2878
5. of 出现的次数为 2342
6. her 出现的次数为 1818
7. in 出现的次数为 1669
8. I 出现的次数为 1657
9. you 出现的次数为 1588
10. he 出现的次数为 1499
11. his 出现的次数为 1356
12. that 出现的次数为 1335
13. it 出现的次数为 1290
14. she 出现的次数为 1248
15. on 出现的次数为 1094
16. was 出现的次数为 962
17. is 出现的次数为 924
18. with 出现的次数为 882
19. at 出现的次数为 851
20. for 出现的次数为 846

②长单词中使用频率最高的前 20 个单词

Romeo and Juliet

1. BENVOLIO 出现的次数为 63
2. MERCUTIO 出现的次数为 62
3. LAWRENCE 出现的次数为 55
4. Montague 出现的次数为 27
5. Mercutio 出现的次数为 20
6. banished 出现的次数为 18
7. Romeo’s 出现的次数为 16
8. Benvolio 出现的次数为 15
9. gentleman 出现的次数为 14
10. Thursday 出现的次数为 14
11. MUSICIAN 出现的次数为 14
12. daughter 出现的次数为 13
13. tomorrow 出现的次数为 13
14. MONTAGUE 出现的次数为 12
15. Farewell 出现的次数为 12
16. Therefore 出现的次数为 12
17. BALTHASAR 出现的次数为 12
18. marriage 出现的次数为 11
19. Lawrence 出现的次数为 11
20. therefore 出现的次数为 11

Price and Prejudice

1. Elizabeth 出现的次数为 593
2. Catherine 出现的次数为 109
3. pleasure 出现的次数为 92
4. feelings 出现的次数为 88
5. Longbourn 出现的次数为 87
6. Gardiner 出现的次数为 84
7. daughter 出现的次数为 77
8. therefore 出现的次数为 74
9. Netherfield 出现的次数为 73
10. happiness 出现的次数为 71
11. something 出现的次数为 70
12. attention 出现的次数为 68
13. Charlotte 出现的次数为 68
14. marriage 出现的次数为 66
15. character 出现的次数为 64
16. together 出现的次数为 64
17. certainly 出现的次数为 63
18. acquaintance 出现的次数为 63
19. immediately 出现的次数为 61
20. received 出现的次数为 61

Jane Eyre

1. Rochester 出现的次数为 312
2. something 出现的次数为 125
3. Thornfield 出现的次数为 99
4. continued 出现的次数为 71
5. yourself 出现的次数为 70
6. answered 出现的次数为 69
7. pleasure 出现的次数为 68
8. returned 出现的次数为 64
9. sometimes 出现的次数为 58
10. appeared 出现的次数为 58
10. appeared 出现的次数为 58
11. anything 出现的次数为 56
12. followed 出现的次数为 55
13. Brocklehurst 出现的次数为 53
14. Rochester’s 出现的次数为 49
15. feelings 出现的次数为 47
16. question 出现的次数为 45
17. remember 出现的次数为 45
18. together 出现的次数为 45
19. features 出现的次数为 45
20. character 出现的次数为 45

