给定a、b两个文件，各存放50亿个url，每个url各占64字节，内存限制是4G，让你找出a、b文件共同的url。

50亿*64bytes=320G，远远大于4G内存。

思想是采用分治算法
step1：遍历文件a，对每个url求取hash(url)%1000，然后根据所取得的值将url分别存储到1000个文件(记为a0,a1,...,a999，每个小文件约为300M)，为什么是1000？主要根据内存大小和分治的文件大小来计算，我们就大致可以把320G大小分为1000份，每份大约300M（到底能不能分布尽量均匀，得看hash函数得设计。）
step2：遍历文件b，采取和a相同的方式将url分别存储到1000个小文件，文件a的hash映射和文件b的hash映射函数要保持一致，这样的话相同的url就会保存到对应的小文件中。

然后现在的问题转换为：找出1000对小文件中每一对相同的url（不对应的小文件不可能有相同的url）

step3：然后我们对每一对小文件，先读取a的小文件，建立hash表，然后再读b的小文件，遍历b中每个url，对于每个遍历，我们都执行查找hash表的操作。若hash表中搜索到了，则说明两文件共有，存入一个集合。