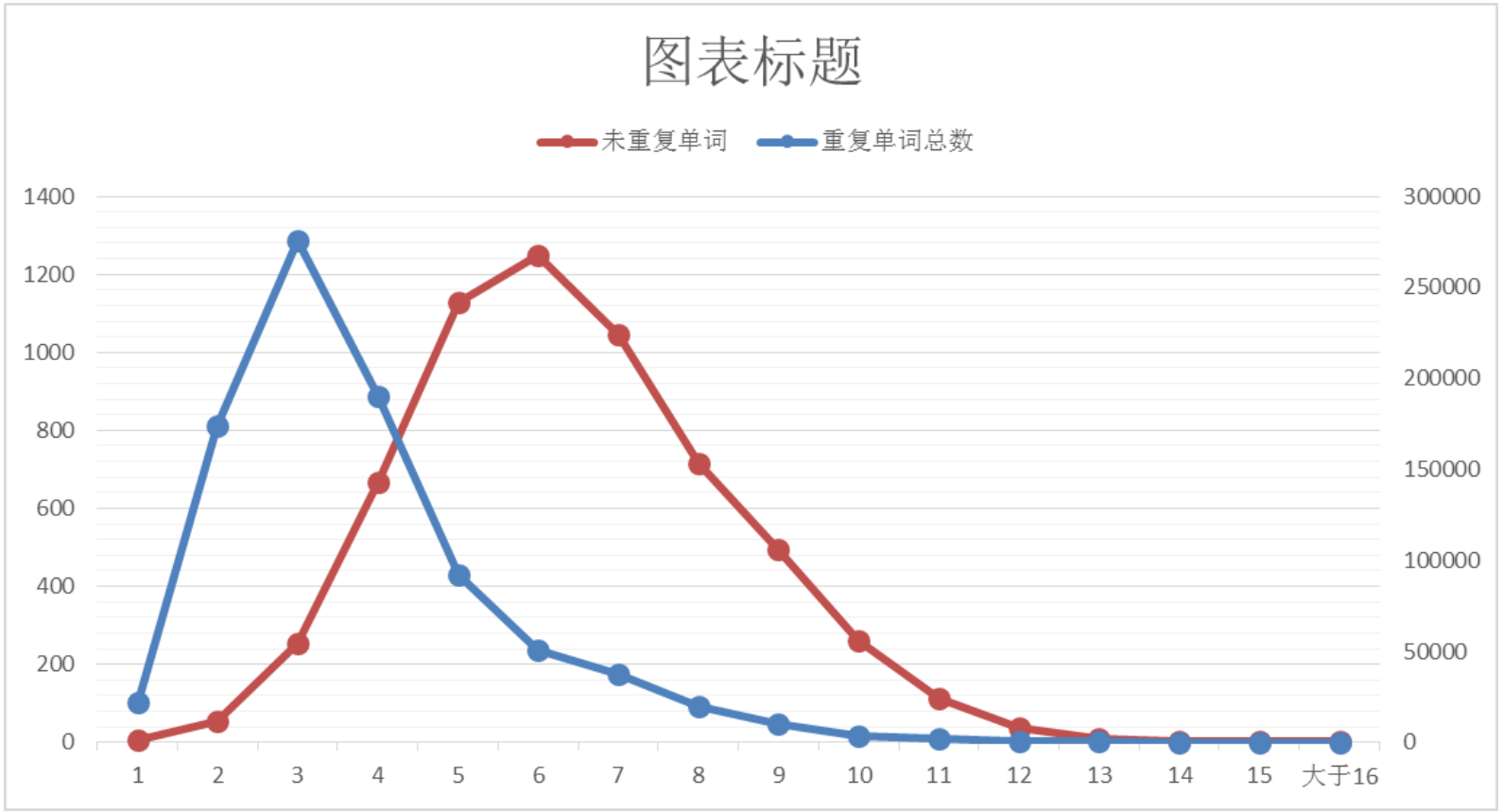
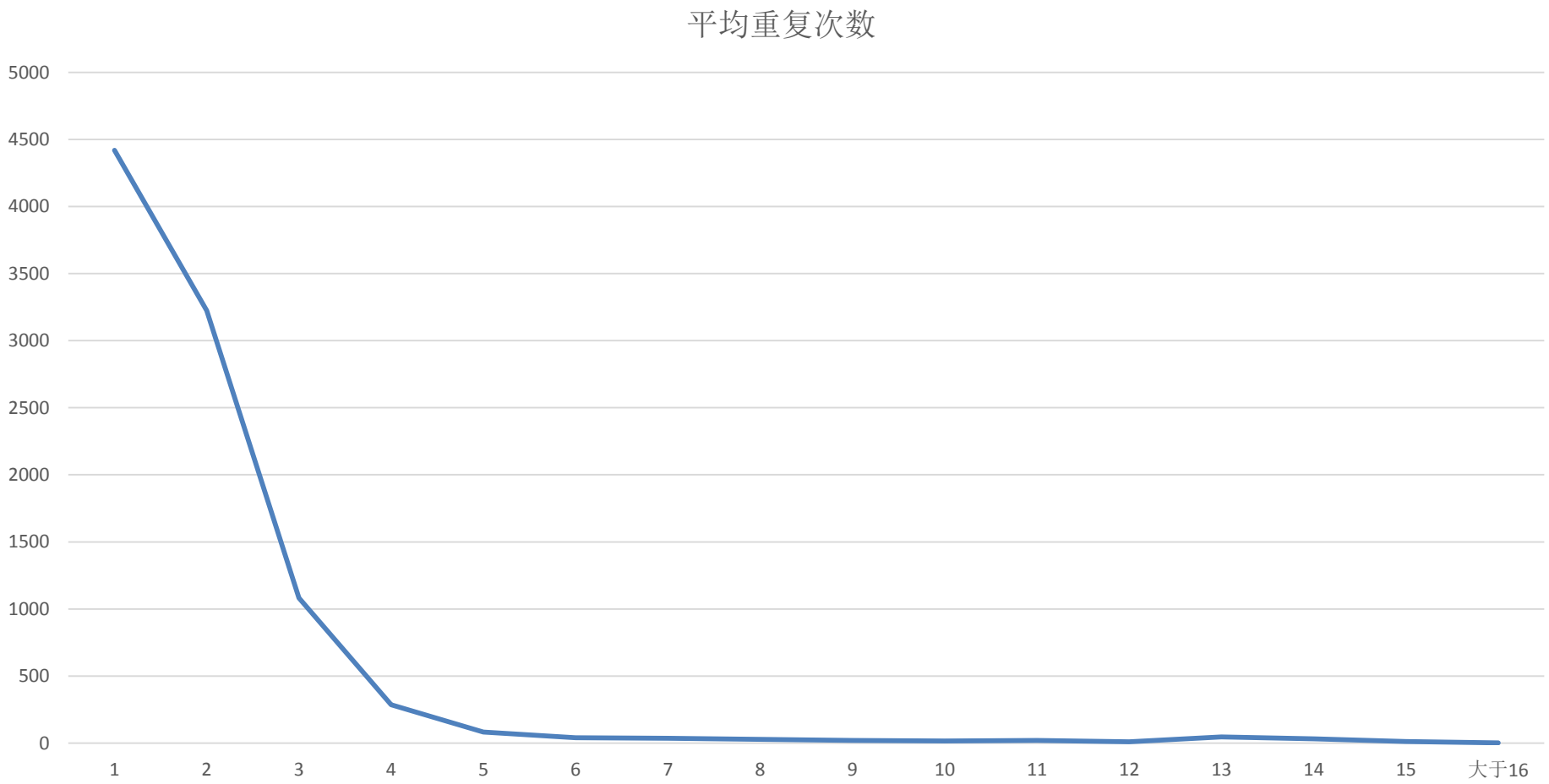
如何高效进行英语学习

