第四次作业

姓名：刘文晨

学号：202222280328

习题3.1.1

设A={1, 2, 3, 4}，B={2, 3, 5, 7}，C={2, 4, 6}。

则，SIM(A, B)=2/6=1/3，SIM(A, C)=2/5，SIM(B, C)=1/6。

习题3.1.2

设A={1, 1, 1, 2}，B={1, 1, 2, 2, 3}，C={1, 2, 3, 4}。

则，SIM(A, B)=3/9=1/3，SIM(A, C)=2/8=1/4，SIM(B, C)=3/9=1/3。

习题3.2.1

3.2节的第一句话的前10个3-shingle是：{"The", "he ", "e m", " mo", "mos", "ost", "st ", "t e", " ef", "eff" }。

若采用基于词的构造方法，则为：{"The most effective", "most effective way", "effective way to", "way to represent", "to represent documents", "represent documents as", "documents as sets", "as sets for", "sets for the", "for the purpose"}。

习题3.2.2

若采用3.2.4节所示的基于停用词的shingle的表示方法，3.2节的第一句话的shingle有：{"The most effective", "way to represent", "as sets for", "for the purpose", "the purpose of", "of identifying lexically", "is to construct", "to construct from", "the document the", "the set of", "set of short", "of short strings", "it If we", "If we do", "we do so", "do so then", "so then documents", "as short as", "as sentences or", "or even phrases", "in their sets", "if those sentences", "in different orders", "in the two", "the two documents", "two documents In", "In this section", "we introduce the", "the simplest and", "and most common", "as well as", "as an interesting", "an interesting variation"}。

习题3.2.3

长度为n个字节的文档中k-shingle最多有2^(4*n)个。

习题3.3.2

哈希函数：

行	S1	S2	S3	S4	x+1 mod 5	3x+1 mod 5	2x+4 mod 5	3x-1 mod 5
0	1	0	0	1	1	1	4	-1
1	0	0	1	0	2	4	1	2

行	S1	S2	S3	S4	x+1 mod 5	3x+1 mod 5	2x+4 mod 5	3x-1 mod 5
2	0	1	0	1	3	2	3	0
3	1	0	1	1	4	0	0	3
4	0	0	1	0	0	3	2	1

1. 初始化签名矩阵：

	S1	S2	S3	S4
h1	∞	∞	∞	∞
h2	∞	∞	∞	∞
h3	∞	∞	∞	∞
h4	∞	∞	∞	∞

2. 考虑第0行：

	S1	S2	S3	S4
h1	1	∞	∞	1
h2	1	∞	∞	1
h3	4	∞	∞	4
h4	-1	∞	∞	-1

3. 考虑第1行：

	S1	S2	S3	S4
h1	1	∞	2	1
h2	1	∞	4	1
h3	4	∞	1	4
h4	-1	∞	2	-1

4. 考虑第2行：

	S1	S2	S3	S4
h1	1	3	2	1
h2	1	2	4	1
h3	4	3	1	3
h4	-1	0	2	-1

5. 考虑第3行：

	S1	S2	S3	S4
h1	1	3	2	1
h2	0	2	0	0
h3	0	3	0	0
h4	-1	0	2	-1

6. 考虑第4行：

	S1	S2	S3	S4
h1	1	3	0	1
h2	0	2	0	0
h3	0	3	0	0
h4	-1	0	1	-1

最终得到的最小哈希签名矩阵为：

	S1	S2	S3	S4
h1	1	3	0	1
h2	0	2	0	0
h3	0	3	0	0
h4	-1	0	1	-1

习题3.3.3

(a)

哈希函数：

行	S1	S2	S3	S4	2x+1 mod 6	3x+2 mod 6	5x+2 mod 6
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

1. 初始化签名矩阵：

	S1	S2	S3	S4
h1	∞	∞	∞	∞
h2	∞	∞	∞	∞
h3	∞	∞	∞	∞

2. 考虑第0行：

	S1	S2	S3	S4
h1	∞	1	∞	1
h2	∞	2	∞	2
h3	∞	2	∞	2

3. 考虑第1行：

	S1	S2	S3	S4
h1	∞	1	∞	1
h2	∞	2	∞	2
h3	∞	1	∞	2

4. 考虑第2行：

	S1	S2	S3	S4
h1	5	1	∞	1
h2	2	2	∞	2
h3	0	1	∞	0

5. 考虑第3行：

	S1	S2	S3	S4
h1	5	1	1	1
h2	2	2	5	2
h3	0	1	5	0

6. 考虑第4行：

	S1	S2	S3	S4
h1	5	1	1	1
h2	2	2	2	2
h3	0	1	4	0

7. 考虑第5行：

	S1	S2	S3	S4
h1	5	1	1	1
h2	2	2	2	2
h3	0	1	4	0

最终得到的最小哈希签名矩阵为：

	S1	S2	S3	S4
h1	5	1	1	1
h2	2	2	2	2
h3	0	1	4	0

(b)

这些哈希函数中只有h3是真正的排列转换。

(c)

Jaccard相似度	S1-S2	S1-S3	S1-S4	S2-S3	S2-S4	S3-S4
估计值	1/3	1/3	2/3	2/3	2/3	2/3
真实值	0	0	1/4	0	1/4	1/4

Jaccard相似度的估计值和真实值差距很大。