1. 统一资源标识符，统一资源定位器

HTTP超文本传输协议

HTML

1. A
2. 网页结构化是为了让用户更顺利地、分层次地获取网页中的信息。
3. 连通图，结点，边

挥发性，半结构性，隐蔽性

1. 网页抓取使用的是宽度优先策略，因为它①可以保证对浅层的，通常更重要的网页首先处理，②不会陷入WWW中的无穷尽的深层文档中出不来，③可以在两个HTML文件之间找到最短路径，④有利于多爬虫合作抓取，抓取的封闭性较强。
2. MD5签名函数是一个散列函数，它可以将任意长度的数据流转换为一个固定长度的数字(通常为4个整型数，即128位)。抓取网页时，它用来记录网页访问历史：如果某网页在过去的某个时刻已经被抓取，则将其对应的槽位（哈希表中）的值置为1，反之则为0。
3. 参考网页：<https://blog.csdn.net/qq_34333481/article/details/84024513>

该博客中给出了Jaccard的定义、公式、两个例子以及python实现源码。

例子1：对两集合A = {1, 2, 3, 4}, B = {3, 4, 5, 6};

有A ∩ B = {3, 4}, A ∪ B = {1, 2, 3, 4, 5, 6};

则Jaccard(X, Y) = |{3, 4}|/ |{1, 2, 3, 4, 5, 6}| = 1/3

例子2：假设手机这一物品有3个特征：分别是①是否国产；②是否有“刘海”（指屏幕上方大块的摄像头区域）；③价格是否大于￥5000。所以可以用3维的向量表示手机A：坚果pro2和B：iPhone8的特征。分别为：A{国产，无刘海，否} 和 B{非国产，有刘海，是}，则二者Jaccard系数为0.

文中进一步推广，表示特征的向量中“肯定/否定”可以用“1/0”来表示，则

Jaccard = p / (p + q + r)，其中

P：样本A与B都是1的维度的个数

q：样本A是1，样本B是0的维度的个数

r：样本A是0，样本B是1的维度的个数

s：样本A与B都是0的维度的个数

另外，原网页中的python实现使用的是矩阵，只能将样本特征先转换为1/0表示再计算Jaccard系数；我将其改为集合，可以有更直观的体现：

