《大数据计算机基础》

2024年秋季学期大作业选题

大作业说明：

大作业是检验学期教学的重要方式，务必重视，其在期末总成绩中权重较高。大作业选题可根据自己兴趣确定，但必须报备获得任课老师认可。

大作业的评分关注重点包括：选题质量、选题难度、完成程度、代码注释、文案完备性以及界面美观等等诸多方面。文案包括：简要需求分析、简要项目说明、重点难点分析、简要测试数据、简要操作手册等等。

另外，大作业可以单人完成，也可以双人或三人完成组队完成。相对来讲，组队模式的选题难度要大于单人模式。

备选题目：

1. 【可两人组队】计算最佳新地铁线路。利用北京公共交通数据（地铁和地面公交）设计最佳北京新地铁线路。
2. 利用提供的新浪新闻数据，发现词语。注意，不得利用jieba等第三方自然语言处理库。可考虑多个n-Gram组合利用。计算出的词汇，需要与jieba等分词库进行对比研究；
   a) 核心目标，探索n-Gram在分词中的价值，结合大数据人工智能技术；
   b) 首先建立特定字集，比如：借此、语气词、时态词、数量词等特定字集，也包括各种标点符号前后的词语，都是一些功能性字，如：说等等；利用上述字，寻找特定组合，如：洗了洗、实际上是洗洗的变体；
   c) 形成多个N-Gram比如：2-Gram、3-Gram、4-Gram、5-Gram等，探索利用这些数据，形成词汇。这是一个数据化过程，也是在此基础上的应用；
   d) 结果与jieba进行分析对照；

3. 【可两人组队】下载Harvard University的https://projects.iq.harvard.edu/cbdb的数据库，结合网络上各种文献，分析该数据库结构，各个字段（属性）含义。开发Web界面，可在地图上显示多人迁移路径，人际关系等多种应用。至少达到如下目标：
   a) 数据库设计，尤其是表间关系；
   b) 数据库所有字段的含义；
   c) 探索中国历史人物如何数据化，以及数据如何存储，要求以案例说明；
   d) 历史人物的维度，以及每个维度的各朝代数量分布；
   e) 在地图上呈现多个历史人物的迁移关系；
   f) 可对某个专题深入分析，比如：隋唐时期的门阀政治、西晋门阀政治的区别等等；
4. 【可3人组队】对39所双一流高校网站评估。利用网络爬虫，对39所双一流高校（可减少数量，但不少于10所）官网进行多维度评估，包括但不限于：网页规范、图片规范、信息更新、栏目设置、栏目数量，热词跟踪等等；
   a) 优先爬取2-5个网站，并根据相应数据建立数据维度；
   b) 在数据维度基础上，建议评价指标；
   c) 更多网站数据爬取及其评价；
5. 基于作业的货车GPS数据，做如下计算：
   a) 货车OD(Origin-Destination)分析，利用算法，寻找货车前20名聚集地，并可视化呈现该聚集地的主要OD（排名前20），即来自和去往；
   b) 基于a)，做时序分析，即聚集和离开与时间的关系，比如：聚集时间分布、离开时间分布；
   c) a)和b)最好都呈现之上，或高德地图、或百度地图，或OSM地图；
   d) 上述应用如能结合地图路网数据更好。
6. 利用老师提供的脱敏GPS数据，分析交通维度：
   a) 区分步行、骑行、电动自行车、摩托车、自驾车、公交等交通模式；
   b) 每种交通模式的距离；
   c) 不同交通行为的时间分布；
   d) 职住识别，即工作位置、居住位置等；
   e) 可在此基础上充分利用数据做出更加有特色的应用；

7. 利用公开股票交易数据，做如下分析：
   a) 相关股票分析，即分析哪些股票之间的历史交易数据相关；
   b) 利用人民日报、中央电视台新闻联播，抽取关键字，分析事件与股票的关系，或者说分析哪些股票与国际政治形势相关；
   c) 利用已有历史股票数据，寻找异常交易股票；
   d) 分析多地上市股票的交易数据的相关性，寻找错配股票
8. 爬取股吧数据，做如下分析：
   a) 利用已有情感分析数据和其他分词软件，对股吧聊天记录做情感分析，分析出持有、重仓、抛售等倾向数据；
   b) 与历史交易数据进行比对，分析股吧聊天信息与交易记录之间的关系；
   c) 查找哪些聊天人是黑嘴或者托儿。