**同济大学计算机系**

**中文信息处理课程实验报告**

****

**学 号 1552215**

**姓 名 田甘迅**

**专 业 计算机**

**授课老师 卫志华**

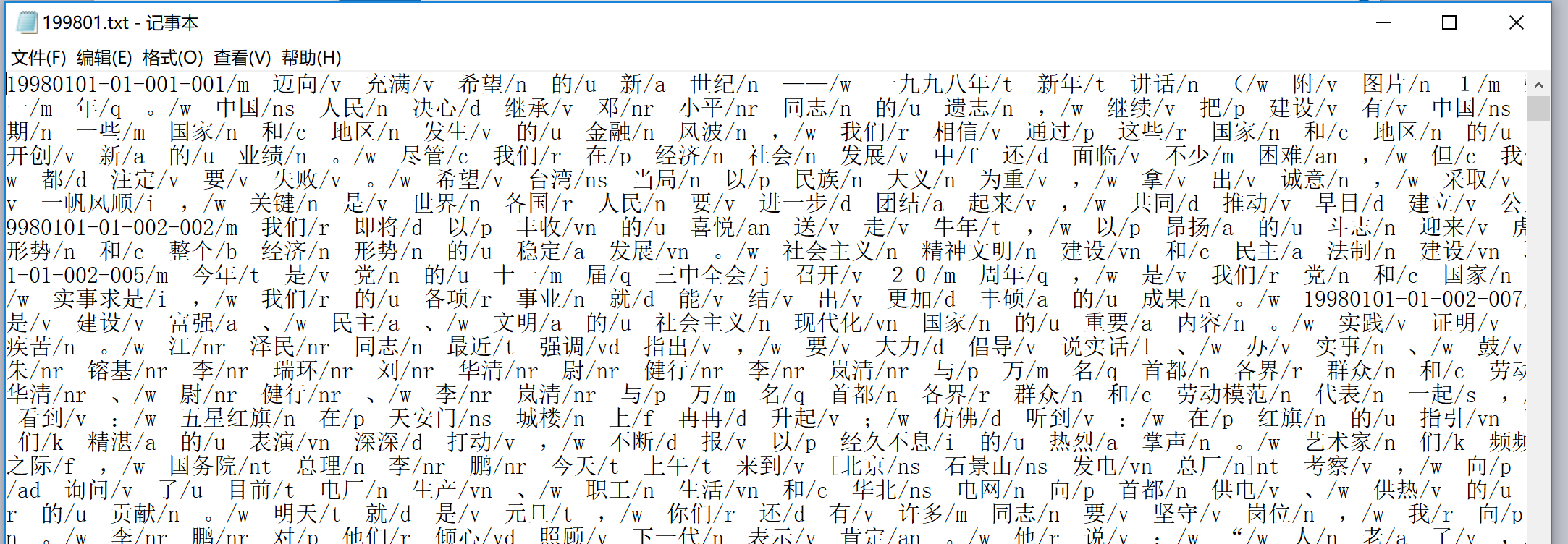
## 一、实验内容

## 词性标注（Part-of-Speech tagging 或POS tagging)，又称词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或其他词性的过程。在汉语中，词性标注比较简单，因为汉语词汇词性多变的情况比较少见，大多词语只有一个词性，或者出现频次最高的词性远远高于第二位的词性。据说，只需选取最高频词性，即可实现80%准确率的中文词性标注程序。

## 二、实验过程

1.语料库

训练语料采用了北大人民日报语料库切分语料



2.模型

使用了一阶隐马模型，在这个隐马尔可夫模型中，隐状态是词性，显状态是单词，使用viterbi算法进行分析，即利用转移矩阵和词频计算出HMM中的初始概率、转移概率、发射概率，进而完成求解。

3.viterbi算法

## 三、主要文件

199801.txt--北大人民日报语料库

Test.txt--测试文件（必须为分词后文件）

Pos.dat--词性标注结果

ictCtrlDlg.cpp--mfc源码

PosTag.cpp--算法程序源码

## 测试样例

可进行训练，词性标注及准确度的计算

## 

## 心得体会

Viterbi算法只需计算词频等数据即可得到结果，对算法要求不是很高。分词与词性标注一体化难以结合，只能以分此后结果进行标注，即test文件必须是分此后结果，希望有机会可以更加完善。