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, lifelines

教育背景

- 格拉斯哥大学，数据科学-硕士，Degree with merit 2019-2021
- 广东外语外贸大学，计算机科学与技术-本科 2014-2018

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($P<0.05$)，尤其在 SKCM 达到 $P<0.005$，表明该特征能明显区分患者的生存情况。
- 使用 COX 模型多变量分析，排除协变量如年龄、临床阶段的影响，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。

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，对 Reddit 博文内容和网友言论进行分类，实现博客板块和用户言论的自动分类，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

- 基于 NRC 词典情感策略设计情绪分析系统对 Twitter 内容进行分类，帮助公司了解用户体验和进行舆论分析。
- 爬取数千条带有情绪词标签的推文并数据预处理，使用 NLTK 包进行 tokenize，stemming，根据情绪词标签初步分类为 6 大基本情绪，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 NRC 处理后的数据进行 crowdsourcing，每个情绪集合分类准确率基本在 70%以上，相比仅用情绪标签分类的数据（48%），提高了 22%。

Terrier 检索系统 2019.11-2019.12

- 设计文档排序 LTR 模型，在海量文档数据集中检索查阅内容相关的文档，并分析检索系统的效率。
- 基于 PL2 函数模型的排序结果，设计文档相关性新特征如查询词在文档中的最小距离，平均距离，使用 LTR 模型对文档重新排序。
- MAP 作为评估指标，结果显示在所有测试查询语句中，LTR 模型的 MAP 高于 PL2 模型 (23%)，准确率达到 61%。