# 段凌霄

## 算法工程师

15712974785 duanlingxiao1219@sina.com 北京 西城区黄寺大街 23 号 askwitionary.github.io

## 教育背景

### 圣路易斯华盛顿大学

**计算机科学与技术 硕士**

2015.08 – 2017.06 密苏里，美国

- 数据挖掘与机器学习方向
- 对学校医学院数据库中 Alzheimer's Disease 病人的 DNA 数据进行挖掘

### 惠提尔学院

**数学与应用数学 学士**

2010.09 – 2014.05 加利福尼亚，美国

- 辅修物理与科学数据计算，选修化学与五种人文社科。2011 曾获 Dean's List 提名
- 连续四年获得半额奖学金（每年 22500 美金）

## 工作经历

### 数据分析师/数据部负责人

**北京量码博信数据技术有限公司**

2018.12 – 至今 北京

- 带领六人小团队一年时间从 Python 零基础到可以解决实际问题
- 支持心血管疾病国家重点实验室科研工作，整理，清洗全国收集来的数据。包括但不限于入库，OCR，数据转换，问题数据" 康复"
- 参加"中国心电智能大赛"，利用数字心电数据和深度学习做多标签分类器，实现心电图的初步自动诊断分析。并在初赛取得全国第 9，复赛取得全国第 33 的成绩
- 对已有数据进行统计分析，指标运算，为类似 BI 系统做后端
- 对医院存档的纸质心电图进行清理与转换（图像到 12 导联时间序列）便于患者病历，检查数据的保存与处理

### 算法研究员

**慧影医疗科技（北京）有限公司**

2018.11 – 2018.11 北京

- 训练简单 X 光片部位分类模型（手、脚、手腕、脚踝等分类）模型达到 95% 以上验证准确率

### 算法工程师

**北京三点一刻科技有限公司**

2018.05 – 2018.09 北京

- 研发并调试平台项目、甲乙方的多维征信、匹配、推荐逻辑与算法
- 以 Tensorflow 的 CNN 为基础，搭建图像识别模型，帮助爬虫与自动化测试同事解决自动识别图形验证码的问题
- 图片文字转文本的优化、聊天记录截屏中 emoji 识别
- 探索、学习各个相关开源项目的代码和文档并基于可能用到的场景做 demo，如 LSTM 聊天交互、自动写稿、趋势预测；决策树、随机森林分类；CNN 文本、图像分类；NLP 相关：分词、句法分析、词向量、关键词提取等

## 个人期望

**期望行业**

兴趣：医疗 专业：大数据

期望：医疗大数据行业

**期望职位**

机器学习; 人工智能; 数据挖掘 (大数据相关)

算法工程师、研究员

## 编程及相关语言

- Python ●●●●●
- JAVA ●●●○○
- SQL ●●●●●
- LaTeX ●●●●●
- Linux/Shell ●●●●●
- MediaWiki ●●●●○

## 常用软件及工具

- PyCharm/IntelliJ ●●●●●
- Anaconda ●●●●●
- MS Office ●●●●●
- Mathematica ●●●●○
- Maple ●●●●○
- MATLAB ●●●○○
- IDL ●●●○○
- Blender ●●●○○
- Photoshop ●●●○○
- Unity ●●●○○

## 自然语言

- 英文 ●●●●●
- 西班牙语 ●○○○○

## 数据挖掘与机器学习平台及算法

Tensorflow SK-Learn Caffe CUDA cuDNN MySQL PostgreSQL HBase Hadoop Keras

决策树 随机森林 LSTM 卷积神经网络 GBDT 线性、非线性规划 自然语言处理相关

## 项目经历

### 基层高血压质控管理指标计算、核对与可视化

**北京量码博信数据技术有限公司 数据分析师**

2019.06 – 至今　　北京

根据研究中心项目组的要求对云南全省与陕西韩城市的基层高血压数据进行清理、整理。并对八大考核指标共十多个细分指标进行高效率运算

- 清理数据。由于数据采集系统的缺陷以及部分诊所医生的不当操作，可能造成数量可观的数据由于不符合标准而难以使用。我和我的小组在这些问题数据中积极寻找规律，恢复问题数据五万余条
- 在与研究中心数据部交接、核对过程中找到数十条逻辑问题、程序 bug 等影响指标准确度的问题
- 在开发算法的同时研究指标运算内在的本质逻辑，优化程序结构，提高运算效率。中心数据部需要处理、运算一天多的数据经过优化只需不到五分钟
- 开发类似 BI 系统的数据管理系统的后台接口以及指标可视化