Wuthering Heights

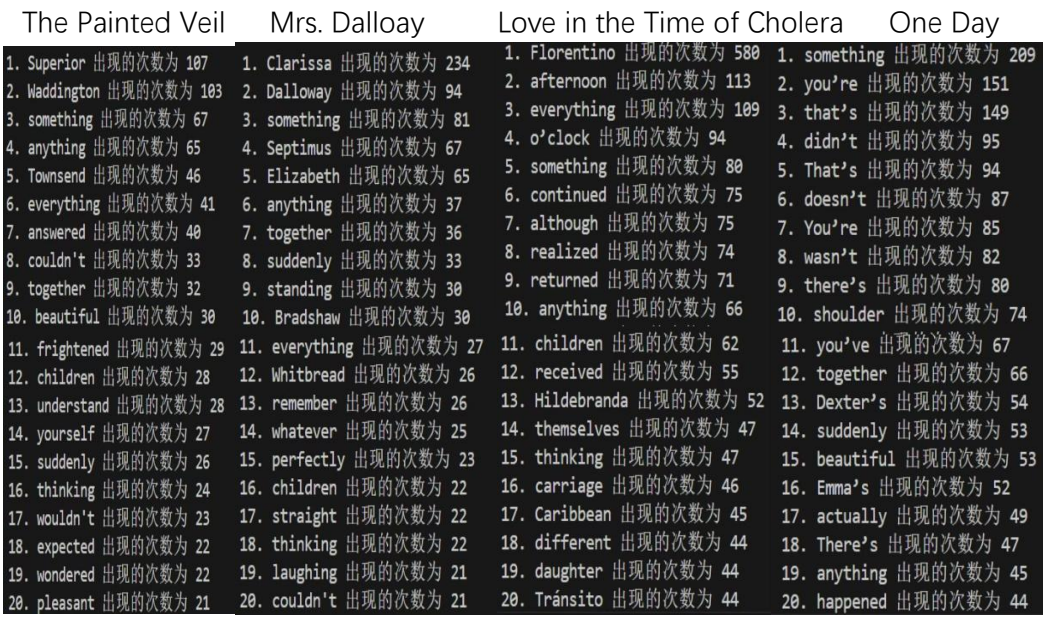
1. Heathcliff 出现的次数为 410
2. Catherine 出现的次数为 332
3. answered 出现的次数为 155
4. Earnshaw 出现的次数为 109
5. exclaimed 出现的次数为 72
6. continued 出现的次数为 66
7. returned 出现的次数为 62
8. Wuthering 出现的次数为 61
9. Linton’s 出现的次数为 58
10. you’ll 出现的次数为 57
11. Heathcliff’s 出现的次数为 54
12. Isabella 出现的次数为 54
13. didn’t 出现的次数为 52
14. something 出现的次数为 49
15. mistress 出现的次数为 46
16. wouldn’t 出现的次数为 45
17. upstairs 出现的次数为 43
18. Catherine’s 出现的次数为 43
19. yourself 出现的次数为 41
20. observed 出现的次数为 39

Little Women

1. didn’t 出现的次数为 149
2. anything 出现的次数为 135
3. something 出现的次数为 134
4. gentleman 出现的次数为 75
5. Laurence 出现的次数为 75
6. wouldn’t 出现的次数为 74
7. children 出现的次数为 70
8. couldn’t 出现的次数为 70
9. pleasant 出现的次数为 68
10. everything 出现的次数为 65
11. answered 出现的次数为 65
12. returned 出现的次数为 62
13. together 出现的次数为 61
14. everyone 出现的次数为 53
15. expression 出现的次数为 53
16. doesn’t 出现的次数为 53
17. splendid 出现的次数为 50
18. yourself 出现的次数为 49
19. That’s 出现的次数为 48
20. you’ll 出现的次数为 47

Tess of the d’Urbervilles

1. d’Urberville 出现的次数为 134
2. Durbeyfield 出现的次数为 118
3. Tess’s 出现的次数为 109
4. something 出现的次数为 89
5. didn’t 出现的次数为 59
6. themselves 出现的次数为 58
7. children 出现的次数为 58
8. together 出现的次数为 57
9. Gutenberg-tm 出现的次数为 55
10. thinking 出现的次数为 53
11. anything 出现的次数为 52
12. continued 出现的次数为 51
13. murmured 出现的次数为 50
14. dairyman 出现的次数为 49
15. remained 出现的次数为 48
16. distance 出现的次数为 47
17. Clare’s 出现的次数为 45
18. followed 出现的次数为 44
19. position 出现的次数为 43
20. afternoon 出现的次数为 42



第一部分数据分析：（全体单词）

① 数据共性：

所有作品中使用最多的前 20 个单词都是**常用虚词**，包括冠词、介词、连词等，这符合英文写作的习惯，因而**不具有写作风格的代表性**。

② 数据特性：

在 Romeo and Juliet 中频繁出现 **thou**，其他文章中却没有见到。这是因为古英语中第二人称常用 thou,后来被 you 代替。该数据反映了**英语写作习惯用词随时间的变化**。

第二部分数据分析：（长单词）

① 数据共性：

所有作品都频繁出现**重要角色或重要地点的名字**，说明尽管存在时间和体裁等多方面的不同，但作者写作时都**遵循小说的三要素原则**。

② 数据特性：

不同作者频繁使用的**表示“说”的单词不同**，反映了不同作品中的**人物特点。**如：Jane Eyre 中使用 answered 符合女主角是家庭教师而男主角是贵族这一身份设定；Wuthering Heights 中使用 exclaim 符合男主角性格古怪的特点，使用 observed 符合叙述者是外来旅客因此常常观察这一身份特点；Tess of the d'Urbervilles 中使用 murmur 则符合苔丝独自一人在外漂泊的经历。

常用长单词也能反映作者的**语言特色**。同为 19 世纪的作品，Little Women 相较于其他四部作品，所用的词汇更为简单，且缩写词居多，这与 21 世纪的 One Day 类似，反映出两位作者朴实无华、偏口语化的写作风格。

