杨睿

热爱数据与编程，熟练使用 Python，C#，较强的数据挖掘功底

数量经济学 • 上海对外经贸大学

男 | 23岁 | 硕士 | 应届毕业生 | 上海

18818231378 | yangruipis@163.com

实习经历

上海通金投资有限公司（通联数据旗下私募）

量化研究员 2017.04-2017.11

● 公开策略的复现与效果评价，如兴业证券“水晶球”期权择时策略、华泰金工的风险收益一致性策略等。

● 股指期货跨期套利策略开发、调试以及模拟盘和实盘对接(ctp 接口)，策略在100万实盘下稳定运行，并且在第一个月获得了12%的年化收益，后期因为合约换季造成较大亏损难以解决，策略暂时停止。

● ctp 接口 python 版本(https://github.com/lovelylain/pyctp）的封装与测试。使用户可以无需学习 C++，而是通过简单的 python 语法进行 CTA 策略的开发与模拟盘对接，现已推广至整个投研部门。

● 期权波动率曲面相关的择时策略研究

Kantar Media CIC(中国领先的网络口碑研究和咨询公司)

数据挖掘与数据分析 2016.03-2016.10

1. 数据采集：独立完成 Facebook 数据爬取(python, C#)，为公司每笔项目每年省下约四万元；抓取手机 APP 端数据如小红书、In、Nice、Lofter 等(WireShark 抓包，python 处理)；其他静态、动态网页抓取。
2. 数据清洗与数据预处理：
3. 熟练对接公司数据库，通过朴素贝叶斯等算法对海量电商数据去除非自然人贴；
4. 公司自己的分词器开发，C#界面，python 底层，调用 Jieba 内核，很好的对接搜狗细胞词库，并且能够在分词后直接绘制想要的词云图。
5. 数据分析：
6. 文本挖掘：根据公司的业务需求，独立完成文本挖掘软件开发（开发周期两个月， C#，本地运行，大大减小了公司服务器的压力，在公司内部被广泛使用）
7. 微博低质量粉识别器（python, SVM 算法）

8. 情感分析：利用朴素贝叶斯算法对帖子情感度进行分析，并在python的SnowNLP自然语言处理库基础上进行大量优化（对接公司语料库，去除无关词，拉普拉斯平滑等等）。
9. 微博账号影响力的评价模型建立：主成分分析，统计建模。

Paypal

人力资源分析师助理　　2015.06-2015.12

对每周各个客服部门接电话的数量、时长、客户评价等数据进行整理与汇总，制作汇报PPT并实现Excel与PPT的自动化的更新。

教育经历

上海对外经贸大学

硕士 · 数量经济学　　2019年毕业(预计)

上海对外经贸大学

本科 · 统计学　　2016年毕业

项目经验

信贷违约预测模型

特征工程、非平衡数据处理与模型构建　　2018.01-2018.02

根据贷款机构的放贷数据，建立信用评估模型，用以预测单个贷款者违约的风险，从而降低授信成本。该项目主要问题有：
1. 数据不平衡，训练集中违约样本数目要远小于未违约样本数目，这将导致分类算法极易将其全判为未违约，从而得到较高的准确率。我们在诸多抽样算法中筛选后，选择了Smote + Temok的ensemble方法进行处理。
2. 模型主要目的是降低授信成本，我们通过引入全样本的违约损失矩阵，将其加入优化方程，建立了一个Cost-Sensitive Logistic Regression模型(ICMLA 2014)进行研究，并作为stacking的一个子模型。

最终得到的AUC值为0.861，相对较好。

一个简单的机器学习常用算法实现

所有　　2017.12-2018.01

1. 特征工程，主要是Filter方法进行特征选择
2. 模型评价，包括一些评价得分(metrics)，分类结果作图以及交叉验证函数
3. 分类算法，包括所有分类算法的接口规范，以及具体的分类算法实现，如（Knn, Logistic, Naive Bayes, CART, Random Forest, SVM, BP network）
4. 聚类算法， 包括K均值聚类和层次聚类

