# 第十章 人工智能

## 一．数据挖掘

#### **1G 的文件，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。**

使用生成器读取文件。每次读取 65536 行，一共进行 1500 次，当读取不到内容时关闭文件。每

次读取，最终要得到 100 个频数最高的词。每 500 次，进行一次合并和统计，得到最多 50000 个

词，对这 50000 个词提取其中频数最高的 100 个词。最终对最多 300 个筛选出来的词，进行合并

和统计，提取最终频数最高的 100 个词。

筛选出 100 个高频词的步骤：

1、统计每个词出现的次数。维护一个 Key 为词，Value 为该词出现次数的 hash 表。每次读取一个词，如果该字串不在 hash 表中，那么加入该词，并且将 Value 值设为 1；如果该字串在 hash 表中，那么将该字串的计数加一即可。

2、据每个词的引用次数进行排序。冒泡、快排等等都可以。

#### **一个大约有一万行的文本文件，每行一个词，要求统计出其中最频繁出现的前 10 个词，请给出思想和时间复杂度分析。**

用 trie 树统计每个词出现的次数，时间复杂度是 O(n\*le)(le 表示单词的平均长度)。然后是找出出现最频繁的前 10 个词，可以用堆来实现，时间复杂度是 O(n\*lg10)。所以总的时间复杂度，是 O(n\*le) 与 O(n\*lg10)中较大的哪一个。

详情扫描二维码了解。<http://www.mamicode.com/info-detail-1037262.html>

|  |
| --- |
|  |
|  |  |

#### **怎么在海量数据中找出重复次数最多的一个？**

算法思想：

先做 hash，然后求模映射为小文件，求出每个小文件中重复次数最多的一个，并记录重复次数。然后找出上一步求出的数据中重复次数最多的一个就是所求。

参考博客一：



<https://blog.csdn.net/u010601183/article/details/56481868>

|  |
| --- |
|  |
|  |  |

参考博客二：

<https://www.cnblogs.com/lianghe01/p/4391804.html>

|  |
| --- |
|  |
|  |  |

#### **给 2 亿个不重复的整数，没排过序的，然后再给一个数，如何快速判断这个数是否在那 2 亿个数当中？**

unsigned int 的取值范围是 0 到 2^32-1。我们可以申请连续的 2^32/8=512M 的内存，用每一个 bit 对应一个 unsigned int 数字。首先将 512M 内存都初始化为 0，然后每处理一个数字就将其对应的 bit 设置为 1。当需要查询时，直接找到对应 bit，看其值是 0 还是 1 即可。

# 

# 二．机器学习

# 三．深度学习