# 吴子俊

资深算法工程师

性别:男 电话: 13681209353 现所在地: 北京

开始工作时间: 2015 邮箱: puppyapplewzj@126.com

政治面貌: 共青团员 民族: 汉族



## 教育经历

2012.09 - 2015.07

北京航空航天大学

控制工程 | 硕士

2013.09 - 2014.10

巴黎中央理工大学(法国TOP2)

系统工程 | 硕士(双学位)

2008.09 - 2012.06 北京航空航天大学

数学与应用数学 | 本科

#### **一** 工作经历

北京字节跳动 2021.01 - 至今

资深算法工程师

#### 工作描述:

- 1. 负责百科线索生产平台项目,整体规划平台架构、数据通路、策略/算法;实现线索生产及策略线索迭代的自动 化,线索生产整体提效50%+;
- 2. 推动百科线索生产相关基础策略/算法能力沉淀,提高可复用性(文本相关性、语义重复、事件抽取等);
- 3. 负责公司内部企业知识图谱的搭建,推动并参与工程(图谱构建)和算法(实体对齐、实体去重、属性抽取等) 的技术探索及落地,为公司内部多个B端业务提供稳定数据服务和下游策略/算法支持。
- 4. 指导安排初级工程师的工作。

2019.09 - 2021.01

北京搜狗网络技术有限公司

#### 高级算法研究员

#### 工作描述:

- 1.作为技术负责人,带领团队从0到1构建智能对话平台(内部创业项目), 完成平台整体算法架构设计及工程落 地(包括但不限于对话引擎、文本分析、语音交互等模块),半年时间实现平台上线及商用;
- 2.负责文本聚类、意图识别/语义匹配、文本纠错、文本增强等NLP算法模块的算法研究与落地,为平台提供服务:
- 3.参与并主导对话平台各模块的产品设计(对话管理、后台算法、前端交互等)。
- 4.接手信息流业务NLP相关的算法方案设计与调研。

2018.04 - 2019.09

因果树网络科技有限公司(B轮)

NLP算法工程师

工作描述:

- 1. 作为核心成员参与公司产业互联网人工智能大数据平台的算法探索及研发工作:
- 2. 基于千万级企业文本数据和NLP算法构建行业/产业标签体系,独立负责企业信息库、一级市场企业-投资机构间投融资匹配系统等核心平台组件的算法相关模块实现;
- 3. 对接来自B端客户的新业务场景需求,提供NLP算法模型解决方案,并与数据及工程部门协作进行落地,集成到平台功能模块。

#### 2016.12 - 2018.04

人人行科技股份有限公司(B轮)

#### 算法工程师

工作描述:

- 1. 作为风控算法项目负责人带领成员对接业务,进行算法解决方案设计和落地:
- 2. 基于亿级用户数据从头构建风控用户画像,同时基于机器学习/深度学习构建用户信用评估模型及反欺诈模型,降低平台违约率约50%(10%=>5%),对欺诈用户或团体召回率达到53%。

#### 2015.09 - 2016.12

法国标致雪铁龙上海研发中心

项目管理&数据分析

工作描述:

- 1. 负责某新车型整车研发阶段的各类数据分析,总结性报告输出;
- 2. 负责整车工业化总装阶段的项目管理与协调。

## ■ 项目经历

2021.01 - 至今 百科线索生产平台

策略平台架构、项目管理、文本质量算法、事件挖掘、实体链接

#### 项目Leader,整体负责百科产品的线索生产平台搭建,应用策略/算法提效内容生产及提高内容质量:

- 1. 设计百科词条线索生产平台技术架构并推动工程落地,实现线索生产策略迭代的流程与数据自动化(测试、回溯、版本管理、上下线),支持策略/算法的细粒度复用,大幅减少人工介入成本,提效词条生产;
- 2. 主导百科词条质量类线索生产的算法策略开发,包括但不限于语义重复识别、文本流畅度判定、文本相关性判断等词条质量模型,通过机器筛选节省50%以上人工审核成本;
- 3. 探索事件挖掘及实体链接算法,基于全网文本信息提取实体、事件线索关联百科实体,转换为有效生产线索以提升运营效率,提升百科产品时效性。
- 4. 搭建数据追踪看板,整体追踪线索生产平台策略准召率、线索覆盖率,以驱动策略/算法迭代。

知识图谱、实体对齐、属性/事件抽取

#### 项目Leader,整体负责企业知识图谱构建并作为中台服务支持公司内部业务:

- 1. 搭建生产引擎、构建引擎模块组成通用知识图谱构建平台,支持多源数据融合、属性择优算子;
- 2. 梳理多个企业源数据(工商信息、行业标签、投融资、品牌集团),定制实体对齐策略算法并基于构建平台实现企业知识图谱实时构建服务:
- 3. 借助公司站内海量新闻图文数据构建企业属性(别名、标签)及事件抽取(品牌新闻、投融资信息)、品牌集团 挖掘、企业-产品关联挖掘等算法服务,用于图谱SPO扩充以及舆情分析等场景;
- 4. 其他基于图谱的下游算法场景支持(相似企业、行业关联等)。

