

王倪东



电话: 15675879934 性别: 男 年龄: 26

邮箱: xesdiny@gmail.com GitHub: <https://github.com/xesdiny>

学历及方向

- 应届硕士研究生(湘潭大学 信息工程学院 计算机技术)
- 自然语言处理、数据挖掘

专业技能

- 精通 JAVA/Python
- 熟悉 kvm、Docker 等分布式虚拟容器
- 熟练使用 NLP 常用 SVM, DT, CRFs 算法实现标注抽取
- 擅长 Hadoop、Zookeeper、Kafka、Rabbitmq 工具
- 擅长分布式、高并发系统开发
- 熟悉 MySQL、Mongodb、HDFS 等数据库
- 了解图数据库 Neo4j 和分布式全文检索工具 ES

实践经历

➤ 2016.05.10~至今 湖南奇点创智数据科技有限公司

实践部门: 数据挖掘部

职位名称: 算法工程师

个人职责: 工作任职于湖南奇点创智数据科技有限公司, 从事大数据方向学习与研究, 在数据挖掘部做 NLP 方向算法研究与实现, 先后参与了关于医疗、金融等大数据研究工作。先后参与了智搜搜在线系统、湘雅放射科数据结构化项目, 数据采集平台, 授信文本分析平台, 金融知识图谱与智能问答, 集群虚拟化构建平台, 风险传导预警项目, 珠江人寿保险产品分析项目, 招商银行舆情预警金融科技项目。熟悉自然语言处理方向常用技术, 如分词、词性标注、命名实体识别, 关系抽取, 句法分析等; 熟悉信息抽取相关的算法和逻辑熟练操作 Debian GNU/Linux 系统, 熟悉 Hadoop 生态环境; 解决问题能力强。主要负责各个项目的核心 NLP 问题的算法实现以及部分系统搭建。

证明人: 周晴宇

证明人电话: 18711346405

项目经历

➤ 2018.04~至今 FinTech 舆情预警金融科技项目

项目描述: 与招商银行深度合作。基于网络爬虫搭建舆情采集, 分析以及预警系统。

设计了一套企业舆情分析与预警系统, 以实现: 企业互联网负面信息监控。实时监控网络空间中的目标企业舆情信息, 采集并提取其中的负面相关信息。企业舆情指标体系构建。针对舆情信息分析构建指标体系, 并从中抽取相应指标值, 细致化的定量分析舆情内容。企业舆情严重程度评级与预警。挖掘舆情事件范式, 根据舆情指标对舆情事件进行严重程度评级, 对满足范式与评级策略的舆情事件进行预警。

个人职责: 在深圳招商银行风险 IT 部驻场学习交流。

主要是公司简称生成模型实现及优化; 舆情数据噪声、数据遗漏分析处理;

辅助参与舆情分类到预警信号映射; 网页正文、时间抽取优化; 摘要抽取等。

关键技术: 简称生成; 数据处理; 数据抽取。

➤ 2017.09~2018.01 珠江人寿保险产品分析项目

项目描述: 与珠江人寿保险公司构建大数据项目平台建立, 围绕对舆情数据进行采集以及人格分析, 客户分类方案, 保单托管, 保险产品标签体系四大部分

个人职责: 从保险产品名称中提取机构简称、产品昵称、设计类型; 保险条款 PDF 解析; 搭建基于机器学习的中文分析平台引擎。引擎主要使用自然语言处理技术, 是机器从语言学的角度理解自然语言, 因此在过程中涉及到了语言学的一些基本任务如词法、句法、语法等。对中文分词、词性标注、命名实体识别等词法分析的基本任务, 分别采用 N-最短路径、隐马尔科夫、条件随机场等方法, 这些方法将任务转换为序列标注问题进行处理。依存句法分析是句法分析中浅层句法分析, 基于最大生成树的分析方法是以整句为最优依存结构的搜索单位, 以整句依存最优为目的, 算法是依赖词法分析的结果。

关键技术: 简称生成; PDF 解析; 序列标注。

➤ 2017.03~2018.04 大数据集群虚拟化平台

项目描述: 基于 kvm 与 docker 的计算机集群虚拟化平台。包含虚拟化节点操作, 各项分布式应用搭建, 集群的 HA (高可用), 硬件资源以及网络资源最大利用率算法, 管理虚拟化节点等功能。

个人职责：WEB 端虚拟机 VNC 管理功能集成；**VPN 管理功能：**虚拟机网络编辑；VPN 开放注册（管理员审核）；物理机主动上线（注册）；VPN 接入（使用账号密码认证，且提供证书）；虚拟机内存、CPU 和网络资源修改功能。

