中国大学生计算机设计大赛



软件开发类作品文档简要要求

作品编号：　　　　　2022045891

作品名称：　　　　三位一体汉语词性分析系统

作　　者：　　　　　陈思 石庆波 赵开心

版本编号：

填写日期：　　　　2022.5.4

填写说明：

1. 本文档适用于**所有**涉及软件开发的作品，包括：软件应用与开发、大数据应用、人工智能应用、物联网应用；
2. 正文一律用五号宋体，一级标题为二号黑体，其他级别标题如有需要，可根据需要设置；
3. 本文档为简要文档，不宜长篇大论，简明扼要为上；
4. 提交文档时，以PDF格式提交本文档；
5. 本文档内容是正式参赛内容组成部分，务必真实填写。如不属实，将导致奖项等级降低甚至终止本作品参加比赛。

目 录

[第一章 需求分析 6](#_Toc100040660)

[第二章 概要设计 6](#_Toc100040661)

[第三章 详细设计 6](#_Toc100040662)

[第四章 测试报告 6](#_Toc100040663)

[第五章 安装及使用 6](#_Toc100040664)

[第六章 项目总结 6](#_Toc100040665)

[参考文献 7](#_Toc100040666)

# 需求分析

在中文信息处理领域，汉语词法分析是一项重要的基础性研究课题。它不是句法分析、语义分析、篇章理解等深层中文信息处理的基础，也是机器翻译、问答系统、信息检索和信息抽取等应用的关键环节，语词法分析包括汉语分词、词性标注与命名实体识别三项子任务，在国内外一些相关的评测中，常常将它们为三个独立的子任务进行评测。在已有的研究中，大部分研究者也习惯将三项子任务独立起来加以考虑，习惯于将汉语分词和词性标注依次处理，分词之后再在序列基础上考虑词性标注问题。这种将汉语词法分析的三项子任务独立处理的方法，虽然符合人们对汉语词法分析的认知规但容易造成错误向上传递放大累加，且多类信息难以整合利用的不足。针对这一问题，本文提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析过程看作字序列的标注过程，在每个字的标记中融合了词位、词性、命名实体三类信息，采用序列数据标注模型之一的最大熵模型实现了汉语分词、词性标注、命名实体识别三位一体的汉语词法分析

# 概要设计

本次任务是基于隐马尔科夫模型实现的分词，词性标注，以及命名实体识别，通过功能函数的调用实现窗口读取电脑文件

# 详细设计

通过 tk界面化，以及文件路径的选择，文件的打开和写入，利用隐马尔科夫模型和维特比算法实现三位一体汉语词法分析。主要难点为隐马尔科夫模型的训练和维特比算法的使用

# 测试报告

该作品有四个汉语词性标注语料。首先是将原词性标注语料拆分为一字一标记的格式，此时的标记形式为“词位＿词性类别”，然后再根据命名实体语料将所有命名实体的那部分字的标记修改为“词位＿命 名实体类别”。然后 将一部分作为训练语料，一部分作为测试语料。统计发现，这些语料中共有字标记257种，采用工具包进行模型训练时，还需要对融合后的语料进行预处理，按照设定的样本窗口和特征模板集 将语料处理为一行一个事件的语料，也就是对每一个样本按照特征模板集扩展出相应的上下文特征作为一个事件。

# 安装及使用

【填写说明：简要说明安装环境要求、安装过程、主要流程等。建议包含默认安装和典型使用流程。】

# 项目总结

在中文信息处理领域，汉语词法分析是其中一项重要的基础性研究课题。针对汉语词法分析中分词、词性标注、命名实体识别三项子任务分步处理时多类信息难以整合利用，且错误向上传递放大的不足，本文提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析过程看作字序列的标注过程，将每个字的词位、词性、命名实体三类信息融合到该字的标记中，采用最大熵模型经过一次标注实现汉语词法分析的三项任务。实验结果表明，三位一体字标注方法的分词、词性标注、命名实体识别的性能都有不同程度的提升。今后将

进一步完善该方法，力争能在中文信息处理的实际任务中推广应用。

# 参考文献

【1】 于江德，胡顺义，余正涛 三位一体字标注的汉语词法分析