

**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目：基于深度学习的古汉语实体识别系统**

**专 业 计算机科学与技术**

**姓 名 刘鸣宇**

**学 号 20201102467**

**指导教师 林民**

**日 期 10.28**

**计算机科学技术学院制**

**说 明**

1. **开题报告主要内容**
2. 课题来源及研究的目的和意义
3. 课题来源

在当前时代，随着科学的进步和人工智能的发展，自然语言处理和知识图谱技术迅猛发展，不少人利用互联网技术来学习语言文字。而作为文化古国，我们的古代文学中蕴含着大量的古文，让一些初学者望而却步。但是随着人工智能的发展，实体识别技术随之而生。在科技发展到一定阶段的今天，我们可以利用现代科技手段，开展对古籍文献的研究和探索。然而由于古代汉语和现代汉语在语义和语法上存在较大差异，阅读古文需要消耗较多的时间以及精力。而实体的关系抽取是自然语言理解的基础，实体识别是实体关系抽取的前提。在阅读文本的时，需要先进行所需理解的信息的词性的识别和分类，从而提取实体关系，为此，古文实体识别成为了在古文学习有待解决的首要任务。

1. 研究的目的和意义

利用计算机对古文进行自动化的信息实体识别，系统自动识别实体并进行标注。从而提高读者阅读古籍的效率，加深读者对古文的理解。同时它也是篇章理解的一个核心技术，对信息检测、问答系统、信息过滤、机器翻译等有非常积极的意义。

1. 个人研究目的

本人负责部分是，研究如何将已有的语言数据提取出实体，以及在软件页面上显示实体分类、事件提取、事件聚类等功能，并且为用户提供可自定义实体的服务。

1. 国内外在该方向的研究现状及分析
2. 国外

在命名实体识别方向中，首先是英国对实体识别的研究比较早。在1991年首次提出实体识别，1999年提出了基于隐马尔可夫模型的英文命名实体识别任务，并且达到了较高的准确率。2011年，collobert等人采用了神经网络框架来构建命名实体识别，这是早期把神经网络的技术应用到实体识别中。后续国外各学术界不断更新，最新的是谷歌在2020年联合CMU、Deep Mind提出了XTREME，它覆盖了40多种语言，基于子序列的深度主动学习方法，该方法大幅度减少实现高性能所需要的数据。只需要原始数据的25%就可以实现最佳性能。由于英文本身是不需要考虑分词问题，因此就比较容易。实体三个指标也可以达到较高程度。标签集的划分，各个类型划分的标签集数量也是不一致的。但是是向标签集个数越来越多的趋势发展。

1. 国内

汉语方面，在1995年开始进行人名识别分析，后续直到2020年，研究重点逐渐由模型从实际应用转化，并且指标也达到了90%左右。但由于起步较晚，应用领域较少，在各学科领域的应用还不是特别成熟。

