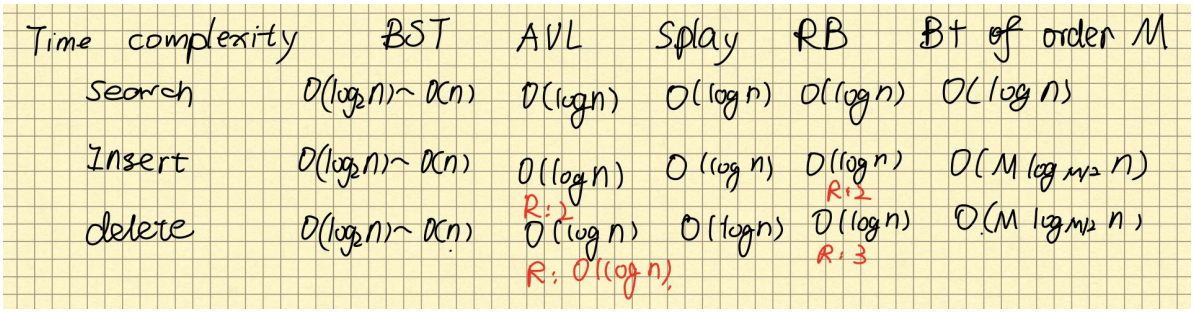
lesson1-2-3整理



AVL vs RB

AVL: search的时候更快，因为更加平衡一些

RB：insert 和delete更快，因为AVL计算平衡因子比较慢

为什么需要NIL?NIL为什么是黑色的?

定义的需要：每一个红色节点的children都是黑的；到 leaf 的 length 都是一样的；

eg:n个节点的红黑树，连续做m次insertion，时间复杂度是什么?

O(M Lg n) ? 太大了 应该是 O(n+m)

m:大部分都是O(1)操作，只有爹爹和叔叔都是红色，才会向上递归

n: ???

inverted file index

download document: hash表记录是否下载；BFS，有限时间内先下载重要的东西；调度系统，优先级序列；

index：布尔运算，term:30万，to be or not to be 不能删啊；

rank：PageRank算法，现在的算法计入用户，经济的影响；

PageRank算法：根据链接到它的网站数量排名

如何确定网页的相关性？TF-IDF

IDF=Log(D/Dw)

D:所有文件；Dw:w个包括 term 文件；

相关性：TF1 * IDF1 + TF2 * IDF2