### 中国心电智能识别大赛

**北京量码博信数据技术有限公司 算法研究员**

2019.01 – 2019.07　　北京

与两位协和医院心内科医生组成小组，利用主办方提供的心电数据，通过深度学习的方法设计并训练了两个分类模型

- 初赛为二分类，将心电信号分成正常与异常。达到了 90.7% 的验证准确率，全国排名第 9
- 复赛为多标签 9 分类，将不定长度的心电信号标注为正常或 8 种心电异常的一种或多种。达到了 83.3% 的验证准确率，全国排名第 33。与前 10 名相差不到 1.7% 的验证准确率，竞争十分激烈
- 学习了心电的基础知识
- 尝试使用了多种深度学习架构，包括但不限于 CNN、LSTM、CNN+LSTM
- 与两位医生探讨并尝试了数十种特征提取方法

### 颈动脉超声数据录入与清理

**北京量码博信数据技术有限公司 算法研究员**

2019.09 – 至今　　北京

整理、读取、统一化中心提供的各省市县多个医院上报的颈动脉超声报告的数据

- 读取 doc, docx, xls, xlsx, pdf, 图片等各种格式各种样式的颈动脉超声报告
- 提取标准化信息
- 图片 OCR 文本识别
- 数据清理与纠错

### 图片验证码识别

**北京三点一刻科技有限公司 算法工程师**

2018.06 – 2018.07　　北京

根据各个以图片为验证码的网站，包括数字、大小写字母识别；鼠标滑块；按顺序选字等图像人机验证的生成逻辑，实现对人机验证的突破

- 分析验证码生成逻辑
- 仿制验证码生成器从而提供无限量训练数据
- 根据不同验证码特点设定突破步骤
- 搭建模型并进行训练
- 根据训练过程和测试结果调参重复训练
- 将正确率合格的模型封装上线
- 对本平台网站图片识别率达 99%
- 对外部平台的图片验证码识别率根据复杂程度从 50% ～ 85%

### 数据分析与征信

**北京三点一刻科技有限公司 算法工程师**

2018.05 – 2018.08　　北京

- 根据平台数据以及爬取的外部平台数据实现对营销需求、甲方、乙方的多维度评分。评分系统通过比较分类中其他项目归一化，并实时更新，基本不用维护。
- 评论文本打标签、提取句子主干、分类
- 需求统一化分类
- 分析各分类下需求描述的关键信息点，使用 TF-IDF 和 TextRank 为基础的模型对需求描述中非结构化信息点进行分析评价，并估算描述信息完整度
- 对每个用户（公司/个人）根据百度精确搜索获得的条目数确定其品牌、影响力
- 根据相关信息，使用统计学方法获取其他维度（共 13 个）评分

### 基于深度学习对图像、视频中物体进行识别与追踪

**个人项目**

2018.01 – 2018.04　　加利福尼亚，美国

- 基于 Google 开源项目，对图片及视频中出现的常见物体进行识别及追踪
- （视频）通过连续位置追踪确定目标并对其做简单行为判定（对视频质量要求高，运算成本大）

### 对 Alzheimer's Disease 病人基因与进行数据挖掘

**圣路易斯华盛顿大学医学院/计算机学院 个人项目**

2016.01 – 2016.12　　密苏里，美国

- 使用华盛顿圣路易斯大学医学院老年痴呆症病人以及对照组正常人的基因片段进行数据挖掘，对病人与非病人的基因进行分类
- 对较大规模数据进行清洗与标准化
- 对 8000 多个基因维度进行维度分析，缩小可能为致病基因的范围
- 学习并练习了数据挖掘相关的基础方法，包括各种回归分析、决策树、临近算法、SVM 等

### 生物医学图像处理几何运算

**圣路易斯华盛顿大学计算机学院/医学院 个人项目**

2015.08 – 2015.12　　密苏里，美国

- 使用 Mathematica 对学校医学院的图片素材进行分析与处理，主要达到的功能包括精确寻找 CT 骨骼结构和寻找 MRI 肿瘤位置与轮廓等
- 使用 Mathematica 对骨骼、器官进行 3D 建模并进行以下处理：寻找主体、体积计算、表面积计算、模型锐化与平滑化、模型化简等

### 吃豆小人（Pacman）AI 设计

**圣路易斯华盛顿大学 个人项目**

2015.08 – 2015.12　　密苏里，美国

- 使用 Python 设计了吃豆人 AI，使用启发法，强化学习法和 Bayes Nets 等完成了对基本吃豆，躲鬼吃豆的快速导航 AI