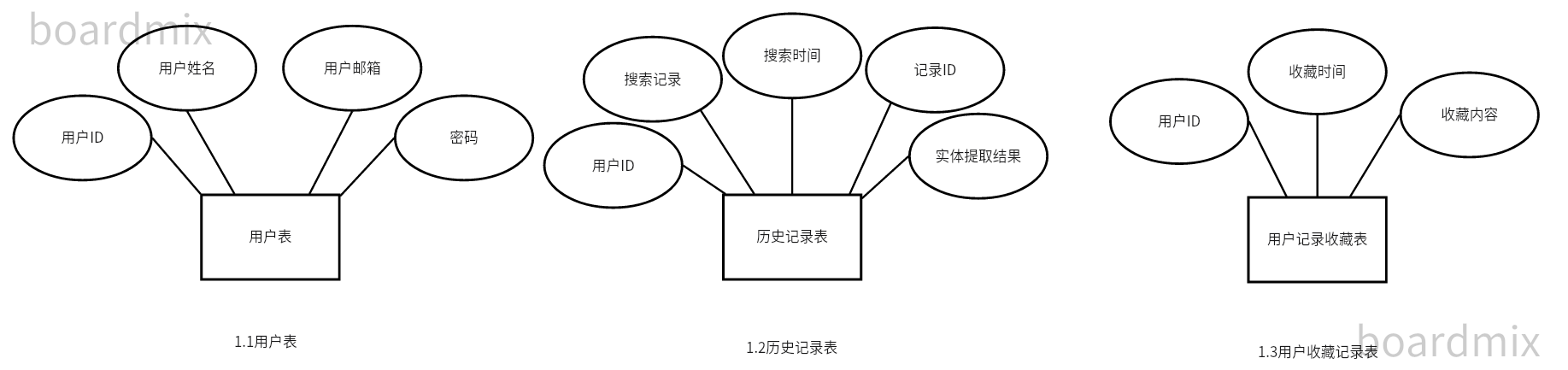
1. 主要研究内容
2. 针对的对象

除了一些需要专业古籍的古文学术研究者，还可以扩充到所需要阅读古籍的人群，其中包括：需要阅读古籍寻找早期的文献资料、历史研究学者、对古文感兴趣的群体、需要古文启蒙的学生…等等人群。

1. 通过展示不同事件来划分相同实体的不同表达

通过对学校文学院的同学的咨询，了解到，当他们在读古文的时候，对应相同的人物，在不同时间，不同的地点所表达感情是可能不一样的。也就是说，即使是实体相同，表达的情感也可能不一样，关于这点我们可以通过区分发生的事件来展现不同的实体。这样的话也有助于用户理解古文，而不是仅仅是在翻译的基础上，通过实体分类展示出来实体之间的联系，这样，我们就可以十分清楚的掌握文章的人物情感或是人物之间发展的故事，能够让读者更好地理解古籍。相应的，故事类型的古文，用图来展示关系，是比较好的方式，在其中可以加入人物的情绪。

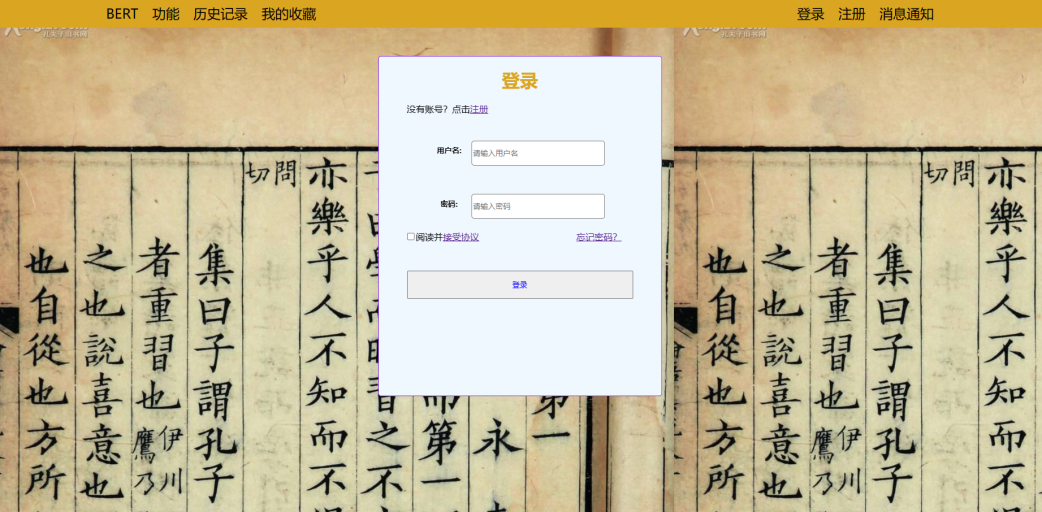
1. 研究任务
2. 用户信息模块：通过注册功能收集用户信息，通过登录功能验证身份，可以在个人中心中修改信息；
3. 用户历史记录模块：用户可以在历史记录功能中查看历史记录；
4. 用户收藏夹模块：可以在个人收藏夹内查看收藏的信息；
5. 用户上传文章模块：可通过文件上传至输入栏，也可以将文章或段落粘贴复制到功能界面，自动生成实体提取效果，并可根据用户需求实现；
6. 实体提取模块：将文章上传到系统中后，系统自动生成实体提取结果，提取结果存入到数据库中；
7. 事件提取模块：将文章上传到系统中后，可以选择事件提取功能，系统自动划分事件，同时标注实体；
8. 实体聚类模块：实体提取结果生成后，可以选择实体聚类功能，系统页面显示实体出现次数；
9. 用户自定义实体模块：若用户对系统生成结果不满意，可在系统生成结果上修改，修改后结果记录到数据库中；
10. 数据库设计：



