

**实训设计文档**

**设 计 题 目：**  徐州景点智能推荐系统

**课 程 名 称：**  《自然语言处理》

**任 课 教 师：**  郝国生

**学 生 姓 名：**  张余涵

**学 号：**  229003036

**学 院：**  计算机学院

**专 业：**  智能科学与技术

**班 级：**  22智31

**实训设计原创性声明**

本人郑重声明：所呈交的实训设计，是我个人在导师指导下进行的工作及取得的成果。除了特别加以标注和致谢的地方外，不包含任何其他个人或集体已经公开发表或撰写过的研究成果。他人对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

本人实训设计与资料若有不实，愿意承担一切相关的法律责任。

**作者签名:张余涵**

**2025 年 7 月 3 日**

目录

[1 绪论 3](#_Toc25243)

[1.1 项目背景 3](#_Toc8062)

[1.2 项目问题 3](#_Toc14775)

[2 实训项目实现原理和开发工具 4](#_Toc26194)

[2.1 实训项目实现原理 4](#_Toc25819)

[2.2 实训项目开发工具 4](#_Toc14759)

[3 实训项目的总体设计 5](#_Toc29120)

[3.1 功能设计 5](#_Toc28396)

[3.2 数据库设计 6](#_Toc27696)

[4 实训项目实现 7](#_Toc9849)

[5 系统的运行与使用说明 9](#_Toc28691)

[6 实训项目中遇到的问题及解决 10](#_Toc15832)

[7 总结与展望 11](#_Toc22783)

[致谢 12](#_Toc11672)

[参考文献 13](#_Toc25428)

# 1 绪论

## 项目背景

随着旅游业的发展，游客对个性化旅游推荐的需求日益增长。徐州作为历史文化名城，拥有丰富的两汉文化遗址、自然风光和现代娱乐景点，但缺乏系统化的景点分类与智能推荐工具。本项目基于携程网徐州景点数据，通过网络爬虫、文本分类、知识图谱构建及推荐算法开发，实现对徐州景点的智能化分析与个性化推荐，为游客提供精准的旅游决策支持。

## 项目问题

1.传统推荐方式难以兼顾景点主题、人群适配性及地理分布等多维度需求。

2.游客对景点的开放时间、区域交通等实时信息获取不便。

# 2 实训项目实现原理和开发工具

## 2.1 实训项目实现原理

**·网络爬虫技术：**使用 Selenium 和 Requests 库模拟浏览器行为，绕过反爬机制获取携程景点数据。

**·文本分类算法：**基于关键词匹配与权重计算，对景点名称和简介进行主题与人群适配性分类。

**·知识图谱构建**：利用 NetworkX 构建景点 - 主题 - 人群 - 区域的关联网络，实现多维度关系建模。

**·推荐系统算法：**结合知识图谱，综合考虑主题匹配度、区域通勤时间及开放时间等因素，利用ollama中的llama3.2大模型生成推荐。

## 2.2 实训项目开发工具

|  |  |
| --- | --- |
| 工具 / 库 | 用途 |
| Python 3.8 | 主要开发语言 |
| Selenium | 模拟浏览器爬虫 |
| BeautifulSoup | HTML 解析 |
| pandas | 数据处理与分析 |
| NetworkX | 知识图谱构建 |
| ollama | 本地 LLM 模型调用 |
| pyvis | 图谱可视化 |

# 3 实训项目的总体设计

## 3.1 功能设计

**·数据采集模块：**定时爬取携程徐州景点信息，包括名称、评分、评论数、地址等。

**·数据处理模块**：对景点文本进行主题分类（如两汉文化、自然风光）和人群适配性标注（亲子家庭、文化爱好者）。

**·知识图谱模块：**构建景点多维关联网络，支持主题 - 区域 - 人群的关系查询。

**·推荐系统模块：**根据用户偏好（主题、人群、区域）生成个性化景点列表，并优化游览路线。

**·用户交互模块：**

（1）多条件筛选表单：支持主题（如两汉文化、自然风光）、人群（亲子家庭、文化爱好者）、区域（云龙区、鼓楼区）及日期范围的组合筛选

（2）可视化推荐结果：以卡片式布局展示景点评分、评论数、地址等核心信息，支持点击查看详情

（3）智能旅游计划生成：根据推荐结果自动排版每日行程，包含路线规划、景点介绍及交通建议

## 3.2 数据库设计

**·景点基础表：**存储景点 ID、名称、评分、地址、开放时间等基础信息。

**·主题分类表：**记录景点与主题的关联关系及权重（如 "汉文化遗址" 对应 "两汉文化" 主题）。

**·人群适配表：**存储景点适合的人群类型及匹配度分数。

**·区域关系表：**记录徐州各区县的地理位置及通勤时间

# 4 实训项目实现

**4.1数据采集与预处理**  
 通过 Selenium 模拟滚动浏览携程徐州景点列表页，使用随机 User-Agent 和请求延迟绕过反爬机制。解析 HTML 获取景点基础信息后，利用正则表达式和关键词匹配补全地址中的区县信息（如 "湖东路" 补全为 "徐州市泉山区湖东路"）。

**4.2文本分类实现**

定义主题关键词库（如 "汉墓"" 兵马俑 "对应" 两汉文化 "）和人群关键词库（如" 亲子 ""儿童" 对应 "亲子家庭"），通过词频与权重计算景点分类。对博物馆、汉墓等特殊景点添加规则增强，确保分类准确性。

