|  |  |
| --- | --- |
| **自然语言处理及应用**  **实验报告** | |
|  | |
|  |  |
| **名称** | 中文分词和词性标注 |
| **姓名** |  |
| **班级** |  |
| **学号** |  |
| Email |  |
| **日期** | 2021-10-2 |

# 实验目的

(1） 实现基于HMM的分词方法和基于最短路径的分词方法。

（2） 实现基于HMM的词性标注器。

（3） 对分词结果、词性标注结果进行评价。

(4)用户可以自己输入句子，然后对句子先进行分词，在对分词结果标注，最终输出标注结果。

# 实验环境

Python语言，使用的Spyder编译器。

Json模块、tqdm模块。Json模块用来将实验中产生的矩阵以json格式存储，之后可以将参数保存到本地。Tqdm模块是一个可扩展的Python进度条，在进行评估时使用，用户可以查看当前模型分词和标注的进度。

# 实验方法

Hmm模型可以看成一个五元组，包括状态值集合、观测值集合、转移概率矩阵、发射概率矩阵、初始状态概率。其中三个概率是最为重要的。

**(1)基于HMM的中文分词方法：**

使用BMES标注方法，将分词任务转换为字标注的问题，通过对每个字进行标注得到词语的划分。用“B、M、E、S”四种标签对词语中不同位置的字符进行标注，B表示一个词的词首位置，M表示一个词的中间位置，E表示一个词的末尾位置，S表示一个单独的字词。 字标注的问题可视为隐马尔可夫模型中的解码问题。句子的BMES标注序列作为隐藏状态序列，句子的字符序列作为可观测序列。通过对数据集学习得到隐藏状态的转移概率矩阵trans、发射概率矩阵emit 、初始状态矩阵start。

使用 Viterbi 算法预测，用动态规划的方法求解最优的标注序列。先初始化，然后递推记录路径，在通过回溯路径产生最优的路径。

**(2)基于最短路径的在中文分词方法：**

将句子中的所有字符当作节点，根据字典找出句子中所有的词语，将词语两端的字符连接起来，构成从词首指向词尾的一条边。通过找出所有的候选词，构建出一个有向无环图（DAG）。找到从句首字符到句尾字符的最短路径，即可作为句子的分词结果。

**(3)基于HMM的词性标注方法：**

词性标注是序列标注问题，将词性序列作为隐藏序列，将词语序列作为观测序列，通过Viterbi算法预测最优的词性序列。初始状态概率、发射概率、转移概率都是从数据集中学习得到，状态值集合为26各基本词类标记，

使用的是北大计算语言学的分词资料，包括形容词a、区别词b、连词c、副词d、叹词e、方位词f、语素g、前接成分h、成语i、简称j、后接成分k、习惯用语l、数词m、名词n、拟声词o、介词p、量词q、代词r、处所词s、时间词t、助词u、动词v、标点符号w、非语素字x、语气词y、状态词z、

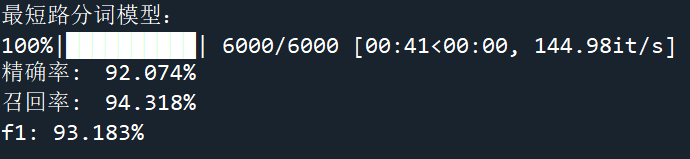
1. **实验结果**

**(1)分词的结果评估：**

取人民日报数据集的前六千行进行分词，将分词结果与正确结果进行对比，产生精确率、召回率、f1-score等评价指标。第一个为基于Hmm的中文分词模型，精确率为60%，召回率为73%，基本可以完成分词任务。



第二个是基于最短路径的分词模型，精确率为92%，召回率为94%，能很好完成分词任务。

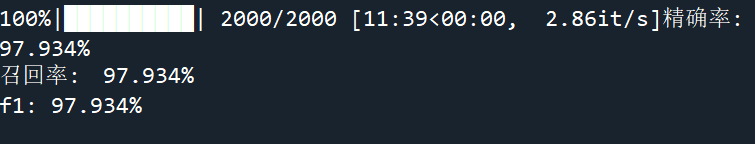


分析：基于最短路径的分词模型测试效率要优于基于hmm的模型。这是因为基于最短路径方法是贪心法，他会找到路径最短的结果，而Hmm模型是基于统计方法，他需要大规模数据来增加准确率，当数据集较少时，准确率就不是很理想。

另外Hmm模型实现复杂、计算量较大，本实验在进行中文分词时，基于最短路径方法明显由于基于Hmm方法。

**(2)词性标注评估：**

训练好模型之后，对人民日报集合前两千行进行词性标注，将标注结果与正确结果进行对比，得到测试结果。精确率为97.9%，召回率为97.9%，基于hmm的词性标注的效果很好，已经能很好完成标注任务。



**(3)用户输入句子，进行分词和标注展示：**

使用最短路径对句子进行分词，使用hmm分词器对句子进行分词，之后使用基于hmm的词性标注器进行标注，输出结果。结果发现，当分词结果准确时，词性标注结果更加准确。