1. 系统功能
2. 用户登录注册功能：

输入：用户名，用户密码等信息进行账号注册或用户登录功能；

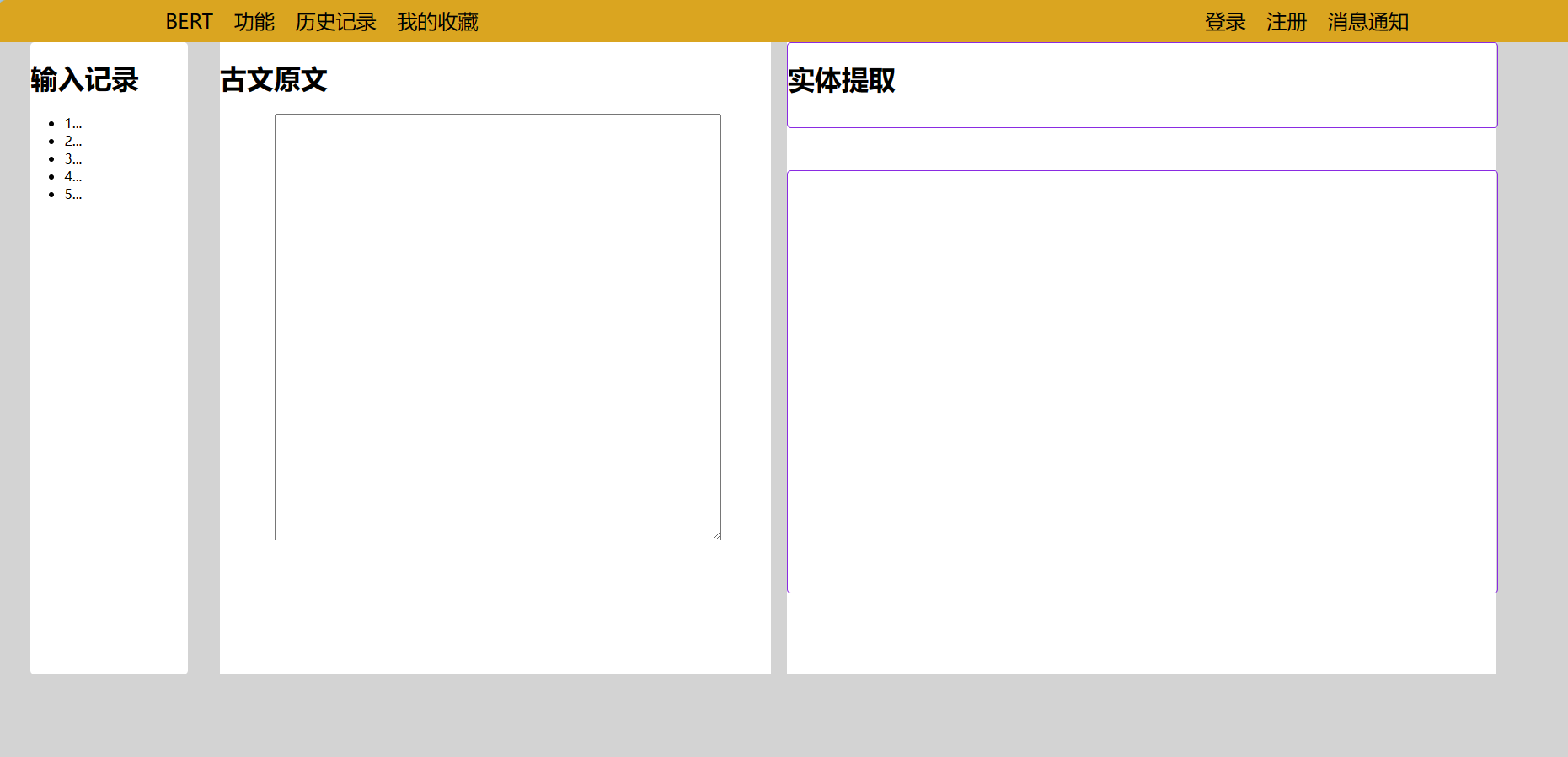
输出：注册成功/失败 或 登录成功/失败；



1. 用户查看历史记录：