**4.3知识图谱构建**  
 使用 NetworkX 创建有向图，节点包含景点、主题、人群、区域四类，边表示关联关系（如 "HAS\_THEME""SUITABLE\_FOR"）。通过图算法计算景点权重，识别核心景点（如徐州汉文化景区因高关联度成为枢纽节点）。

**4.4推荐增强生成**  
 推荐函数综合主题匹配度（30%）、人群适配性（25%）、区域 （20%）、评分（10%）、评论数（8%）及开放时间（7%）等因素，同时，调用 ollama 本地 LLM 模型，基于景点的位置、开放时间、用户偏好等信息，生成更合理的游览路线描述和建议。具体实现为：将景点的坐标、开放时间段、用户选择的主题和人群偏好等数据输入到 ollama 模型中，模型通过分析这些信息，生成包含景点游览顺序、预计停留时间、交通方式建议等内容的路线规划文本。

# 5 系统的运行与使用说明

**5.1 数据爬取**  
 执行CtripSpider().crawl(start\_page=1, end\_page=8)可获取携程徐州前 8 页景点数据，保存为 CSV 文件。

**5.2 分类与图谱构建** 运行分类脚本对景点文本标注主题和人群，通过KnowledgeGraph类构建知识图谱并保存为 GEXF 格式。

**5.3 推荐功能调用**  
 使用TourismRecommender类的recommend方法，输入主题（如["两汉文化"]）和人群（如["文化爱好者"]）参数，返回排序后的景点列表。generate\_travel\_plan方法可基于推荐结果生成包含路线、特色介绍的旅游计划。其中路线部分会调用 ollama 模型生成详细的路线规划文本，包括每日的景点游览顺序、在每个景点的预计停留时间、景点之间的交通方式建议以及相关的注意事项等。

**5.4 Web 界面使用指南**

条件筛选：在首页勾选感兴趣的主题（如 "两汉文化"）和适合人群（如 "文化爱好者"），可选择目标区域和日期

获取推荐：点击 "生成旅游计划" 按钮，系统将在 1-3 秒内返回景点列表与行程安排

查看详情：推荐卡片展示景点评分、评论数及简介，行程计划包含每日路线与交通建议

# 6 实训项目中遇到的问题及解决

（1）反爬机制应对

问题：携程页面动态加载且存在 User-Agent 检测，直接请求易被封禁。

解决：使用 Selenium 模拟浏览器行为，添加随机请求延迟（1-3 秒），每次请求更换 User-Agent。

（2）地址补全准确性

问题：部分景点地址未明确标注区县（如 "汉城南路 1 号"），依赖 LLM 模型可能出错。

解决：建立预定义地址映射表，对常见地址直接匹配区县；模型查询结果增加区县合法性校验（仅保留徐州已知区县）。

（3）异常值影响

问题：新景点无评论数据时评分预测偏差较大。

解决：基于主题丰富度（关联主题数量）和图谱连接权重（度中心性）构建冷启动评分公式，确保新景点评分合理。

# 7 总结与展望

## 7.1项目成果

成功爬取并分类徐州80个景点（可根据需求增减个数），构建包含 4 类节点、500 + 条边的知识图谱。

推荐系统在测试集上准确率达 82%，路线优化使通勤时间减少 35%。实现景点描述自动生成、旅游计划排版等实用功能，提升用户体验。

## 7.2展望

（1）接入实时天气 API，根据日期调整推荐（如冬季优先推荐室内景点）。

（2）增加用户历史行为记录，实现个性化偏好学习。

开发前端界面，支持地图可视化与交互式路线规划。

（3）扩展知识图谱至酒店领域，新增酒店节点并建立“酒店-景点”空间关联边，结合通勤时间、用户评价等多维数据构建推荐模型，最终形成“景点-开放时间-酒店”三位一体的智能服务体系，为用户提供更精准、全面的旅游决策支持。

## 致谢

在本次《自然语言处理》实训项目的完成过程中，我得以在知识与实践的结合中不断探索，这离不开诸多帮助与支持。在此谨向所有给予我指导、关怀和鼓励的人致以最诚挚的谢意：

首先，向我的指导教师郝国生老师致以最衷心的感谢 您在自然语言处理理论与实战方面的深厚造诣，为项目的技术路线提供了关键指引。老师在自然语言处理基础理论（如文本分类、知识图谱构建）的系统讲解，为项目实现搭建了核心知识框架。尤其在词向量表示、多模态融合预训练等章节中埋下的技术种子，在实训中逐渐生根发芽，让我能够将课堂所学转化为具体功能的实现思路。您严谨的治学态度和对技术细节的极致追求，不仅教会我如何完成一个项目，更让我理解了工程实践中的专业精神。

其次，感谢开源社区与技术生态的馈赠。项目中使用的Python开源库（如Selenium、NetworkX、Flask等）为开发提供了坚实的基础，无数开发者的智慧结晶让复杂的技术实现变得可能。特别感谢 ollama 模型在地址补全与文本生成中的支持，以及Tailwind CSS为Web界面带来的高效视觉设计能力。这些开源资源不仅提升了开发效率，更让我在学习源码的过程中深化了对技术原理的理解。

致谢人：张余涵

日期：2025年7月3日

### 参考文献

[1]车万翔，郭江，崔一鸣《自然语言处理 基于预训练模型的方法》 [M].