4.1. 给出文档“no need to light a night light on a light night”的4-shingle集合

- no-need-to-light
- need-to-light-a
- to-light-a-night
- light-a-night-light
- a-night-light-on
- night-light-on-a
- light-on-a-light
- on-a-light-night

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：

- 避免已经采集的网页被重复采集，降低效率

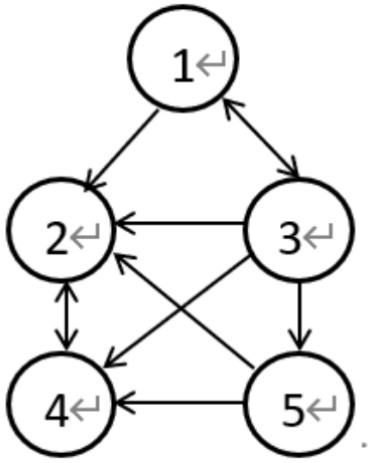
内容去重：

- 避免大量重复的内容被重复采集

关系：

- URL去重是一种显而易见的避免重复采集、提高效率的方法。然而实际上的情况是，会有许多分布在不同URL上的网页充斥着重复的内容，采集这些网页不仅效率低还降低了用户体验，因此需要二者结合以实现更好的去重效果

4.3. 对于如下web图



a). 取 α=0.3，求其转移概率矩阵。

$$L=\begin{pmatrix}0&1&1&0&0\\0&0&0&1&0\\1&1&0&1&1\\0&1&0&0&0\\0&1&0&1&0\end{pmatrix}$$

$$N(L)=\begin{pmatrix}0&0.5&0.5&0&0\\0&0&0&1&0\\0.25&0.25&0&0.25&0.25\\0&1&0&0&0\\0&0.5&0&0.5&0\end{pmatrix}$$

$$P=0.7N(L)+0.06=\begin{pmatrix}0.06&0.41&0.41&0.06&0.06\\0.06&0.06&0.06&0.76&0.06\\0.235&0.235&0.06&0.235&0.235\\0.06&0.76&0.06&0.06&0.06\\0.06&0.41&0.06&0.41&0.06\end{pmatrix}$$

b). 求按ppt35页的公式迭代两次之后的hub值向量和authority值向量。

$$A=\begin{pmatrix}0&1&1&0&0\\0&0&0&1&0\\1&1&0&1&1\\0&1&0&0&0\\0&1&0&1&0\end{pmatrix}$$

初始化：

$$\vec{a}=(1,1,1,1,1)$$
$$\vec{h}=(1,1,1,1,1)$$

第一次迭代：

$$\vec{a}=(1,4,1,3,1)$$
$$\vec{h}=(5,3,9,4,7)$$

Normalize:

$$\vec{a}=(0.189,0.756,0.189,0.567,0.189)$$
$$\vec{h}=(0.373,0.224,0.671,0.298,0.522)$$

第二次迭代：

$$\vec{a}=(0.671,1.864,0.373,1.417,0.671)$$
$$\vec{h}=(2.237,1.417,4.623,1.864,3.281)$$

Normalize:

$$\vec{a} = (0.263, 0.730, 0.146, 0.555, 0.263)$$
$$\vec{h} = (0.343, 0.217, 0.708, 0.286, 0.503)$$