输入：账号登录后，点击软件页面“查看历史记录”按钮；

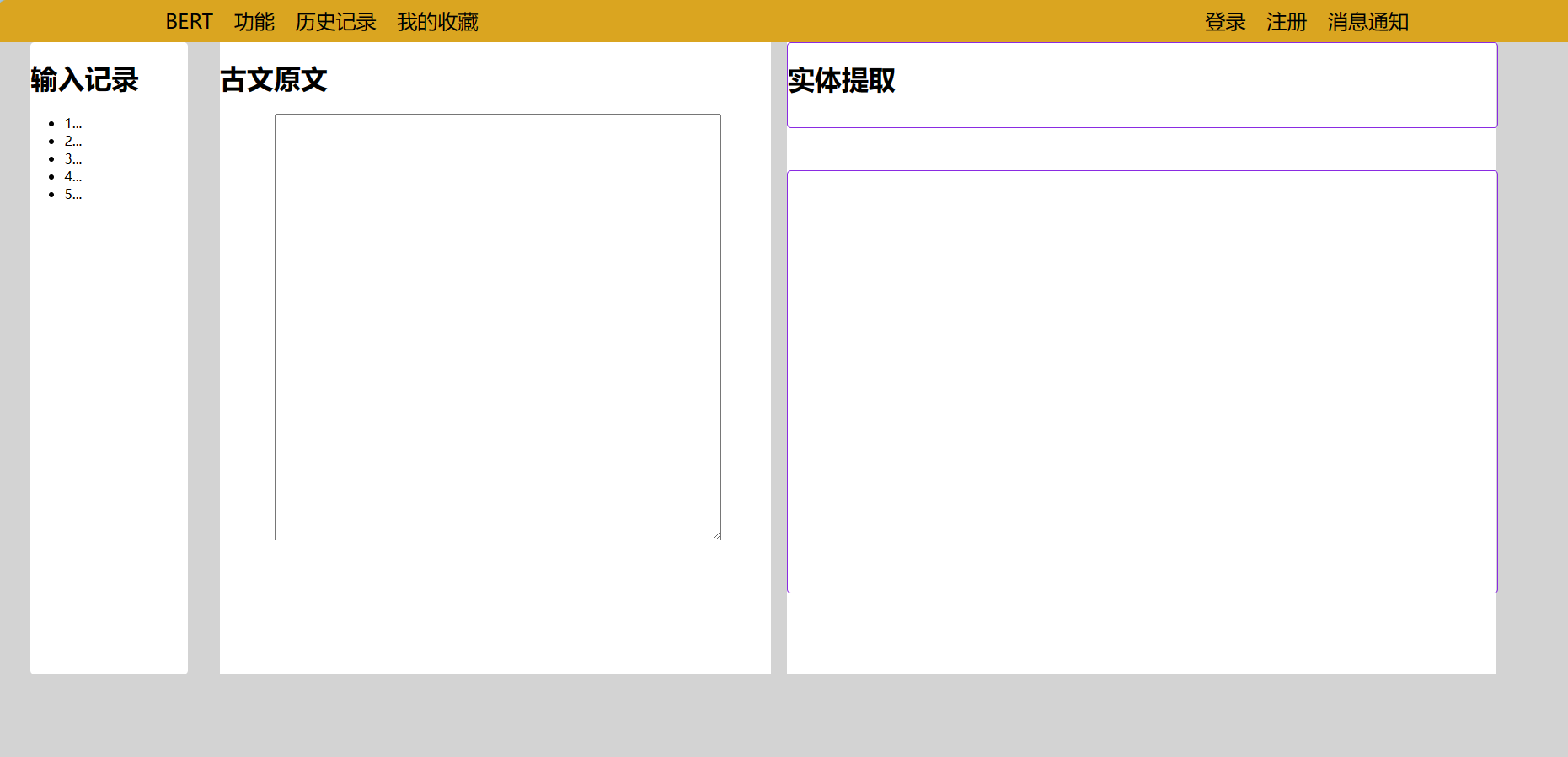
输出：软件进行页面跳转，显示最近一段时间内的使用记录；



1. 用户进行结果收藏：

