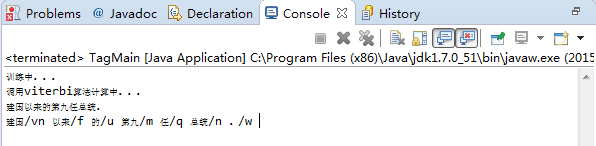
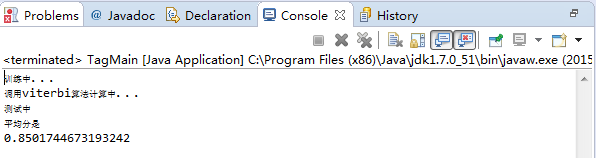
汉语词性标注任务

1. 语料库：北大新闻分词词性标注语料（人民日报1998年1月）
2. 方法：采用一阶隐马尔可夫模型计算参数，采用维特比算法解码.
   1. 训练
      1. 数据预处理：
         1. 语料中，每一行前面有诸如“19980101-01-001-016/m”的日期及其词性。但在训练中这个信息不需要，因此将其去除；
         2. 语料中存在诸如“[人民/n 大会堂/n]ns”的词，需要将“[”和“]ns”去除。否则会产生“[人民”这样的词，和“n]ns”这样的词性。
         3. 语料中，大部分情况下词-词性对之间是用单个空格分隔的，但是也存在用多个空格分隔的情况。需要统一分隔符。
         4. 语料中存在空行，需要去除。
      2. 计算状态转移矩阵的值：
         1. 采用一阶马尔科夫模型，即假设当前词的词性只与前一个词的词性有关，而与其它词的词性无关。
         2. 统计出每种词性总的出现次数。比如v出现的总次数为c(v)
         3. 统计出任意两个词性前后依次出现的次数。比如v、n依次出现的总次数是c(n|v)
         4. 用极大似然法计算状态转移矩阵中的各项的值。比如q(n|v) = c(n|v)/c(v)。
      3. 计算混淆矩阵的值：
         1. 对每个词，分别统计它标注为各种词性的次数。比如c(人/n)=100，c(人/m)=1等等。
         2. 统计出每种词性总的出现次数。比如n出现的总次数为c(n)。
         3. 对所有词性，用极大似然估计法计算该词性出现时，出现某个词的概率。比如q(人|n)= c(人/n)/ c(n)。得到混淆矩阵。
   2. 测试
      1. 数据预处理：同训练时的预处理。
      2. 利用混淆矩阵和状态转移矩阵，用维特比算法解码：
      3. 回溯找到最优解
3. 评估：先计算每一句话的正确率（标注正确的词数除以总词数），再对所有句子的正确率求平均，得到平均正确率。
   1. 举例输出结果  
      
   2. 平均正确率是85.02%  
      
4. 缺陷：
   1. 没有处理未登录词。如果输入的词在训练语料中没有出现过，就无法给出整个句子的词性
   2. 断句不合理。程序中只是按照语料中的一行作为一个句子。实际上一行应该是一篇文档，包含多个句子。
   3. 未调用分词。程序所需的输入是字符串数组，该数组按顺序包含了一句话中的每个词。