Python语言程序设计
课程大作业说明

何博 2020200425

结课项目概况说明

- 分为三个级别
  - Simple (Oct. the 25th), 35 point
  - Intermediate (Nov. the 22nd), 45 point
  - Challenging (Dec. the 20th), 55 point
- 提交文件
  - 源代码
  - 实验报告
  - 展示幻灯片
  - 演示视频
- 提交方式
  - 将上述文件打包为zip格式
  - 上传到百度网盘或文叔叔，生成文件链接
  - 在 tronclass 上提交作业链接

项目说明——Simple

- 内容
  - 阅读一个Python项目（例如游戏、爬虫、可视化程序）源代码
  - 分析该项目的代码结构
  - 撰写一份能够介绍该项目的报告
  - 制作展示该项目的幻灯片和视频

- 需提交文件
  - 项目源代码
  - 报告
  - 幻灯片
  - 视频

项目说明——Intermediate

- 内容（在以下项目中四选一）
  - Loan Prediction
  - Housing Prices Prediction
  - Titanic Survival Prediction
  - Wine Quality Prediction
- 需提交文件
  - 项目源代码
  - 报告，其内容包括但不限于：问题描述，数据集说明，方法介绍，关键代码细节、运行截图、实验结果、结果分析等
  - 幻灯片，用于展示完成任务的主要步骤及方法描述
  - 视频：对实现方法及运行结果进行演示。

项目说明——Intermediate

- 实验结果——需要包含对应任务的全部指标
  - 二分类问题指标
    - Accuracy
    - Precision & Recall
    - F1 Score & AOC
  - 回归问题指标
    - Mean Squared Error
- 数据
  - 随本文件同URL提供

项目说明——Loan Prediction

- 项目介绍
  - 基于用户的婚姻状况，教育程度，受抚养人数和就业情况，对用户是否可以贷款进行预测。
  - 数据与数据说明见前文链接所提供的文件。
- 需提交文件
  - 项目源代码
  - 报告，其内容包括但不限于：问题描述，数据集说明，方法介绍，关键代码细节、运行截图、实验结果（二分类指标）、结果分析等
  - 幻灯片，用于展示完成任务的主要步骤及方法描述
  - 视频：对实现方法及运行结果进行演示。

项目说明——Housing Prices

- 项目介绍
  - 房屋价格根据车库情况、房间数量等各种因素而变化。根据数据集中给定的因素，预测房屋价格。
  - 数据与数据说明见前文链接所提供的文件。
- 需提交文件
  - 项目源代码
  - 报告，其内容包括但不限于：问题描述，数据集说明，方法介绍，关键代码细节、运行截图、实验结果（回归问题指标）、结果分析等
  - 幻灯片，用于展示完成任务的主要步骤及方法描述
  - 视频：对实现方法及运行结果进行演示。

项目说明——Titanic Survival

- 项目介绍
  - 使用真实的泰坦尼克号数据集，来预测某人是否会在泰坦尼克号船中幸存下来。
  - 数据与数据说明见前文链接所提供的文件。
- 需提交文件
  - 项目源代码
  - 报告，其内容包括但不限于：问题描述，数据集说明，方法介绍，关键代码细节、运行截图、实验结果（二分类问题指标）、结果分析等
  - 幻灯片，用于展示完成任务的主要步骤及方法描述
  - 视频：对实现方法及运行结果进行演示。

项目说明—— Wine Quality

- 项目介绍
  - 通过葡萄酒的化学物质含量预测葡萄酒的质量。
  - 数据与数据说明见前文链接所提供的文件。
- 需提交文件
  - 项目源代码
  - 报告，其内容包括但不限于：问题描述，数据集说明，方法介绍，关键代码细节、运行截图、实验结果（回归问题指标）、结果分析等
  - 幻灯片，用于展示完成任务的主要步骤及方法描述
  - 视频：对实现方法及运行结果进行演示。

项目说明——Challenging

- 明确问题
- 数据分析
- 方法选择
- 评价指标

项目说明——明确问题

- 定义：给定一个信息的标题、出处、相关链接以及相关评论，尝试判别信息真伪。
- 输入：信息来源、标题、相关超链接、评论
- 输出：真伪标签（0: 消息为真，1: 消息为假）