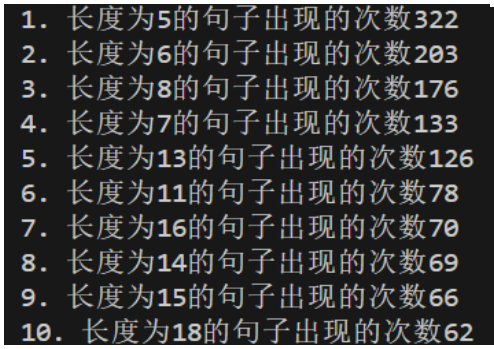
3）句子基本数据：

作品名称	作家名称	出版日期	句子总数	平均每句单词数	感叹号	问号	句号	逗号	冒号
Romeo and	William Sha	1597	3119	8.1927	248	369	2513	2664	69
Pride and P	Jane Austen	1813	6844	17.8229	496	443	6321	9657	43
Jane Eyre	Charlotte Br	1847	10436	17.7071	932	1491	8369	14483	2769
Wuthering H	Emily Bront	1847	7066	16.4067	1329	778	5144	9905	1165
Little Wome	Louisa May	1868	8701	18.2017	662	744	7731	16049	15
Tess of the	Thomas Ha	1891	8274	18.5021	1037	735	6869	10594	127
The Painted	W. Somerse	1925	5325	13.5256	55	448	4963	2995	123
Mrs. Dallow	Virginia Wool	1925	3793	16.9423	346	361	3148	6107	33
Love in the	Gabriel Garc	1985	5065	29.5587	23	31	5136	9258	647
One Day	David Nicho	2009	11521	11.3646	680	1917	9921	9949	150