输入：账号登录后，用户到实体提取页面，输入文档后产生实体提取结果，点击“收藏”按钮即可；

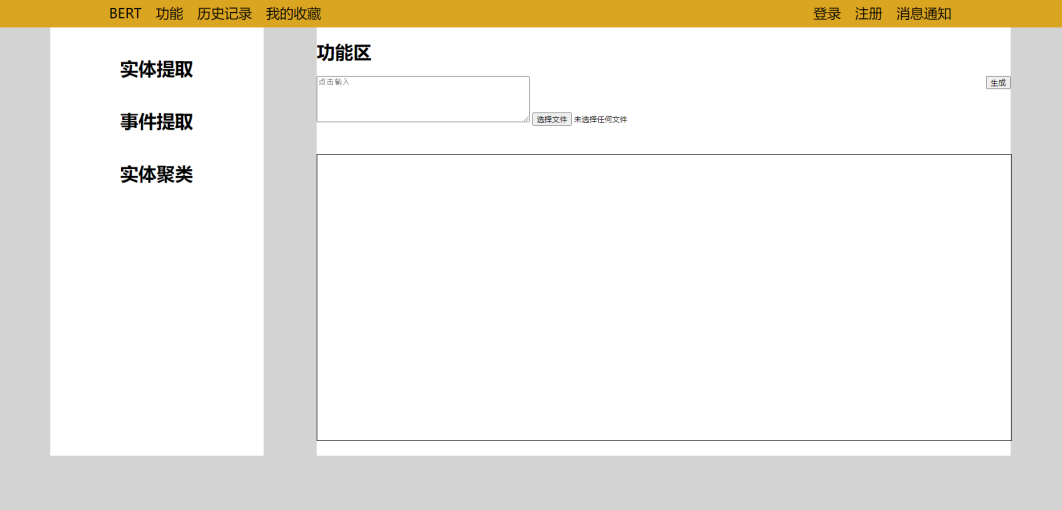
输出：收藏后，将本次结果存放到用户收藏记录表中，可以跳跃到查看收藏界面进行查看；