冯佳武

BIBLE



单词字母个数	1	2	3	4	5	6	7	8	9
重复单词总数	22102	174118	275843	190253	91947	50584	37293	20049	9861
未重复单词	5	54	255	668	1128	1250	1047	714	494
平均重复次数	4420.40	3224.41	1081.74	284.81	81.51	40.47	35.62	28.08	19.96
单词字母个数	10	11	12	13	14	15	大于16	总计	
重复单词总数	3826	2157	356	417	97	33	1	878937	
未重复单词	260	114	38	9	3	3	1	6043	
平均重复次数	14.71538	18.921053	9.368421	46.3333	32.3333	11	1	145.45	


平均重复次数
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
大于16

剔除123

摘要

- 我不喜欢的长度
- A-Z，你最讨厌哪一个
- 你最烦谁
- 我的语料库
- 讨论---文本增加对于单词量的增加

文本来源

小说

名称	文件大小
1.圣经 n1	4467663
2.简爱 n2	1114997
3.雾都孤儿 n3	981553
4.罗密欧和朱丽叶 n4	145397
5.星球大战 n5	1033366
6.哈利波特 n6	2315757
7.安徒生童话 n7	2020791
8.悲惨世界 n8	1666567
9.圣诞颂歌 n9	189407
10.伊索寓言 n10	65752

