冼永裕 广东广州 | 15915790273 | wentby719@gmail.com

专业技能

- **Languages:** Python, Java, SQL
- **Model:** Logistic Regression, Linear Regression, Decision Tree, SVM, KNN, K-means
- **Tools:** Scikit-Learn, Pandas, Numpy, matplotlib, MongoDB, MySQL, lifelines

教育背景

- 格拉斯哥大学，数据科学-硕士，Degree with merit 2019.09-2020.12
- 广东外语外贸大学，计算机科学与技术-本科 2014.09-2018.06

研究经历

癌症进化动力特征与生存率关系研究 2020.06-2020.09

- 分析 GDSC 中癌症患者的簇类别特征与生存率的关系，为医学界在癌症生存率的预后诊断中提供帮助。
- 协助医学生提取癌细胞进化动力特征来构建簇类特征，用生存分析研究不同簇类的患者生存率的差异性。
- 使用 KM 模型和 COX 模型分析簇类间的患者生存率关系，结果显示在 OV，SKCM，GBM，PAAD 癌症中，不同簇类患者的生存率均具有差异性($P<0.05$)。在可视化结果中，簇类 3 的患者生存率明显低于其他患者，结论表明簇类别特征可作为特定癌症患者生存率的预后指标。

实习经历

广州宸祺出行科技有限公司 数据实习生 2020.11-2020.02.25

- 使用 Davinci 平台对公司用户运营部的客户数据需求进行视图可视化，使用 PostgreSQL 对需求的数据进行清洗、整理、筛选。
- 对公司的客户数据库中的每个数据表进行字段整理，梳理数据库结构并汇总到表格。
- 协助用户运营部进行客户数据分析与挖掘，并绘制每周周报。

项目经历

Reddit 网站内容类别预测 2020.03-2020.04

- 设计文本语言分类模型，实现对 Reddit 博文内容和用户言论自动分类，帮助网站规范管理社区言论。
- 从数据集中提取分类信息如博客标题和内容，使用 spacy 对原数据进行预处理，并构建词向量特征。
- 对比 LR，SVM，决策树多个模型，基于 bigram ，对模型正则化 C 值和调整最大特征量，以 TFIDF 为词向量的 LR 模型预测博文类别的准确率最高，达到 72%。
- 使用 LR 回归模型，设计评论深度，回复 id 一致性等特征，预测网友言论类别，最终模型准确率提升至 60%。

Twitter 情绪检测分析 2020.02-2020.03

- 设计文本情绪分析系统对 Twitter 内容进行分类，帮助公司了解用户体验和进行舆论分析。
- 爬取数千条带情绪标签的推文，使用 NLTK 对数据预处理，初步归入 6 种情绪类，使用 MongoDB 数据库存储。
- 采用 NRC 词典策略对推文的每个词进行统计得出每个类的相应分数，最高分数的类作为推文的最终情绪类。
- 取 20%的数据结果进行众包，情绪集分类准确率均在 70%以上，比仅用情绪标签分类的数据 (48%) 提高了 22%。

CMT 共享自行车管理系统 2019.09-2019.11

- 与团队使用 python 开发设计一款共享单车服务系统，实现用户端租还功能和后台数据管理和可视化的功能。
- 使用 Tkinter 作为 GUI 框架，实现前端注册、租还、管理等功能和后端城市区域、人员管理等功能的界面呈现。
- 使用 pandastable 将后台数据可视化到操作界面上，实现基于 GUI 界面上城市状态、位置，付费等的修改功能。