

XX

技术方向：自然语言处理算法工程师

电话：xxxxx

E-mail: xxxxx@126.com

## 教育背景

2008.09-2011.07	XXX	计算机科学与技术	工学硕士
2003.09-2007.07	XXX	应用数学	理学学士

## 专业技能

- 熟悉常用机器学习与深度学习模型与原理，如 LR, DT, MLP, RNN/LSTM/GRU, CNN 等
- 熟悉常用的自然语言处理技术，如词袋, TF-IDF, word2vec, fastText, TextRank, LDA 等
- 熟悉文本分类相关算法与模型
- 熟悉 Python，熟练使用 numpy, pandas 等模块，了解 TensorFlow、Pytorch、sklearn 等框架
- 具有扎实的数学基础，熟悉数据结构和常用算法
- 了解设计模式及常用代码重构技术，熟练使用 C++/Java 进行程序开发
- 具有良好的英语听说读写能力，熟练阅读英文技术资料

## 工作经历

xxxx.06--至今 XXX 软件公司

- 参与用户画像职业标签模型的构建
- 参与智能邮箱系统的设计与实现

xxxx.06--xxxx.05 北京 XXXX 技术股份有限公司

- 负责建筑设计软件数据库设计，数据库管理系统设计及实现
- 参与建筑设计软件建模模块的设计与实现

## 项目经验

- 项目名称：用户画像职业标签模型构建（多分类）  
开发工具：xgboost, fastText  
项目描述：基于用户信息和行为数据，判断用户职业，推荐 XX 与流量变现。  
责任描述：参与用户画像职业标签模型的构建，具体工作内容如下：
  1. 数据预处理：数据清理，归一化，离散化，中文分词与去停用词，生成词向量，特征提取
  2. 特征选择：基于树模型的特征重要度选择
  3. 模型选择：GBDT(xgboost), fastText
  4. 模型融合：概率等权重融合
  5. 模型评估：分别计算准确率，F1 值和 AUC 值
- 项目名称：智能邮箱系统设计与实现（过滤垃圾邮件）  
开发工具：Python, Java, C++, JavaScript, sklearn, Bag of Words, Naive Bayes, N-gram  
项目描述：支持主要邮件收发协议的中文版邮箱客户端。  
责任描述：参与 XXX 智能邮箱系统的设计与实现，负责垃圾邮件识别模型的构建，具体工作内容如下：
  1. 系统构建脚本编写
  2. 关键模块的设计和代码实现，包括会话，写邮件，通讯录，日历等模块
  3. 垃圾邮件识别和自动分类
  - 1) 邮件收集，构建正负样本
  - 2) 对数据做分词，去停用词处理
  - 3) 数据补齐或截断
  - 4) 基于朴素贝叶斯算法训练分类器
  - 5) 模型评估：AUC 约 86%
  - 6) 模型移植：保存模型为 pb 文件，移植到邮件客户端工程

- **项目名称: XXX 建筑设计软件数据库管理系统设计与实现**  
**开发工具:** MySQL, ODBC, C++, 设计模式, 代码重构  
**项目描述:** ARCH 建筑设计软件的数据库管理系统, 可以定制构件属性、数据及图形表示  
**责任描述:** 负责数据库设计、DBMS 的设计与实现和数据库角色和权限系统的设计与实现
- 

## 其它项目

- **项目名称: Sentiment Analysis on Movie Reviews(Kaggle)**  
**开发工具:** gensim, nltk, sklearn, word2vec, TF-IDF, SVM, TextCNN, TextRNN  
**项目描述:** 根据带情感标签的影评数据, 训练影评自动情感分析模型。  
**主要工作:** 基于影评数据训练自动情感分析模型, 具体工作内容如下:
  1. 读入数据, 删除 html 符号, 标点符号等, 去停用词
  2. 对数据做补齐或截断处理
  3. 模型选择:
    - bow/TF-IDF+SVM
    - word2vec+TextCNN
    - word2vec+TextRNN
  4. 模型融合: 概率加权融合
  5. 模型使用: 用训练好的模型对未标记数据进行自动情感分析
- **项目名称: 用户评论情感分析分类器**  
**开发工具:** TensorFlow, TextCNN, TextRNN, LSTM, GRU, fastText  
**项目描述:** 基于带类别标签的用户评论数据, 训练出文本分类模型。  
**主要工作:** 基于用户情感数据训练文本分类器, 具体工作内容如下:
  1. 数据预处理: 对数据做补齐或截断处理, 分词, 去停用词
  2. 模型选择:
    - FastText
    - TextRNN(GRU + FC)
    - TextCNN(Conv1 + Relu + MaxPooling + conv2 + FC)
  3. 模型融合: 多数表决融合
  4. 模型使用: 用训练好的模型测试未分类文本。
- **项目名称: 公司组织的用户性别预测比赛**  
**开发工具:** sklearn, Logistic Regression, Random Forest, DNN/MLP  
**项目描述:** 根据带标签的用户数据(机型, 地域, app 列表), 训练模型预测用户性别。  
**主要工作:** 基于用户数据训练用户性别预测模型, 具体工作内容如下:
  1. 读入数据, 数据清洗
  2. 特征选择: 基于 Random Forest 得到的特征重要度选择
  3. 模型选择:
    - baseline model: LR, improved model: MLP
  4. 模型融合: 基于 f-score 值作为权重对模型结果进行加权平均融合
  5. 模型使用: 用训练好的模型对未标记数据进行性别预测
  6. 模型评价: 准确率从 71%上升到约 79%