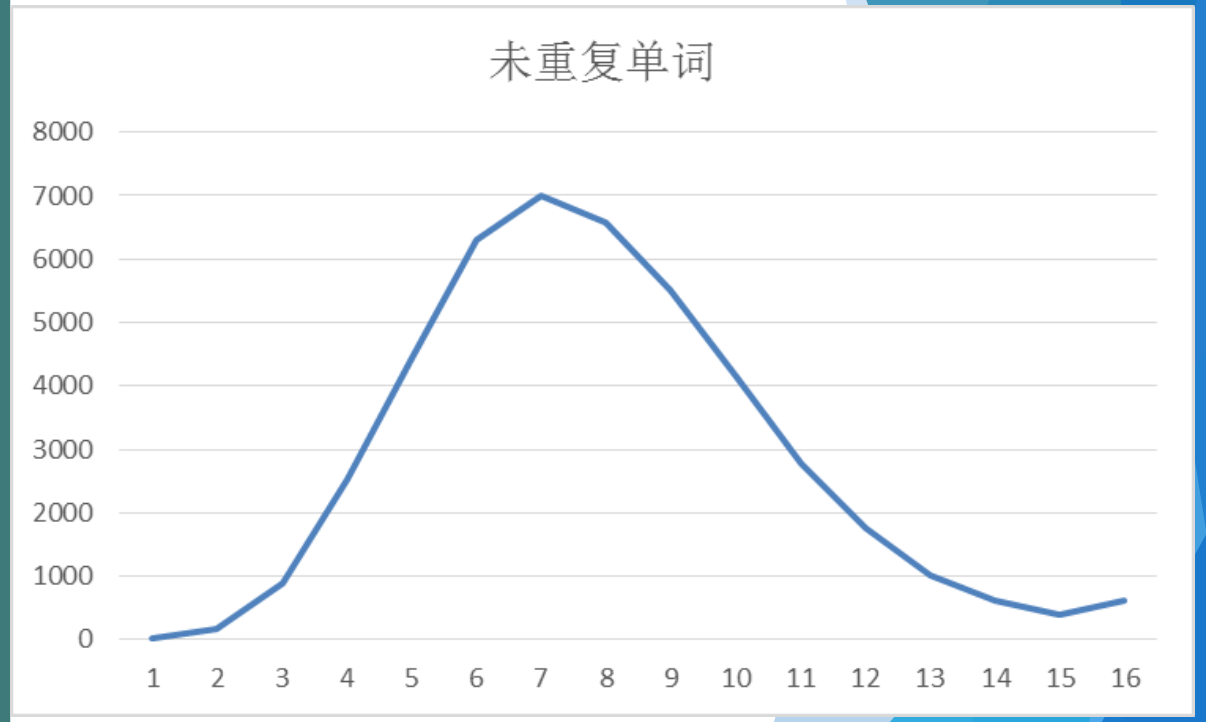
生物专业文献摘要

名称	文件大小
1. 3D genome	5604902
2. Integrative structural biology	4787244
3. Lung cancer dna methylation	4940000

我不喜欢的长度

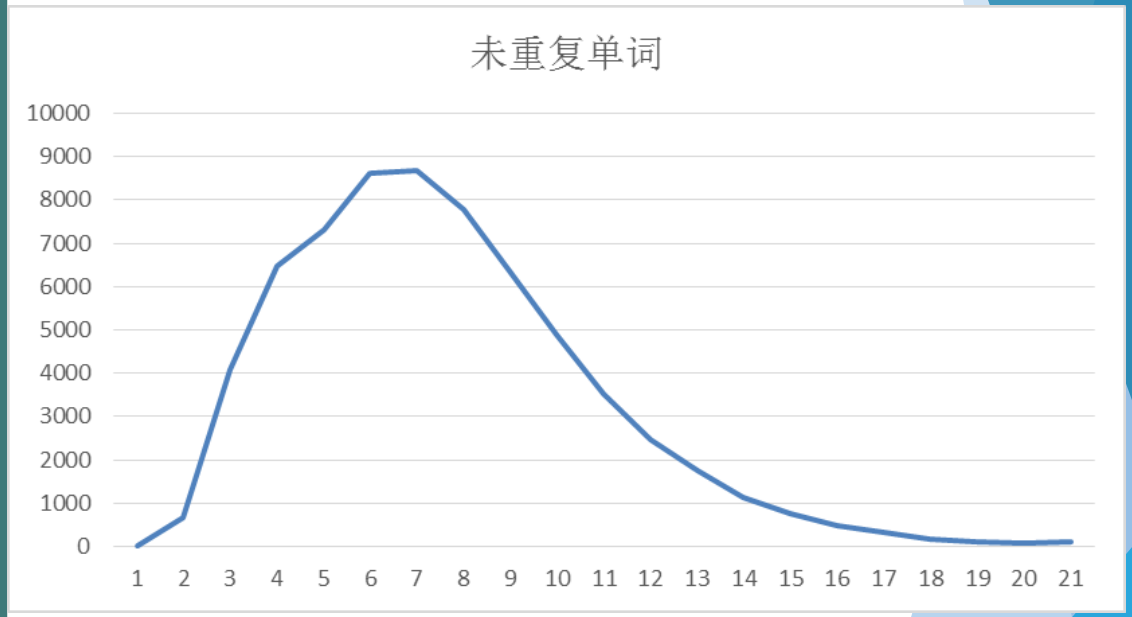
小说

单词字母个数	重复单词总数	未重复单词	平均重复次数
1	101690	25	4067.6
2	442205	172	2570.9593
3	670052	891	752.022447
4	501961	2516	199.507552
5	285780	4452	64.1913747
6	190588	6312	30.1945501
7	145664	7010	20.7794579
8	91632	6569	13.9491551
9	55629	5514	10.0886834
10	28501	4146	6.87433671
11	13762	2778	4.95392369
12	6630	1770	3.74576271
13	3458	1021	3.38687561
14	1129	615	1.83577236
15	601	398	1.51005025
大于16	672	614	1.09446254
总计	2539954	44803	56.6916055