1. 实体提取功能：

输入：可以通过上传文档，或者直接在软件界面输入文章或段落，输入后点击“生成”按钮

输出：页面内输出提取结果，可选择导出图片



1. 事件提取功能：

输入：使用左侧功能区“事件提取”，可以通过上传文档，或者直接在软件界面输入文章或段落；

输出：通过系统运算，先输出文章中一共发生了几个事件，然后输出文章中发生的不同事件中的实体；

1. 实体聚类功能：

输入：使用左侧功能区“实体聚类”，可以通过上传文档，或者直接在软件界面输入文章或段落；

输出：通过系统运算，输出出现频率最高的十个实体；

1. 自定义实体识别：

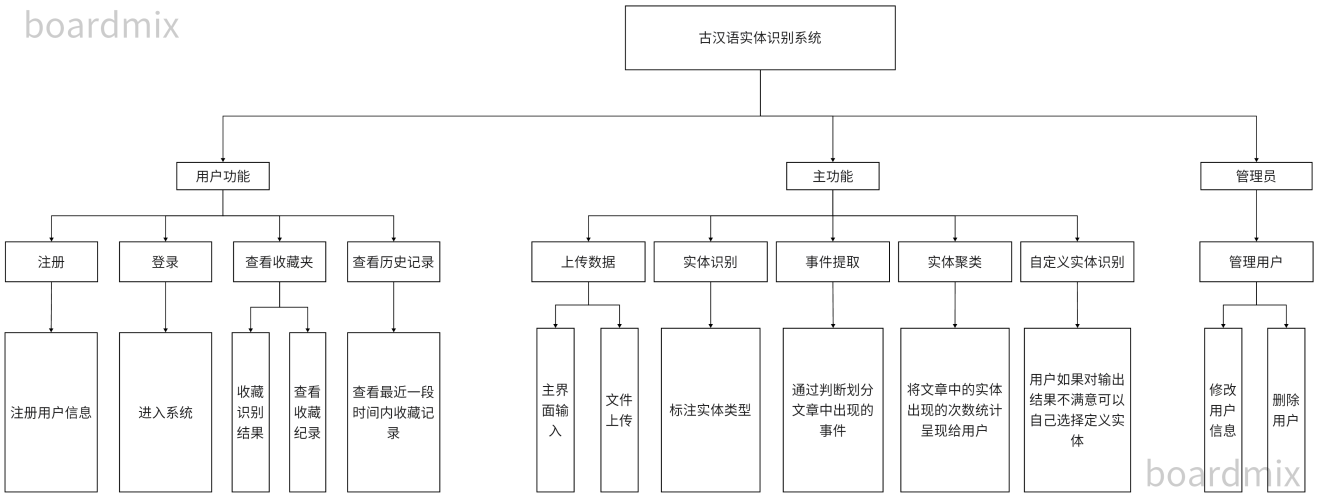
输入：用户如果对实体标注的结果不满意，可直接在文章段落上进行标注并将结果记录到数据库中；

4. 研究方案

1. 总体研究方案

本系统主要功能为对古文进行实体识别及对收集到的数据进行处理，从而在前端生成概括古文大意的可视化结果及文字结果，可以在功能页面上传，对于可视化结果可以以PDF格式导出，或者导出图片，文字结果可以复制及收藏；次要功能为用户信息的处理包括注册登录及修改信息功能，历史记录功能，用户收藏夹功能。

1. 个人研究
2. 历史记录：为用户提供搜索语句和搜索文字结果，以及搜索时间。
3. 实体识别的文字表现：将抽离的实体关系、实体事件与实体，利用文字排版清晰地展示，并按序排列文章中出现次数前十的实体，并提供修改文本的选项，将修改后的结果一致存入数据库中。
4. 文件上传：实现用户输入的多样化，可进行txt，word文件的上传，对文件中的数据进行分析处理。



5．进度安排，预期达到的目标

1. 第一阶段（2023/10-2023/12）

功能需求确定，学习开发需要使用的技术，前端界面搭建完成。

1. 第二阶段（2023/12-2024/2）

后端接口对接成功，学习后端基于深度学习完成的实体关系抽取模型。

1. 第三阶段（2024/2-2024/3）

测试代码，优化页面，性能测试，以及讨论是否有必要进行功能修改。

1. 第四阶段（2024/3-2024/4）

论文修改，准备答辩。

6．课题已具备和所需的条件、经费

小组人员，一人一部笔记本电脑，以及安装相应开发工具。

7．研究过程中可能遇到的困难和问题，解决的措施、

1. 可能遇到的问题：

关于具体功能的实现、模型训练不够、可能会遇到较多的专业型错误、对接口功能了解不够透彻对接出现问题。

1. 解决措施:通过充分的小组讨论，与老师交流，以及在网络中查找相关资料文献。

8．主要参考文献

[1]韩普,顾亮.基于混合深度学习的中文医学实体抽取研究[J].图书情报工作,2022,66(14):119-127.DOI:10.13266/j.issn.0252-3116.2022.14.012.

[2]袁悦. 面向实体抽取的先秦典籍词性标记方法研究[D].南京农业大学,2019.DOI:10.27244/d.cnki.gnjnu.2019.002356.

[3]环科尤. 基于深度学习的格萨尔史诗命名实体识别关键技术研究[D].青海师范大学,2022.DOI:10.27778/d.cnki.gqhzy.2022.000638.

