海量数据topk问题

笔记本:	机器学习		
创建时间:	2019/6/20 18:15	**更新时间:**	2020/5/30 11:25
作者:	beyourselfwb@163.com		
标签:	面试经验		

类别一：top k个数值最大的数
类别二：top k个出现频次最大的数

参考： https://www.jianshu.com/p/cea0deb449e1
https://www.cnblogs.com/xudong-bupt/archive/2013/03/20/2971262.html
https://gist.github.com/wbbeyourself/6e8d47b902e0b252f48b0120cc0b646f

总体思路：分治+hashmap+最小堆(小顶堆)

第一步，都是按照数字/词语hashcode将大文件拆分成小文件

第二步：用HashMap统计词频, 最小堆对每个文件求topk（多线程优化）

第三步：合并top k, 再来一次最小堆求top k

最优的解决方案应该是最符合实际设计需求的，进行时间或空间的取舍：
参考： http://doc.okbase.net/zyq522376829/archive/169290.html