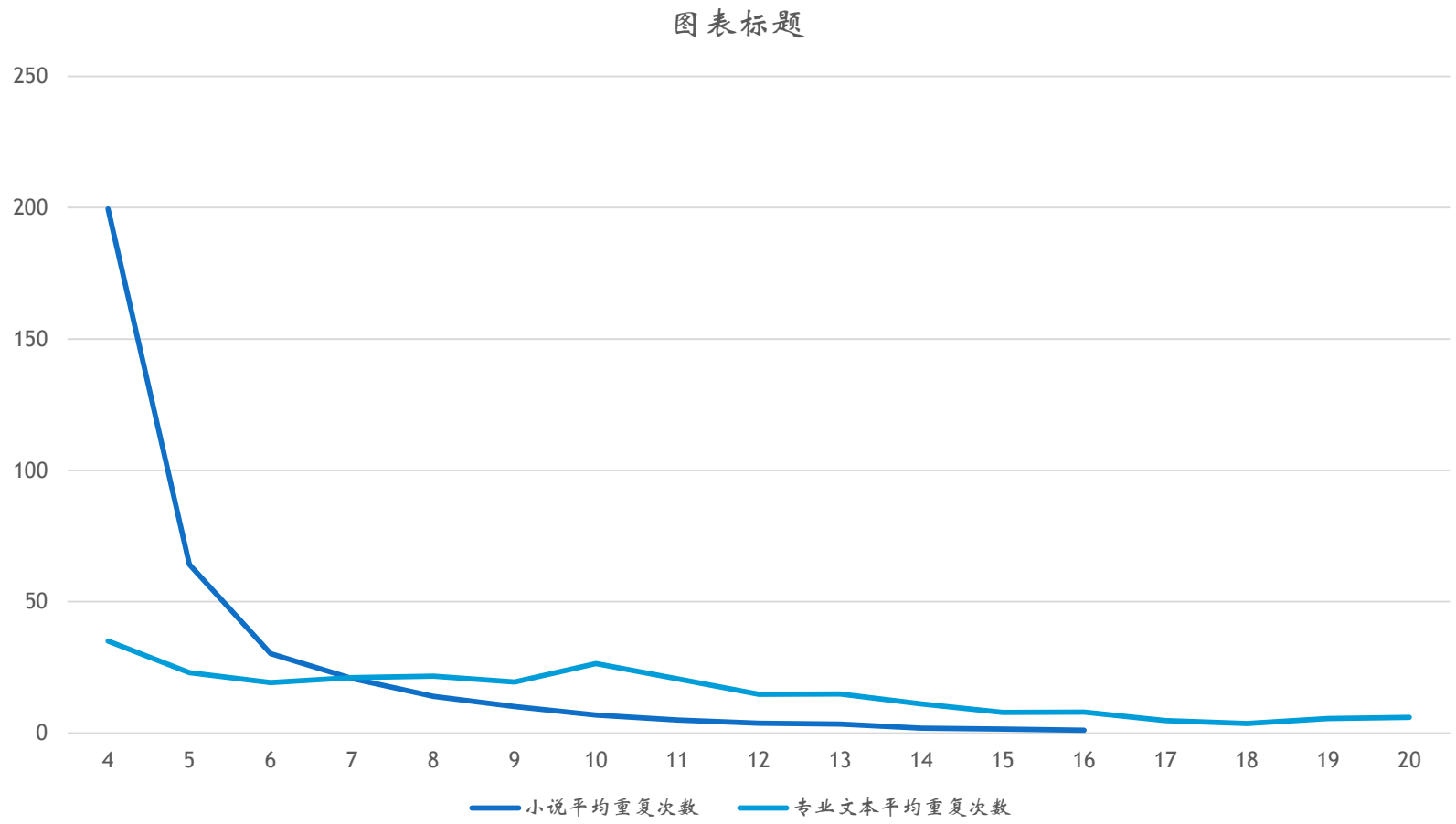


我不喜欢的长度

生物专业文本 单词字母个数	重复单词总数	未重复单词	平均重复次数
1	120451	26	4632.73
2	308068	664	463.96
3	331553	4075	81.36
4	226086	6465	34.97
5	168027	7323	22.95
6	164834	8607	19.15
7	183624	8692	21.13
8	169082	7792	21.70
9	123269	6327	19.48
10	128732	4878	26.39
11	72634	3520	20.63
12	36365	2470	14.72
13	26131	1753	14.91
14	12714	1145	11.10
15	6089	770	7.91
16	3886	486	8.00
17	1537	321	4.79
18	680	185	3.68
19	573	103	5.56
20	505	84	6.01
大于21	242	113	2.14
总计	2085082	65799	31.69



我不喜欢的长度


图表标题
250
200
150
100
50
0
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
小说平均重复次数
专业文本平均重复次数