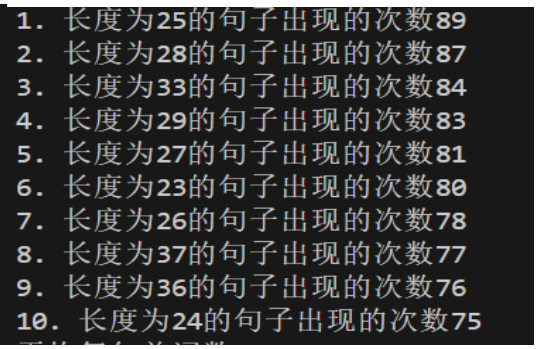
注：该表格转换时作家名称未完全展示，全名可参照报告第四页。

不同长度句子分布情况：

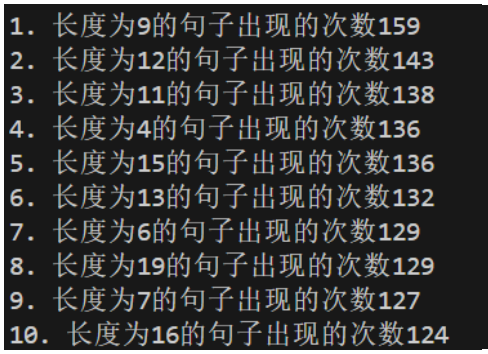
Romeo and Juliet



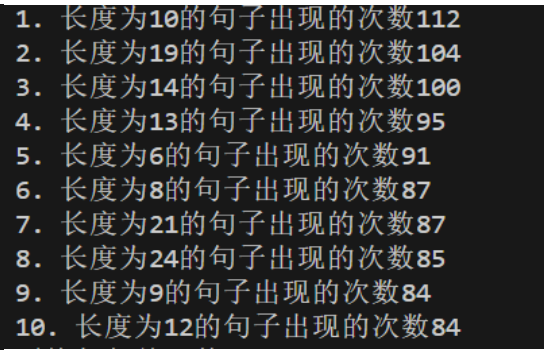
Price and Prejudice



Jane Eyre

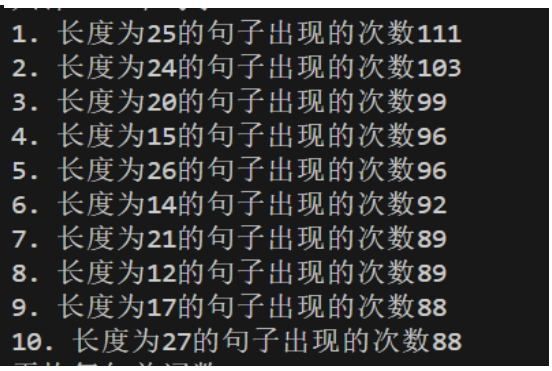


Wuthering Heights

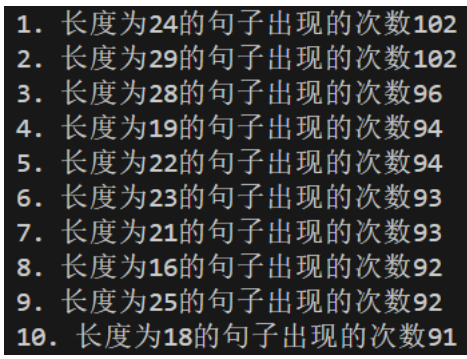


Little Women

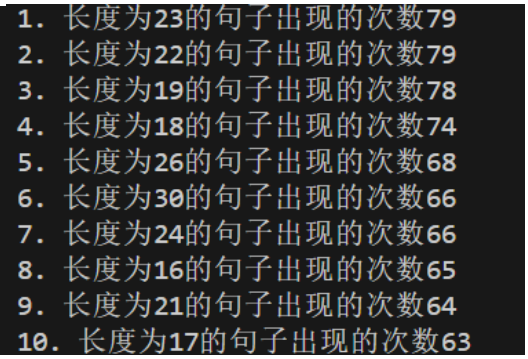
Tess of the d'Urbervilles



The Painted Veil



Mrs. Dalloay



Love in the Time of Cholera

```
1. 长度为105的句子出现的次数39
2. 长度为140的句子出现的次数39
3. 长度为132的句子出现的次数37
4. 长度为87的句子出现的次数35
5. 长度为52的句子出现的次数35
6. 长度为70的句子出现的次数35
7. 长度为126的句子出现的次数35
8. 长度为144的句子出现的次数33
9. 长度为158的句子出现的次数33
10. 长度为109的句子出现的次数32
```

One Day

```
1. 长度为11的句子出现的次数285
2. 长度为8的句子出现的次数284
3. 长度为15的句子出现的次数272
4. 长度为20的句子出现的次数266
5. 长度为13的句子出现的次数262
6. 长度为10的句子出现的次数262
7. 长度为9的句子出现的次数260
8. 长度为16的句子出现的次数257
9. 长度为7的句子出现的次数251
10. 长度为18的句子出现的次数238
```

数据分析结果：

① 句子长度与作者写作风格强相关：

Shakespeare 精于戏剧写作，内容围绕人物对话展开，句子最少且长度最短。19 世纪的五位作家文风类似，句子数量和长度也接近。Maugham 和 Marquez 则形成鲜明对比，二者句子总数接近但句子长度却明显不同，其中 Marquez 的作品平均句子长度位居首位。这体现出前者简洁明快的现实主义写法和后者诗意瑰丽的魔幻现实主义风格。21 世纪的 David Nicholls 则文风朴素，句子很多但长度很短。

② 标点符号的使用与作者的写作情感强相关：

通过分析可以看出，相较于男性作家而言，所选取的女性作家在作品中使用感叹号的频率更高，也即情感流露更多。特别需要注意的是，同样以爱情为主题，作家的情感诠释并不相同。使用感叹号频率最高的 Emily Bronte 塑造的 Heathcliff 因失去爱人变得歇斯底里，情感的表达较为激烈；而使用感叹号频率最少的 Marquez 塑造的 Florentino 同样爱而不得，但以其老者身份展开的叙述显然更为和缓。

4）特殊词汇分析

作品名称	作家名称	出版日期	love	romance	happiness	dream	freedom	dignity
Romeo and Juliet	William Shakespeare	1597	139	0	1	10	0	0
Pride and Prejudice	Jane Austen	1813	102	0	71	0	4	5
Jane Eyre	Charlotte Brontë	1847	184	3	21	32	3	1
Wuthering Heights	Emily Brontë	1847	115	2	10	13	0	3
Little Women	Louisa May Alcott	1868	195	18	25	6	6	9
Tess of the d'Urbervilles	Thomas Hardy	1891	162	0	21	18	2	3
The Painted Veil	W. Somerset Maugham	1925	159	1	6	6	12	3
Mrs. Dalloway	Virginia Woolf	1925	107	1	6	2	1	7
Love in the Time of Cholera	Gabriel García Márquez	1985	399	0	20	30	3	9
One Day	David Nicholls	2009	143	7	4	2	3	1