[4]程顺航,李志华,魏涛.融合自举与语义角色标注的威胁情报实体关系抽取方法[J/OL].计算机应用:1-11[2022-10-27].http://kns.cnki.net/kcms/detail/51.1307.TP.20220726.1045.004.html

### [5]庄百川 基于深度学习的古文自动断句与标点研究[D]武汉邮电科学研究院,2022. DOI：10.27386/d.cnki.gwyky.2022.000049

### [6][杨泽](https://kns.cnki.net/kcms2/author/detail?v=kMpVSI0yL5daWap-M5Uk2a-WpwUj5vFcxMgNuq28M34hLfWXJx0hMoov7bgcbVWvaEWEL7TZ5_G7Ntx5t0ojH8bddG9OsS93zM1Voron-Go=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank) 中国古典文学文本的命名实体识别及知识图谱构建研究[D]南京邮电大学,2022. DOI:10.27251/d.cnki.gnjdc.2021.000428

### [7][吕子颖](https://kns.cnki.net/kcms2/author/detail?v=kMpVSI0yL5f78EX6HHJaKFApDwevEtL4Hxd2ikqa7InKPnMp2IT-K4pwhDpDRKJAgFd8G_I_wOJvHhVKRFL9FlFrmH94EFN-4e-8wUFHXQ4=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank) 古文命名实体识别的研究与实现[D][大连理工大学](https://kns.cnki.net/kcms2/organ/detail?v=kMpVSI0yL5eH-dI_GMh3K5Q8pdOK2kuO3LmfXPvvL6hTZmV_Loo8V2hhXnl4bKyGJgpiPXvvEt10vLRI4aisvyfKf3zg3L59IZNXGVS4IDKtgnuP9VpL1w==&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank),2021.DOI:10.26991/d.cnki.gdllu.2020.002311

### [8][成晓](https://kns.cnki.net/kcms2/author/detail?v=kMpVSI0yL5d3QmwS_mck_9LUvDf59wiPCXXkaUGl3xTCNespuL5IQgXbFXeuqfXRw1QD0zlsbfFOjTn9tJMJehSYb4SPq4SZnmUfrtQUcV4=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank) 面向旅游领域的蒙古文命名实体识别的研究[D][内蒙古大学](https://kns.cnki.net/kcms2/organ/detail?v=kMpVSI0yL5cF-T2vTTF3R9BlLqtD-l-L2eji1L1Dq0HpQ3AhZl129wDUITXhpAq2ZaDd5N8mAz_qyZpWWk1ioU8DhZi2bo5jxUsN4ZUY3KE=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank),2021.DOI:10.27224/d.cnki.gnmdu.2021.000921

### [9][刘泽庆](https://kns.cnki.net/kcms2/author/detail?v=kMpVSI0yL5ejL1GbTUFcUPwWXChAOX-J8VpuXNX4s-ZiXAtieTHBUgUUjaDKGbvtfTS-gIGkcrtPtmoKPi_GiZU4uW1i3cwiktrB9roch5M=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank) 融合HowNet的深度学习下的中医文本实体识别[D]华北理工大学,DOI:10.27108/d.cnki.ghelu.2021.001018

### [10][詹子依](https://kns.cnki.net/kcms2/author/detail?v=kMpVSI0yL5fYvfi3jTswTOo4-tjbViueWKd113n3mv5iQt7nD7nIGnPRvoPOcVLud9aKERavluJo7ca4agLM0wGyF71anLgIK-yqgwOPCD8=&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank) 面向古汉语领域的命名实体识别[D][东北石油大学](https://kns.cnki.net/kcms2/organ/detail?v=kMpVSI0yL5dO8_x1_HiwlAkcxTEVQDsWt3wvfRuBFE097E8Ge5r7RL_z8PEpNlZuVHjbE6RM3pCN1Flr8JAvb14iCvc1EBGVZKpgK8Eet5nTiSyPiu7U6Q==&uniplatform=NZKPT" \t "https://kns.cnki.net/kcms2/article/_blank),DOI:10.26995/d.cnki.gdqsc.2023.000301