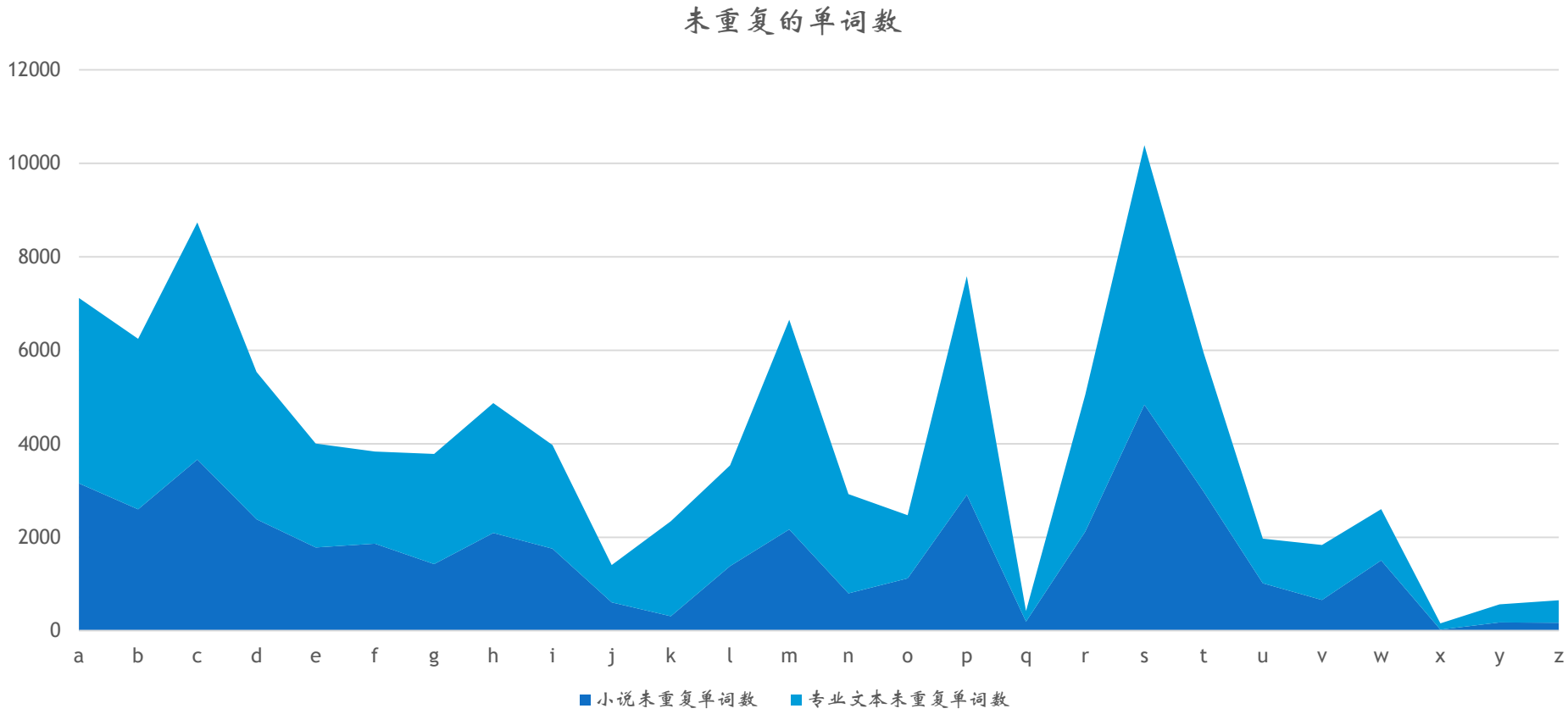
我不喜欢的长度

- 1. 文本的长度维持在7个字母时单词数目最多
- 2. 少于3个字母的重复是十分严重的
- 3. 随着单词的重复长度的增加，重复是递减的，但是专业词汇在10和11个单词时有明显回升，可以推测部分专业名词可能在这个范围内，也有可能是由于文本搜索的关键词影响的。
- 4.专业文本的单词重复是明显低于小说类的
- 5.在6到7个字母以上的长度的时，专业文本单词的长度会重复数有所增加

