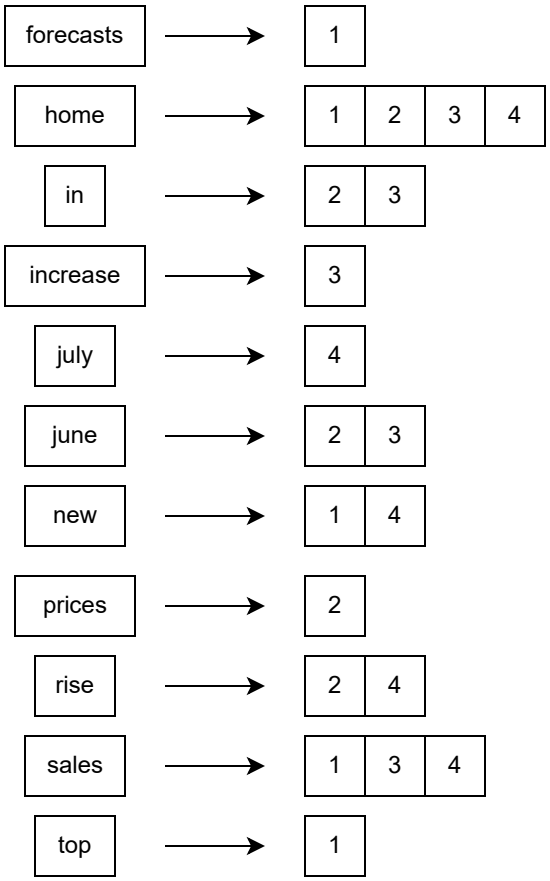
作业一

1

a.

	文档1	文档2	文档3	文档4
forecasts	1	0	0	0
home	1	1	1	1
in	0	1	1	0
increase	0	0	1	0
july	0	0	0	1
june	0	1	1	0
new	1	0	0	1
prices	0	1	0	0
rise	0	1	0	1
sales	1	0	1	1
top	1	0	0	0

b.



c.

a) 文档四

b) 文档三

2

VB编码：
777 : 00000110 10001001
17743 : 00000001 00001010 11001111
294068 : 00010001 01111001 10110100
31251336 : 00001110 01110011 00110111 10001000
777 : 00000110 10001001
16966 : 00000001 00000100 11000110
276325 : 00010000 01101110 11100101
30957268 : 00001110 01100001 00111101 11010100

γ编码：
777 ： 11111111 10100001 001
17743 ： 11111111 11111100 00101010 01111

294068 ： 11111111 11111111 11000011 11100101 10100
31251336 ： 11111111 11111111 11111111 01101110 01101101 11000100 0
777 ： 11111111 10100001 001
16966 ： 11111111 11111100 00010010 00110
276325 ： 11111111 11111111 11000001 10111011 00101
30957268 ： 11111111 11111111 11111111 01101100 00101111 01101010 0

3

a)

不考虑开始，结束符

单词	2-gram集合	2-gram Jaccard系数
bord	bo,or,rd	
border	bo,de,er,or,rd	$\frac{3}{3+5-3}=\frac{3}{5}$
lord	lo,or,rd	$\frac{2}{3+3-2}=\frac{1}{2}$
morbid	bi,id,mo,or,rb	$\frac{1}{3+5-1}=\frac{1}{7}$
sordid	di,id,or,rd,so	$\frac{2}{3+5-2}=\frac{1}{3}$

考虑开始，结束符

单词	2-gram集合	2-gram Jaccard系数
$bord$	b,bo,d,or,rd	
$border$	b,bo,de,er,or,r,rd	$\frac{4}{5+7-4}=\frac{1}{2}$
$lord$	l,d,lo,or,rd	$\frac{3}{5+5-3}=\frac{3}{7}$
$morbid$	m,bi,d,id,mo,or,rb	$\frac{2}{5+7-2}=\frac{1}{5}$
$sordid$	s,d,di,id,or,rd,so	$\frac{3}{5+7-3}=\frac{1}{3}$

b) 首尾添加k-1个首尾标识符，然后查询
$\$_1\$_2\dots\$_{k-1}S_1$ ， $\$_2\$_3\dots\$_{k-1}S_1S_2$ ， $\dots$ ， $S_n\$_{k-1}\$_{k-2}\dots\$_1$

4

a.

	Doc1	Doc2	Doc3
car	44.55	6.6	39.6
auto	6.24	68.64	0
insurance	0	53.46	46.98
best	21	0	25.5

b.

Doc1	0.897369	0.125692	0	0.423002
Doc2	0.0756426	0.786683	0.612705	0
Doc3	0.595268	0	0.706204	0.383317

c.

Doc3 1.30147

Doc1 0.897369

Doc2 0.688348