关键技术：动态负载均衡、虚拟化、openVpn。

➤ 2017.03~2017.08 风险传导预警项目

项目描述：基于行业知识图谱，构建面向 C 端的智能问答平台。包含实体关系抽取、用户问句理解、意图识别、知识推理、大规模知识图谱检索等功能。

个人职责：1.负责知识图谱架构设计。设计基于图数据库 Neo4j 和 ElasticSearch 的行业知识图谱管理系统，系统包含知识图谱构建、图谱索引和大规模图检索。

2.负责智能问答平台架构设计。设计基于知识图谱的智能问答平台，包含用户问句理解、意图识别、知识推理。

关键技术：自然语言处理、实体关系抽取、大规模图检索、分布式全文检索

➤ 2016.10~2017.12 湘雅医疗 CT 报告结构化

项目描述：为湘雅三医院医院的医疗合作项目。此医疗合作项目主要基于医院的医疗需求，结合医学影像、实验室检查、病理数据、临床数据和患者基本信息等五大数据源，采用统计分析、大数据挖掘算法等技术，为实现满足医患人员需要的多维查询、数据导出等功能，研发出一整套具有医疗信息查询、数据统计分析等功能的系统，实现对医疗数据的整理、开发和应用，大大提高医院的管理和服务质量。并将非结构化的文本纳入结构化数据的精准查询中，大大提升了非结构化数据的利用，进一步提高医院的行业竞争力。

个人职责： 1. CT 报告结构化，利用序列标注方法对医学实体进行标注；

2 系统开发，使用 python Flask+gunicorn 框架开发结构化 API 接口；

3 医疗数据清洗，将原始的 mysql 数据转换为 mongodb 的 document 数据存储。

关键技术：自然语言处理结构化、数据清洗

➤ 2016.01~2017.03 智搜搜企业全息画像&实时监控平台

项目描述：“智搜搜”是一个企业全息画像&实时监控平台，具有创新的危机管理理念、独创的关系搜索技术、跨界融合的互联网海量数据（金融、法律、政务等）等特点。作为一个年轻化、创新化的产品，“智搜搜”基于大数据分析，深度挖掘显性及隐性关系，构建目标企业的投融资关系图谱及商务关系图谱。为金融相关行业提供精准营销、客户群分析、风险管理、反欺诈、贷前信审、贷后管理等服务。

个人职责：

1. 对系统的后台数据多进程采集，利用 socket 协议建立连接完成模拟浏览器发送请求；

2. 对数据源的反封锁并设置采集，通过分析请求头以及相关参数和使用代理；

3. 后台采集系统的架构搭建，关于采集后台环境以及资源使用预警监控，之前使用 ganglia，后来采用 open-Falcon。

关键技术：多线程、反封锁、资源监控

➤ 2017.01~2017.03 文本分析平台

项目描述：基于大数据平台架构开发面向公司内部的自然语言处理服务。系统包含中文分词、词性标注、命名实体识别、文本分类、文本情感极性分析和文本结构化等组件。

个人职责：对中文分词、词性标注、命名实体识别等词法分析的基本任务，分别采用 N-最短路径、隐马尔科夫、条件随机场等方法，这些方法将任务转换为序列标注问题进行处理。

依存句法分析是句法分析中浅层句法分析，基于最大生成树的分析方法是整句为最优依存结构的搜索单位，以整句依存最优为目的，算法是依赖词法分析的结果的。

关键技术：词法分析、依存句法分析、标注、隐马尔科夫

➤ 2015.09~2016.02 舆情危机预警系统

项目描述：基于 Hadoop/Kafka/Storm 大数据平台的舆情危机预警系统。针对微博、各大论坛、微信等舆情源进行实时数据抓取和分析，为舆情危机管理提供决策支持。

个人职责：基于 Hadoop 平台的分布式网络爬虫设计与实现。利用 Hadoop 的分布式架构和 map/reduce 编程模型来实现大规模网页抓取，包含 URL 去重和周期抓取、网页数据下载和抽取解析、抓取源反爬虫应对策略和数据存储等。

关键技术：分布式爬虫、反爬虫应对策略、数据抽取、集群运维、文本聚类

论文/获奖经历

-
- 王倪东, 周维, 黄正宇. 基于双层条件随机场与规则推导的中文公司简称生成[J]. 中文信息学报, 编号 2018:-0017(已录用, 待刊出).
 - 周维, 王倪东等. 普通本专科学校教学科研仪器设备维修管理软件(软件著作权, 申请号:2015R11L362210, 已发表).
 - 周维, 王倪东等. 本科学双学位及辅修专业教务教学管理软件(软件著作权, 申请号:2016R11L129124, 已发表).