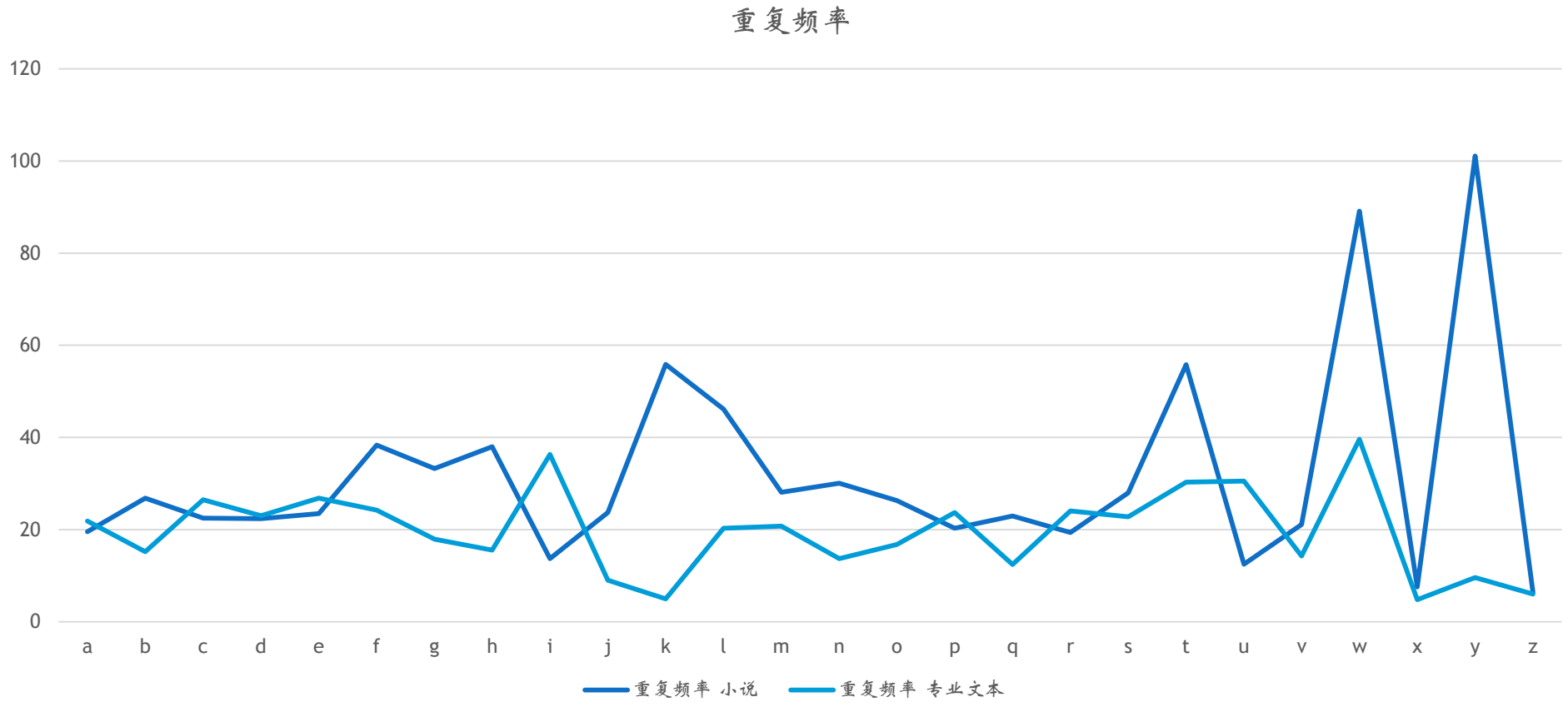
A-Z，你最讨厌哪一个

小说				专业文本			
字母	重复单词数	未重复单词数	重复频率	字母	重复单词数	未重复单词数	重复频率
a	61566	3152	19.53	a	86768	3967	21.87
b	69699	2597	26.84	b	55318	3648	15.16
c	82244	3662	22.46	c	134300	5073	26.47
d	53323	2384	22.37	d	72427	3152	22.98
e	41807	1781	23.47	e	59598	2223	26.81
f	71319	1862	38.30	f	47737	1971	24.22
g	47343	1425	33.22	g	42166	2357	17.89
h	79482	2091	38.01	h	43207	2779	15.55
i	23990	1755	13.67	i	80583	2218	36.33
j	14365	606	23.70	j	7200	799	9.01
k	17317	310	55.86	k	10106	2033	4.97
l	63731	1382	46.12	l	43686	2156	20.26
m	60832	2167	28.07	m	93096	4486	20.75
n	24013	799	30.05	n	29061	2124	13.68
o	29436	1120	26.28	o	22662	1351	16.77
p	58928	2907	20.27	p	110811	4678	23.69
q	4447	194	22.92	q	2812	227	12.39
r	41036	2119	19.37	r	70305	2926	24.03
s	135345	4835	27.99	s	126360	5553	22.76
t	165989	2976	55.78	t	89966	2971	30.28
u	12678	1015	12.49	u	29139	955	30.51
v	13895	658	21.12	v	16803	1176	14.29
w	134025	1504	89.11	w	43427	1097	39.59
x	182	24	7.58	x	637	133	4.79
y	17890	177	101.07	y	3704	387	9.57
z	1081	170	6.36	z	2889	481	6.01

A-Z，你最讨厌哪一个


未重复的单词数
12000
10000
8000
6000
4000
2000
0
a b c d e f g h i j k l m n o p q r s t u v w x y z
小说未重复单词数
专业文本未重复单词数

A-Z，你最讨厌哪一个


重复频率
120
100
80
60
40
20
0
a b c d e f g h i j k l m n o p q r s t u v w x y z
重复频率 小说
重复频率 专业文本

A-Z，你最讨厌哪一个

- 1. 对于A-Z的字母来说，无论对于专业文本和小说，S开头拥有的是最多的单词，则最少的则是以P开头的单词。
- 2. 对于重复出现来说A-Z中对于专业文本和小说来说，分别为C和T，可能这就是关于文本内容的差异了
- 3. 可以明显看出在A-Z中每一个字母开头的单词在专业文本都有所增加。整体上都会比小说中要多。
- 4.对于重复频率来说当然是小说的单词中重复概率会高出很多。

你最烦谁---小说

that
with
said
which
they
their
have
will
them
from
this
were
there
your
come
what
when
then
lord
harry
been
made
over
make
before
give
back
like
into
down
would
came
could
take
went
people
little
about
more
away
very
time
king
house
good
great
place
those