作品名称	作家名称	男主角	女主角	第一人称	第二人称	第三人称
Romeo and Juliet	William Shakespeare	133	56	641	288	290
Pride and Prejudice	Jane Austen	373	593	2307	1349	3644
Jane Eyre	Charlotte Brontë	312	317	7322	2899	3812
Wuthering Heights	Emily Brontë	410	332	3828	1685	3453
Little Women	Louisa May Alcott	459	1053	3184	1888	3763
Tess of the d'Urbervilles	Thomas Hardy	199	775	1933	1504	4843
The Painted Veil	W. Somerset Maugham	416	155	1138	1102	3834
Mrs. Dalloway	Virginia Woolf	164	234	131	112	3041
Love in the Time of Cholera	Gabriel García Márquez	580	454	199	182	6971
One Day	David Nicholls	557	583	1906	1870	4666

数据分析结果：

① 十部作品中 **love** 一词使用频率均远远高于别的词汇，这符合“爱情“的主题。

② **十九世纪**的作品中，除了 pride and prejudice 之外，其余作品使用 **happiness 和 dream** 的频率与 20 世纪末期的 Love in the Time of Cholera 趋于一致。尤其突出的是 Little Women，其使用 freedom 和 dignity 的频率在十部作品中均很高，且女主角名字的频数远高于男主角。这足以说明，随着工业革命的完成，英国和

美国均进入快速发展的时代，女性主义开始萌芽。这一时期的作者对于爱情的看法发生了一定程度的转变，开始追求女性在爱情之外的独立与自由。

③ **Mrs. Dallowa**y 虽然出版于 20 世纪，但故事的主人公设定为维多利亚时代的上层贵妇，通过数据分析也可以看出该文本中给定的特殊词汇出现的频次均很少，这事实上是女性主义进一步发展之后作者站在新的角度对过去时代的分析与映照当下的反思。作者此时的写作主题看似围绕“爱情“展开，实际上是对女性精神内核的探索。

④ 从**人称**上看，Charlotte Brontë 写作以第一人称为主，Emily Brontë 和 Louisa May Alcott 第一人称和第二人称兼重，其余作者都偏重于使用第三人称。

六、实验心得

1.编程能力的进步与新知识的探索。

从最开始的算法分析到编程实现再到修改完善，在做大作业的过程中我多次修改探索更优方案。在此过程中我的模块化编程能力显著提升，将不同功能的代码抽象成不同的函数，提高了代码的可移植性和可修改性。除此之外，我对文件读写与指针的结合运用有了更深刻的认知，同时学习到了字符串处理相关的更多函数，如 strdup 函数等。在此过程中我认识到能够熟练掌握库函数与自己编程实现同等重要。另外，我对数据结构与算法有了初步的了解，学会了使用 Hash 表。尽管最终因其不利于排序这一小缺点而选择运用了自己编写的结构体，但也掌握了一种更优的方法。

2.锻炼了理论与实践结合的思维。

通过这次实验，我深刻地认识到学习技术的最终目的是解决实际问题。在实验的过程中，我的程序曾经出现过没有正确处理内存分配而导致的爆栈问题，也出现过因连字符导致的单词隔断问题，随后我又针对问题思考最终将其解决。在撰写报告的过程中，我发现了一些数据的异常或是特征，这促使我查找自己程序中的逻辑问题以及为程序设置新的分析指标，报告结果因此更完善更深入。通过实践来检验并解决问题是编程能力提升的关键，也是学习任何知识的必要途径。

3.探索文理结合的可能以及未来职业发展的铺垫。

最初选题的时候我便处于对文学作品的热爱选择了文本分析，此次实验让我认识到文理结合的巨大优势。当我用自己学习到的编程知识以理性的方式分析曾经一读再读的感性的文学经典时，我得以站在更加宏观的视角上分析问题，得到的结论也因大量的数据建立在纵深的时间轴之上更具客观性。通过数据分析寻找选取的十篇作品的共性与特性，我得以梳理出一条不同作家在经典爱情文学中的书写风格随着时代发展与女性主义兴起的变化脉络，也意识到作家依然具有独立于时代之外的洞察力与独特性。利用计算机强大的处理能力抽丝剥茧般分析人类文学史上的经典，或许会给我们带来更多有益的启示。身处信息时代，作为科大的信息学院学子，通过这次实验我认识到当我所学的知识与我的兴趣结合在一起之后所迸发出的力量如此巨大，这为我将来的职业发展埋下了一颗小小的种子。

4.实验的可优化部分。

事实上我的实验依然有其不足之处，尤其是算法方面的优化。我

将继续学习正则表达式等库函数、学习更高效的数据结构以及争取在我现有代码基础上实现对中文作品的分析。

依然感谢这次大作业的机会带给我的成长。

谢谢老师认真读完我的实验报告!