13. 正则表达式 课后作业

1）拯救特工 （20 分）

隐私保护十分重要，尤其是在情报机构，内部文件不可直接将情报人员的名字直接写出来。请你将使用*替代原本特工的名字，假设：

msg = '特工张大三让特工李小四把图纸 USB 交给特工王二麻。'

输出效果如下：

张**让李**把图纸USB交给王**。

2）文本统计 （40 分）

频率计数广泛应用于从海量的数据中统计各种事件出现的频率。如，语言学家从文章中发现单词的使用模式、商人从订单中发现重要的客户等。

请从“**Pumas.txt**”中，统计文本包含的段落数（回车结束）、句数（.?!结束）、单词数，统计单词出现的频率，并输出前 10 个单词。**需使用正则表达式进行分隔**。统计词频时，可以自己编写代码，也可以使用 collections.Counter()。输出效果如下：

```
段落数： 2
句数： 13
单词数： 276
频率最高的10个单词：
a          :         14
the        :         14
puma       :          9
in         :          8
to         :          6
it         :          6
and        :          6
of         :          5
that       :          4
were       :          4
```

3）基因预测 （40 分）

基因是生命的本质，生物学家用字母 A、C、T、G 分别代表生物体 DNA 的 4 个碱基。基因由一序列的密码子组成，每个密码子是一个由一系列代表氨基酸的 3 个碱基组成的序列。判断某字符串是否对应一个**潜在的基因**准则：

- 基因长度为 3 的倍数。
- 以 ATG 标识基因的开始。
- 以 TAG、TAA、TGA 标识基因的结束。
- 除结束部分，中间部分不包含：TAG、TAA、TGA。

判断“**gene.txt**”文件中哪些是潜在的基因序列，后 3 条**需使用正则表达式判断**。运行效果：

```
1:ATGCGCCTGCGTCTGTAG
2:ATGCTGCGCCGTCTGTAA
3:ATGCGCCTGCGTCTGTGA
```