你最烦谁---专业文本

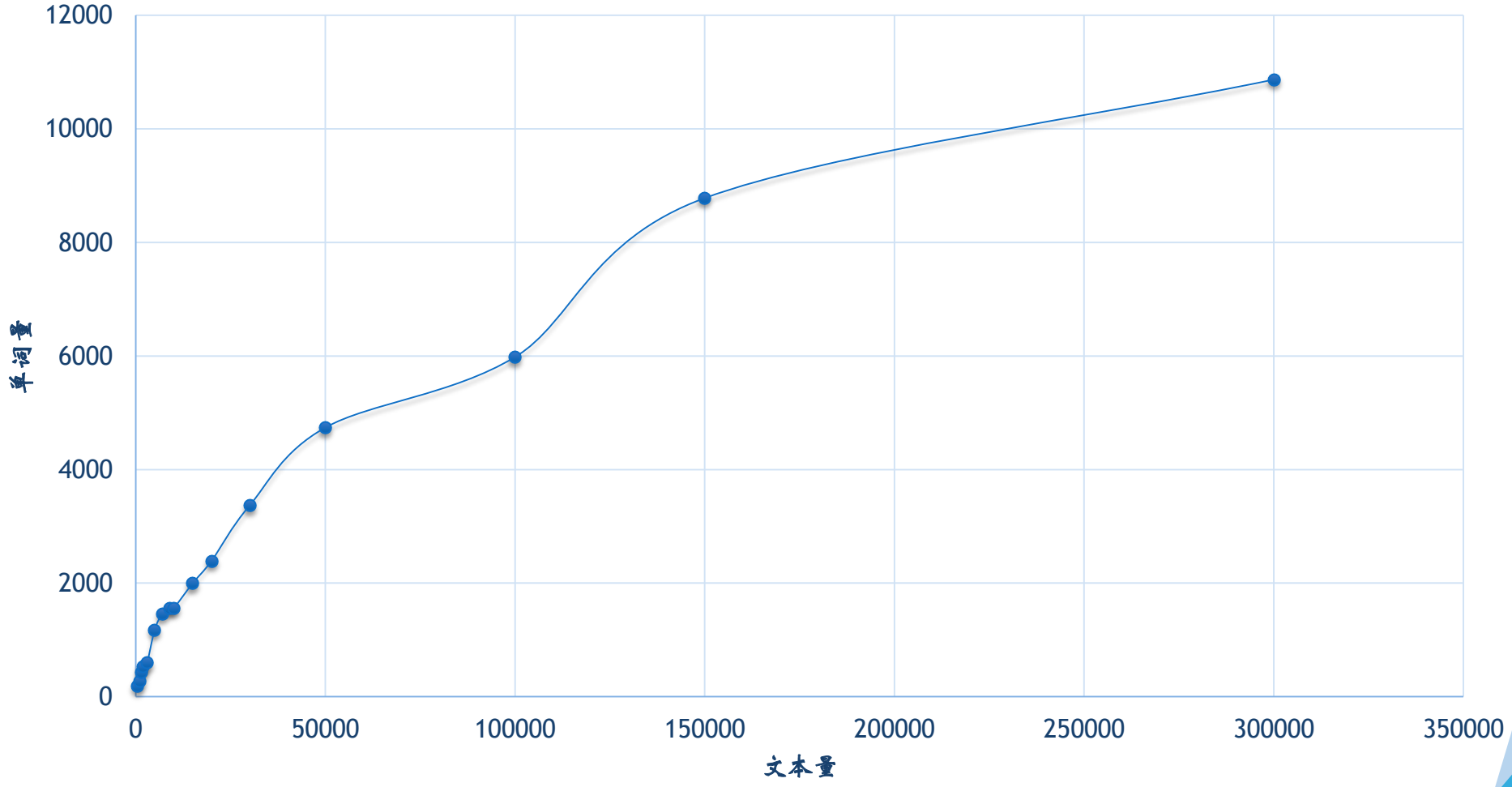
analysis
biology
university
information
cancer
pmid
structural
medline
genome
with
have
human
expression
author
molecular
indexed
pubmed
methylation
department
cell
center
protein
gene
that
were
with
lung
institute
from
genes
cells
these
research
promoter
tumor
medicine

