# 3. 有一个 1G 大小的一个文件，里面每一行是一个词，词的大小不超过 16字节，内存限制大小是 1M。返回频数最高的 100 个词。

方案 1：顺序读文件中，对于每个词 x，取 ，然后按照该值存到 5000 个小文件（记为 ）中。这样每个文件大概是 200k 左右。如 果其中的有的文件超过了 1M 大小，还可以按照类似的方法继续往下分，知 道分解得到的小文件的大小都不超过 1M。对每个小文件，统计每个文件中出 现的词以及相应的频率（可以采用 trie 树/hash\_map 等），并取出出现频率 最大的 100 个词（可以用含 100 个结点的最小堆），并把 100 词及相应的频

率存入文件，这样又得到了 5000 个文件。下一步就是把这 5000 个文件进行 归并（类似与归并排序）的过程了。