汉语分词:最大匹配方法

参考 李正华讲义

分词任务

中文分词的目的是将汉字序列切分为词序列

举例说明：

输入句子：他是研究生物化学的。

可能的分词：他 是 研究生 物化 学 的 。

他 是 研究生 物 化学 的 。

他 是 研究 生物 化学 的 。

合理答案：他 是 研究 生物 化学 的 。

正向最大匹配算法

从左到右寻找词的最大匹配（每次都贪心的找一个最长的词典词）

我们有一个词典，用于存放所有可能的词语，即除了单字，分词结果中的每个词均要在词典中出现。

正向最大匹配算法

从左到右寻找词的最大匹配

从当前位置开始，向右截取最大长度，组成当前词；

和字典中的词逐一进行匹配；

若匹配成功，则进行下次匹配，下次匹配的当前位置则为这次词后面的那个字。

如果未能匹配，就缩短长度（长度减一）重新截取，直到当前词与词典中的词匹配或者当前词是单字；

正向最大匹配算法

从左到右寻找词的最大匹配

初始化： 指针 p1 指向句子的首位置，词的最大长度为 m

算法执行：

(1) 如果 p1 到达句子末尾，分词结束；

(2) 假设当前判断的词的长度是 i，初始化为 m；

(3) p2 = p1 + i；如果 p2 超过句子的末尾，则 i--，直到 p2 到达句子末尾之前；

(4) 如果 p1 和 p2 之间的字符串 S'在词表中不存在，i--，重复(3)；

(5) 如果 p1 和 p2 之间的字符串 S'在词表中存在，则 S'是一个词，p1 = p2+1，转(1)；

正向最大匹配算法

例子：我是中国人

词典中包括【中国、中国人】

假设：最大词长为3

正向最大匹配算法

例子：我是中国人

第一轮：

第一次："我是中"是选取的词，在词典中未找到匹配项

第二次："我是"是选取的词，在词典中未找到匹配项

第三次："我"是选取的词，是单字，匹配成功

正向最大匹配算法

例子：我/是中国人

第二轮：

第一次："是中国"是选取的词，在词典中未找到匹配项

第二次："是中"是选取的词，在词典中未找到匹配项

第三次："是"是选取的词，是单字，匹配成功

正向最大匹配算法

例子：我/是/中国人/

第三轮：

第一次："中国人"是选取的词，在词典中找到匹配项，匹配成功。

至此，短句中所有字匹配结束，该短句分词结束。

逆向最大匹配算法

从右到左寻找词的最大匹配

与正向最大匹配的区别在于，从句子的末尾开始，向左边截取一定的长度去匹配。

逆向最大匹配算法

从右到左寻找词的最大匹配

初始化：指针 p1 指向句子的尾部位置，词的最大长度为 m

算法执行：

(1) 如果 p1 到达句子首位置，分词结束；

(2) 假设当前判断的词的长度是 i，初始化为 m；

(3) p2 = p1 - i；如果 p2 到达句子的首部，则 i--，直到 p2 到达句子首部之前；

(4) 如果 p2 和 p1 之间的字符串 S'在词表中不存在，i--，重复(3)；

(5) 如果 p2 和 p1 之间的字符串 S'在词表中存在，则 S'是一个词，p1 = p2-1，转(1)；

逆向最大匹配算法

例子：我是中国人

第一轮：

第一次："中国人"是选取的词，在词典中找到匹配项，匹配成功

逆向最大匹配算法

例子：我是/中国人

第二轮：

第一次：因为剩余字数已不足3，小于假定的最大词长，所以选择"我是"，在词典中未找到匹配项

第二次："是"是选取的词，是单字，匹配成功

逆向最大匹配算法

例子：我/是/中国人

第三轮：

第一次：因为剩余字数已不足3，小于假定的最大词长，所以选择"我"，是单字，匹配成功

至此，短句中所有字匹配结束，该短句分词结束。

分词算法评价：正确率/召回率/F值

给定人工标注的分词答案，评价某一算法给出的结果。

正确率(Precision) = 正确识别的词数 / 识别出的个体总数

召回率(Recall) = 正确识别的个体总数 / 测试集中存在的个体总数

F值 = 正确率* 召回率 * 2 / (正确率 + 召回率)

思考：评价程序应该怎么写？

UTF-8编码

utf-8是不定长的，根据左侧位1的个数来决定占用了几个字节，中文一般占2-4个字节

utf-8可以根据字的第一个字节移位推出长度的
0xxxxxxx占1个字节
110xxxxx 10xxxxxx占2个字节
1110xxxx 10xxxxxx 10xxxxxx占3个字节
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx占4个字节

GBK编码

GBK编码方式中文占两个字节，英文占一个字节,根据第一个字节的最高位来判断

如果第一个字节的最高位是1，则是两个字节连在一起为一个字符，否则一个字节为一个字符

中文的编码范围

第一个字节	第二个字节
0x81-0xFE(129-254)	0x40-0xFE(64-254)