你最烦谁

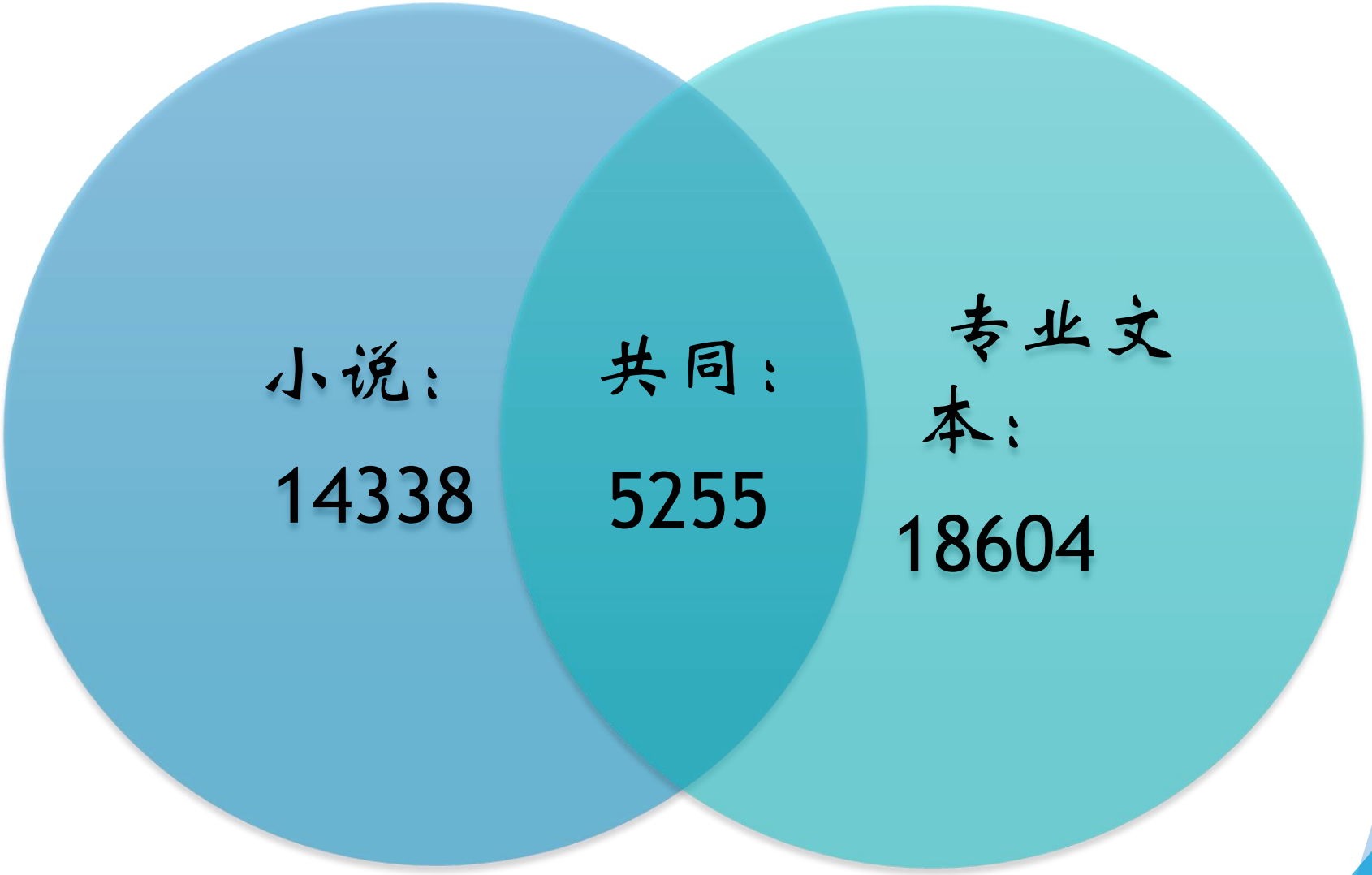
单词	次数	单词	次数
with	15106	structural	4650
that	14763	genes	4549
cancer	10913	molecular	4401
university	10414	center	4123
cell	8924	which	3965
department	8524	human	3960
methylation	8391	genome	3903
lung	8175	analysis	3901
information	7791	have	3751
gene	7108	promoter	3657
pubmed	6950	tumor	3594
from	6937	medicine	3370
pmid	6918	structure	3368
author	6916	epub	3353
expression	6774	using	3155
were	6610	results	3131
this	6558	pmcid	3102
institute	6486	data	3096
protein	6097	proteins	3061
medline	5905	between	2999
indexed	5891	binding	2794
biology	5562	medical	2778
cells	5464	study	2764
these	4907	patients	2716
research	4764	integrative	2708

单词	次数	单词	次数
that	22659	came	4435
with	18065	would	4390
said	15925	before	4084
they	15005	over	3948
will	14292	back	3922
have	14176	down	3900
from	12206	make	3854
your	10771	take	3847
which	10421	could	3836
this	10363	went	3830
them	10213	people	3804
were	9970	little	3691
their	8827	away	3634
there	8616	more	3608
lord	8346	about	3599
what	7424	very	3472
then	7117	time	3446
when	7000	great	3402
into	6456	house	3222
harry	6446	good	3219
been	5833	those	3134
come	5506	king	3089
made	5334	place	3060
give	4953	these	3022
like	4915	after	2996

讨论---文本增加对于单词量的增加



单词合集



合集单词

information
research
analysis
expression
cell
cells
university
author
institute
center
medicine
results
data
binding
specific
that
with
from
were
these
which
using
associated
have
also
lines
medical
role
both
structures
science
national
hospital
different
identified
however
showed
school
three
than
disease
patients
laboratory
development
based
sequence
model
studies
complex
methods
their
other
small
normal
high
more
here
region
found
into
been
study
including
significantly
activity
function
domain
college
university

谢 谢