2019.09 - 2021.01 智能对话平台

NLP、平台/产品设计、系统架构、项目管理

项目Leader, 带领团队从0到1搭建智能对话平台, 半年实现平台上线并在ToB业务中商用:

- 1. 调研市场产品及技术架构,以NLP算法为驱动,结合公司自身ToB业务场景需求制定产品功能设计方案(功能模
- 块,用户交互)及后台技术架构(算法服务、前后端交互、数据管理),并制定开发计划;
- 2. 构建通用文本聚类模块,利用无监督对话语料辅助新业务场景下冷启动,节省80%的人工标注成本;
- 3. 实现基于文本分类的意图分类模块,并引入小样本学习,实现对全新业务场景低资源情况下的快速适配;
- 4. 基于ES、Sentence Bert以及文本匹配模型构建QA场景的召回+精排模块,准确率达到97%;
- 5. 结合平台ASR语音识别的badcase分析,探索基于Bert+汉字+拼音的多场景通用文本纠错模型并上线;
- 6. 基于预训练模型和大规模相似句语料探索文本生成模型,实现本数据增强,大幅缓解新业务场景数据冷启动问题。

2019.02 - 2019.09 投融资匹配系统

协同过滤、图算法、GBDT

#### 项目Leader,负责一级市场投融资匹配系统的算法方案设计及实施:

- 1. 基于协同过滤、Graph embedding等算法进行企业和业务标签Embedding生成;
- 2. 基于Personalized-PageRank、Path-Sim计算节点关联度进行可解释性召回;与金融分析师协作构造企业业务特征;
- 3. 结合Embedding特征和其他业务特征训练GBDT模型对粗召回结果进行精排序,投融资匹配准确率达到68%+;
- 4. 半年内实现系统上线,针对新三板股转系统项目进行定制优化并通过POC招标。

2018.04 - 2019.02 企业信息库

文本分类、命名实体识别

#### 独立负责企业信息库构建中的NLP算法研究及实现:

- 1. 与一级市场金融分析师协作,设计构建TMT行业多层级标签体系;
- 2. 基于SimHash算法对千万级文本(企业描述、企业新闻等)进行存量去重,同时构建增量实时去重通路;
- 3. 使用FastText、TextCNN、Bert等算法及业务标注数据构建企业多层级多标签分类模型,实现94%线上准确率;
- 4. 对企业新闻事件文本进行命名实体识别,结合关系抽取结果以及分类标签构建企业知识图谱;
- 5. 基于Bert进行无监督的短文本关键词排序,优化企业搜索结果。

2017.10 - 2018.04 风控反欺诈模型

异常检测、社区发现、标签传播

- 1. 对接风控业务部门及公安系统,汇总反欺诈业务场景和案底,构建量化特征指标;
- 2. 利用画像特征及行为特征,基于IForest、GMM等无监督算法构建异常检测模型识别异常用户;同时挖掘用户授权通讯录、APP关注关系、设备IP等多维度关系数据构建用户网络,使用社区发现、标签传播算法进行欺诈团体挖掘;
- 3. 异常用户/团体整体召回率达到53%,为风控业务部门提供有力决策依据,有效降低了平台欺诈风险。

2016.12 - 2017.10

用户画像及信用评分系统

用户画像, GBDT, DNN

#### 项目Leader, 带领团队从0到1构建了基于算法的信用评分系统:

- 1. 基于APP端用户基本信息、交易记录、非交易行为等9个维度的TB级别数据构建风控用户画像:
- 2. 基于用户授权的通讯录信息进行非结构化文本挖掘,丰富用户画像及特征;基于AutoEncoder+Bi-LSTM对用户行为序列进行建模,提取用户动态行为特征;
- 3. 组合用户画像的动静态特征,基于XGBoost实现违约概率预测模型并上线;
- 4. 以预测模型为基础与业务方共同建立了20+风控规则引擎方案,通过A/B Test、灰度等手段进行追踪验证,部署为产品级线上应用,最终平台违约率降低约50%(10%=>5%);参与设计模型后期监控方案,生成数据报表。

#### 学习能力:

具有较强的自我学习能力,善于利用各类资源快速地了解并熟悉全新领域的知识,抓住核心部分,并进行应用实践。

### 业务能力:

负责或参与过多个跨部门协作项目,具备了较强的团队协作能力和项目管理领导能力;善于理解业务场景,从中抽象、提炼核心问题,探索并实践解决方案。

#### 性格特点:

较为沉稳但不内向, 乐于交流, 喜欢富有挑战性的工作, 有较强的意志力和自制力。

## □ 相关技能

#### 专业技能:

熟悉Linux开发环境;熟练掌握Python语言;熟练使用Pytorch深度学习框架;熟练掌握Hive、PySpark等大数据相关应用工具;了解Tensorflow,Neo4j图数据库;了解Axure使用。

#### 管理能力:

具备多次项目及团队管理经验,能够合理进行任务拆解及工作分配,并通过有效的时间管理手段保证项目进度。

#### 其他:

英语CET-6; 法语B2。