# 词性标注程序说明

#### 小组成员：

杨欣松，熊炜伟，张冠群

#### 任务定义：

Part-of-speech tagging: 20 points

This data set contains one month of Chinese daily which are segmented and POS tagged under Peking Univ. standard.

Project ideas:

Design a sequence learning method to predicate a POS tags for each word in sentences.

Use 80% data for model training and other 20% for testing (or 5-fold cross validation to test learner's performance. So it could be interesting to separate dataset.)

#### 输入输出：

输入：

1998-01-105-带音.txt

输出：

preTran.txt //前80%

pretest.txt //后20%

train.txt //预处理后的训练文本

test.txt //预处理后的测试文本

answer.txt //标记答案

A.txt //词性转移概率

B.txt //词性概率

result.txt //词性标注结果

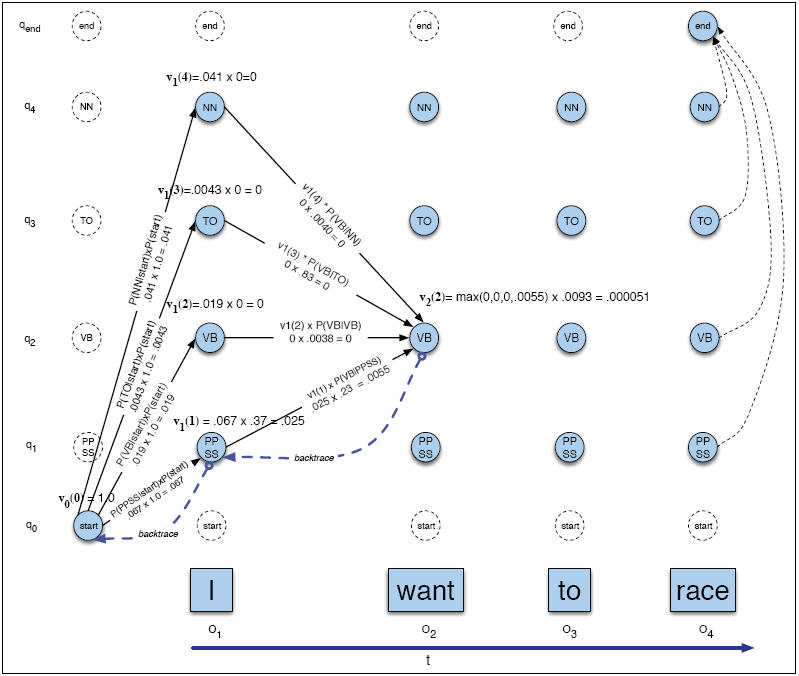
analyse.txt //结果分析

#### 方法描述：

词性标注实质是隐马尔可夫模型的解码问题，即给定观测序列O=(o1o2…oT)，及HMM模型参数λ= (A,B)，如何计算最优的状态序列Q=(q1q2…qT)（i.e., 能最好解释观测O的状态序列）。

使用viterbi算法，计算词性序列的最优路径。





**细节问题：**

A,B若用矩阵则过于稀疏，为提高效率改用MAP存储。为了记录词性序列路径，每个V节点需记录前项节点的词性。

若碰到B中没有的词，则考虑所有词性，将该词标记为V概率最大时的词性。

#### 结果分析：

标记总数：209279

正确个数：186286

正确率：**0.890132**

标记错误可能的原因有**：**

1. 测试文本中的某些词在训练文本中没有出现，处理方式是将该词标记为所有词性中使该节点概率最大的词性，而这个词性很可能并不是该词的词性之一。
2. 概率最大的词性序列并不一定是正确的。

#### 源码运行环境：

VS2013 C++

词性标注.cpp

词性标注.exe