GitHub地址（https://github.com/Yangruipis/simpleML）

文本挖掘工具开发

所有　　2016.04-2016.08

针对公司数据分析业务，开发符合业务流程与逻辑的文本挖掘软件，其中涉及知识主要有： ● 四层逻辑匹配（关键词，排除关键词，near rule(关键词附近必须有的词)，except near rule） ● 表达式解析：near rule 是表达式形式，有"+","-","_"表示且，非，或的逻辑关系，且有括号判断优先级，需要进行完整无误的解析 ● 产品树与属性树解析：产品树包含了每个产品以及其多层子产品，而属性树包含了每个属性以及多层子属性。解析成树状结构并对所有文本进行匹配后，可将每个文本落到描述哪一个产品的什么属性之中。 ● 多种统计结果以及报表输出 ● 多线程窗体界面编写 语言：版本一：python(后台)+C#(界面)，版本二：C#

电商数据情感分析算法研究

电商数据情感分析算法研究　　2016.06-2016.07

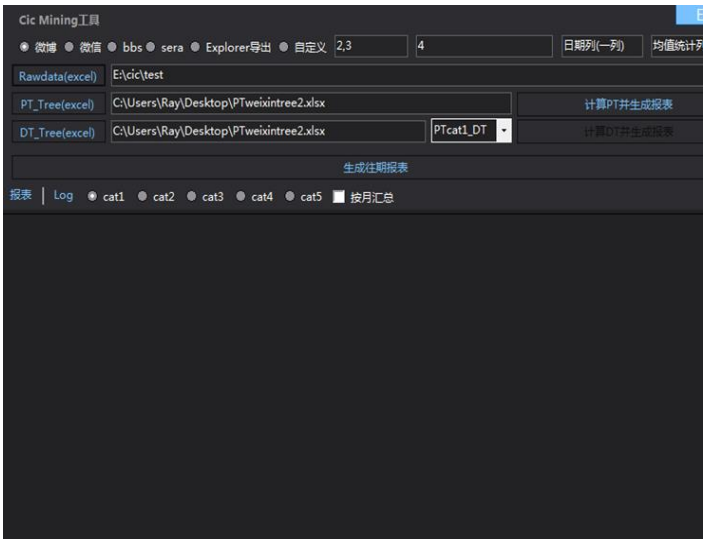
运用公司语料库（包括正负面词库与训练集），对每一个句子（电商买家评价，微博评论等等）进行 1. 情感正负面的判别 1. 情感强弱打分 1. 意见挖掘（将情感落在句子里具体某个品牌或是某个属性上） 开发语言：python 与 C#均进行了开发，同时 c#制作了给分析师使用的界面 使用算法：对比下来使用了朴素贝叶斯算法，（相对支持向量机算法参数易于调整），目前在公司被广泛使用

新型贝叶斯统计软件研发

所有经典统计功能开发，网络数据爬取功能，MCMC 算法　　2015.01-2015.11

● 本项目致力于开发基于贝叶斯统计的、能够高效获取和处理大数据的新型统计软件。 ● 担任所有统计功能的开发和优化工作，包括基本作图、空间作图、描述性统计、区间估计、假设检验、方差分析、相关分析、回归分析、时间序列、灰色预测。MCMC 抽样算法：ARMS 开发 ● 担任所有数据爬取功能与软件对接功能，包括股票基金等数据(新浪网)、宏观经济数据(中国统计局，动态网页)

作品展示



文本挖掘工具作品，目前整个公司(Kantar Media CIC)分析师都在使用此工具。

自我描述

● 参加多次数学建模大赛且均有获奖，有大量数学建模和统计建模经验。熟练掌握机器学习相关算法及原理，熟练使用 Mysql 等数据库，熟悉深度学习平台如 tensorflow， Keras 等，熟悉 XGBoost，了解 Spark、Docker、
● 学习能力强，擅长自学，也有较强团队精神。
● 在校成绩优异，获校十佳学术标兵、上海市优秀毕业生、企业奖学金一等奖等荣誉
● 个人技术博客：http://www.cnblogs.com/yangruiGB2312/
● 极客精神，emacs 使用者，长期使用 Linux(Ubuntu)系统，且拥有个人 CentOS 服务器使用经验（自己租的 VPS）

期望工作

数据挖掘 / 机器学习 ， 实习 ， 上海 ， 2k-5k

技能评价

C#

掌握

R 语言

熟悉

Python

精通

SQL

熟悉

荣誉与奖项

● 2017年“华为杯”研究生建模大赛全国二等奖
● 2016年中国大数据创新行动大赛第一名
● 2016年上海市优秀毕业生
● 2015年美国大学生数学建模二等奖