**4.1** 给出文档“to be or not be that is a question”的4-shingle集合。

{to be or not, be or not be, or not be that, not be that is, be that is a, that is a question}

**4.2**采集器中，为何同时有URL去重和内容去重？试讨论二者关系及各自所起的作用。

URL去重是在采集时查看URL是否有已经被采集或者已经URL池，防止重复采集同一个页面

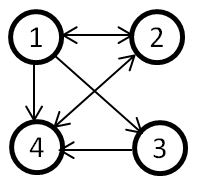
内容去重是为了做索引排序，防止搜索引擎索引重复的网页，降低存储和处理开销。

两者的作用不同，所以需要对每一个不同的URL页面进行内容去重。

而即使URL相同，内容也可能随着时间的流逝而发生变化，即使URL不同，也有可能有相同的内容，所以对二者先后去重是有必要的。

回答合理即可。

**4.3** 对于如下web图：



a. 取α=0.3，求其转移概率矩阵。

b. 求按课本公式21-8或ppt35页的公式迭代两次之后的hub值向量和authority值向量。

0.075 0.308 0.308 0.308

0.425 0.075 0.075 0.425

0.075 0.075 0.075 0.775

0.075 0.775 0.075 0.075

（2）

邻接矩阵为：

0 1 1 1

1 0 0 1

0 0 0 1

0 1 0 0

H = [6 4 3 2]

A = [2 4 3 6]