数据与算法课程实验

实验十 最佳英文排版问题

1.引言：

小白喜欢写英文诗，他写诗很有自己的风格：不用标点符号，而且特别抽象，很难进行理的断句和分段。

有一天小白又有新作品了，他觉得如果能把这诗分成两段，效果肯定棒，但是他很犹豫，不知道该从哪里断开好。于是他只好求助于同事小陈。

小陈当然是读不懂小白的诗了，不过他的思维模式是计算机的（据说他是计算机系毕业的），所以他打算用一种计算机范畴内的“最佳排版”策略来进行排版和分段。“诗嘛，就是要看起来漂亮！”小陈说。

在我们考虑的范围内，英文单词是不能断开的，也就是说一个单词不能分别处于两个行的尾和头（有一种做法是对断开的单词的断开处加横杠以示区别，但我们不采取这种做法，否则就不需要考虑“最佳排版”了）。

2.实验内容：

给定一篇包含 N 个单词的诗，每个单词的物理长度依次为 a[0],a[1],...,a[N-1]。欲将诗分成两个段落，每个段落至少有一个单词，每个段落的排版应遵循以下规则：

每行可容纳的总字符数为 M；

每个单词必须在同一行，不得断开换行；

同一行的每两个单词之间有且仅有一个空格，不考虑其他标点符号；

采用左对齐的方式排版，即每行的第一个单词排在该行最左侧；

假设排版之后共有 m 行，每行末端没有英文单词的空档长度记作 s[0],s[1],...,s[m-1]。

不考虑最后一行的空挡大小 s[m-1]，设计尽可能高效的算法给段落排版，使得

$$P=s[0]*s[0]+s[1]*s[1]+...+s[m-2]*s[m-2]$$

最小。当 m=1 时，P 等于 0.

假设第一段的最优的 P 值为 P1，第二段的最优的 P 值为 P2，要求令 **P1+P2** 最小。

3.输入：

第一行输入两个整数 N 和 M，从第二行开始，输入 N 个整数，代表 a[0]到 a[N-1]。

数据范围：10<=N<=500000, 10<=M<=20000, 1<=a[i]<=M

4.输出：

输出三个整数，第一个整数是最优的 P1+P2 值，第二个整数是第一段落的最后一行的首个单词的索引值，第三个整数是第二段落的第一行的首个单词的索引值。当存在多种可行的解答时，采取这么一种策略：假设输出为 P_total，a，b，当存在多个可行的 a 时，选择最小的 a；a 确定之后，如果仍然存在多种可行的 b，就选择最小的 b。

5.样例：

输入：

10 50

3 1 3 3 6 2 1 3 2 29

输出：

0 0 4

解释：由于该诗可以放入两行里，所以每一行各成一段就行了，此时 P_total 等于 0（末行的空档不计入 P 中）。有多种可行的解，显然 a 应当等于 0，这样之后仍然有多种可能的 b，选择最小的 b=4；如果 b<4，第二行就放不下所有单词了，而且此时不存在摆放三行而使 P_total 等于 0 的方法。

6.提示：

用动态规划的方法，分别得到两个数组 optiValuePre[N]和 optiValuePost[N]，含义如下：

optiValuePre[i]表示以第 i 个单词为第一段的尾行首单词时，P1 的值。

optiValuePost[i]表示以第 i 个单词为第二段的头行首单词时，P2 的值。

然后就可以利用这两个数组，来寻找最佳的分段位置了。这种算法用了三轮动态规划，每一轮的复杂度相似。

积极开动脑筋，也许你能得到更好的解法。

假如你只在大样例上得到了错误结果，请考虑目标函数 P 的数据范围。

7.评分标准：

- 使用 C 或 C++实现
- 共有 5 个测试样例，难度递增，每个 20 分
- 拒绝抄袭

参考文献：

[1]. 何宗林，*基于动态规划策略的英文文档排版算法*