**广东外语外贸大学信息科学与技术学院**

**《文本信息处理工程实践》课程作业**

**第2次作业：编写并测试正向最大分词类**

姓名: \_\_\_高子雄、游畅\_\_\_\_\_\_\_\_\_\_\_

学号： \_20181002989、20181003005\_

所在班级：\_\_\_软工1804\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **作业算法描述**

本次实验的目的为构建一个正向最大分词的Python类，类内包含数据成员：

indexpos、result、MaxLen、Len、Dict；

类内包含方法：

MM(sentence, Len, indexpos)、

GetResult()。

该实验的本质内容是编写具有正向最大分词的类方法MM。下面即对正向最大分词算法进行简要的概括描述：

正向最大分词算法属于机械分词的一种。其基于字符串匹配的原理，在预定义的词典中查询给定sentence可能的分词方式。

（1）、预定义词典。txt文件chineseDict中给出了若干汉语词汇以及其词性，格式为(词汇,词性)。我们使用Python的Dict型数据结构将其记录：词汇为字典的键，词性则为键值；

（2）、正向最大分词的核心思想。正向，即自左向右，也就是汉语的书写与阅读方向。在预定义的词典中做正向最大匹配意味着默认地将尽量长的分词方式判断为正确的分词方式。例如，MM能够识别“掩耳盗铃”而不是“掩耳”与“盗铃”。在算法中，我们可以将这种思想程序式地表达出来。

（3）、正向最大分词算法参数设置。首先我们定义指针indexpos，这表示当前指针指在sentence中的位置。然后定义最大长度Len，最大长度表示下一步所截取的待匹配的字符串的长度。一般而论，为了保证能够识别出词典里最长的词，Len取值应与词典最长词相同。但是在大规模的任务中就会造成可观的多余开销，因此本次实验中，Len的取值为常见的成语长度4。

（4）、具体流程。

①、利用指针和Len对句子切分

②、若Len==1，说明已达到最短长度。添加str[Len]进result，指针位置+1

③、在字典里查找切分片段

④、若查找成功，将该片段添加进result，指针跳跃至(indelpos+Len)处

⑤、若查找失败，则Len-1，回到①

**2.实现要点**

（1）、边界问题。

在上一部分的步骤表示中，并未完整说明对边界情况的处理。在此进行详细的阐述。

本次实验中的边际问题主要有两处：①当一直无法在词典中得到匹配结果，Len减少至1时；②当指针来到sentence的后半部分，sentence所剩余的部分小于Len时；

情况①表示的是待匹配的str是一个单字，例如“你”、“我”、“的”、“在”等词。这种情况下，无需再进行匹配，将str[indexpos]分词，即可。

情况②表示的是指针所指位置到结尾的距离已小于切分长度Len。这种情况下，待匹配的字符串即是str[indexpos, len(str)-1]。将其放入字典匹配的同时，Len需要修改为len(str[indexpos, len(str)-1])，即len(待匹配串长)。

（2）、递归调用。

递归调用始终是函数的一大难点。在处理递归的过程中，需要注意的地方一是调用的时机（位置），二是传入参数的变与不变。

在这次实验里，需要递归的地方只有1处（逻辑上），实际却有3个地方：

①：若查找成功，将该片段添加进result，指针跳跃至(indelpos+Len)时，递归一次。indexpos= indelpos+Len；Len恢复为初始长度（即manLen）；str不改变。

②：查找失败时，若Len已是1，分词后进行递归。indexpos= indelpos+1（等同上一种情况的+Len）；Len恢复为初始长度（即manLen）；str不改变。

③：查找失败，Len大于一时，Len=Len-1，指针位置不变，str不变，进行递归。

**3.程序演示**

例1：“去年今日此门中，人面桃花相映红。”



图（1）：例1结果

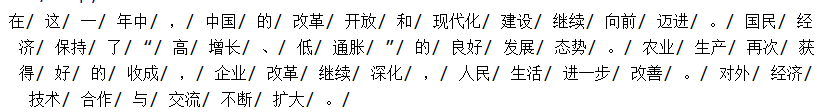
结果应为/人面桃花/。人面桃花是一个成语。这里切分失败的原因在于词典中不包含这个词。

例2：“江畔何人初见月， 江月何年初照人。”



图（2）：例2结果

结果应为江月/何年/初/照/人。



图（3）：例3结果

应为“在/这/一年/中”。错误原因同样是词典中没有“一年”这个词。

**4.总结和体会**

起初没有事先检查词典，想了很久不知如何处理标点符号（用split去除后如何回复原位呢？）在查看词典后才反应过来标点符号也是分词的一部分，并无需要split去除。

另一点则是str[a:b]的闭开区间问题。Python中此用法表示的是左闭右开，表达了str[a]到str[b-1]字符串。