项目说明——数据分析

- 数据获取
  - https://github.com/yaqingwang/WeFEND-AAAI20
  - 随本文件同URL提供
  - 只使用有标签数据
- 数据读取
  - 文件格式为csv格式
  - 可以使用Python自带的文件读取方式，手动分列
  - 可以使用Pandas库进行csv文件读取
  - 文件读取代码可以参考上文提及的git仓库中代码
- 读取代码
  - with open(filename, ‘r’, encoding=‘utf’) as f:
  - Import pandas as pd; dataset = pd.read_csv(filename)

项目说明——数据简要展示

	Ofiicial Account Name	Title
0	环球人物	中国反腐风刮到阿根廷，这个美到让人瘫痪的女总统，因为8个本子摊上大事了
1	西湖之声	腾讯为《如懿传》道歉？这部3亿大剧上映第一天遭网友狂吐槽：愣是拍成村头恋曲...
2	厦门晚报	顺风车司机奸杀20岁女乘客，落网视频曝光！滴滴道歉...
3	腾讯娱乐	偶遇鹿晗关晓彤旅行过七夕，小情侣是真滴甜...
4	腾讯娱乐	赵丽颖和冯绍峰即将公布恋情？网友：腥不腥没区别啊

	News Url	Image Url
0	http://mp.weixin.qq.com/s?__biz=MTAzNDI4MDc2MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/hpcO6kWnPm6cX3M...
1	http://mp.weixin.qq.com/s?__biz=MTA2Mjk0MTE2MA...	http://mmbiz.qpic.cn/mmbiz_jpg/vQCGoQzHAbaAXRr...
2	http://mp.weixin.qq.com/s?__biz=MTA3NzI1Mzg4MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/TxqQX9BtmOMpwDZ...
3	http://mp.weixin.qq.com/s?__biz=MTA5NTIzNDE2MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/9Ju9PZ1NxhfkIZ3...
4	http://mp.weixin.qq.com/s?__biz=MTA5NTIzNDE2MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/9Ju9PZ1NxhdTkXb...

	Report Content	label
0	[内容不符]	0
1	[满口胡言]	0
2	[？]	0
3	[领个屁证，过你妹的七夕，几天前的图在今天拿来博眼球]	0
4	[事件不实。]	0

项目说明——传统方法

- 特征工程
  - 目的是最大限度地从原始数据中提取特征以供算法和模型使用
  - 文本预处理
  - 特征提取
- 分类器
  - 贝叶斯分类器
  - 支持向量机
  - 随机森林

项目说明——深度学习方法

- 数据向量化
  - 将文本转化为深度学习模型所需的数值
  - 先进行文本分词操作（可使用分词工具jieba）
  - 后使用预训练的向量将词语转化为数值
- 数据集划分
  - 所提供数据集已经进行划分

项目说明——深度学习方法

- 模型创建
  - 使用PyTorch或者TensorFlow深度学习框架创建模型
  - 可直接进行内置的简单模型调用和模型搭建（LSTM，GRU，CNN，MLP）
- 模型训练
  - 将数据分成若干批次(batch)，按照批次送入模型
  - 在batch训练之后计算loss
  - 按照使用框架的api使用反向梯度回传降低loss
  - 重复上述步骤，直到训练结束

项目说明——深度学习方法

- 预训练模型
  - 可以使用transformers库，进行预训练模型的调用（Bert）。
  - 使用transformers的api进行数据加载、分词、训练。

项目说明——评价指标

- 分类评价指标
  - Accuracy
  - Precision
  - Recall
  - F1
  - AUC

项目说明——参考资料

- PyTorch: https://pytorch.org/
- TensorFlow: https://www.tensorflow.org/
- Jieba: https://github.com/fxsjy/jieba/
- Transformers: https://huggingface.co/transformers/
- Scikit-learn: https://scikit-learn.org/stable/

项目说明——参考资料

- 机器学习 周志华
  清华大学出版社 ISBN 9787302423287
- 自然语言处理入门 何晗
  人民邮电出版社 ISBN 9787115519764
- NLP汉语自然语言处理原理与实践 郑捷
  电子工业出版社 ISBN 9787121307652