陈晓阳

自然语言处理 & 文本挖掘

联系方式

北京市海淀区
中关村南大街5号
北京理工大学
100081

+86 156 5299 5998

ling032x@gmail.com
http://ling0322.info
Twitter: @ling0322
Github: ling0322

语言

母语-中文
英语

编程语言

♥ Python
C++, Javascript, Go
Java, C#

教育情况

2012至今 **工学硕士** - 北京理工大学
专业: 计算机科学与技术, 研究方向: 自然语言处理
主要课程: 计算语言学、机器学习、数据挖掘、Web智能与社会计算.

2008 - 2012 **工学学士** - 苏州科技大学
专业: 计算机科学与技术
主要课程: 数据结构、操作系统、编译原理、网络原理、Java、C#

个人简介

主要研究方向是自然语言处理和文本挖掘, 除此之外兴趣爱好是学习各种编程语言Python3/Rust/Go/Scala. ♥Coding, 喜欢学习和研究新技术. Keep it simple, stupid 原则的坚定支持者.

实习经历

2014.5至今 **百度** 研发工程师
自然语言处理部门, 智能交互应用基础研究(HCI-BASE)小组, 目前参与的项目有:
- 爱奇艺视频搜索中唯一答案部分Query的省略
- 知识图谱中实体描述性句子的缩写

2012.3 **豌豆实验室(豌豆荚)** 前端工程师
前端项目组实习生, 公司内部协作工具的编写与维护.

项目经历

2012-2014 **MilkCat开源中文依存句法分析以及相关工具包** 开源项目 - goo.gl/uN3MXX
开源中文依存句法分析工具包, 采用Nivre08的arc-eager算法, 目前正在加入Zhang11的柱搜索部分, C++编写, 其他工具包括包括中文分词、词性标注、新词/短语发现等. 提供Java、Python、Go等多种语言的接口.

2013-2014 **WEBDICT中文词库计划** 开源项目 - webdict.info
使用机器学习以及众包方式构建一个无版权形式的中文词库, 可用于中文分词以及中文输入法, 目前已收集到22万细粒度中文词语.
- 网页爬虫(Twitter, 网易新闻, 腾讯新闻)
- 新词发现算法设计与实现(C++)

2013-2014 **新浪微博爬虫cococat** 实验室开源项目 - goo.gl/5gbiHj
支持微博用户信息抽取, 好友与关注着抽取, 用户时间线抽取以及搜索结果的抽取. Python编写, 我完成的部分:
- 模拟用户登录模块, 以及带Cookies的HTTP Request模块
- 用户微博以及评论抓取模块

2013 **Goldeye互联网数据挖掘软件** 实验室项目
网页新闻信息的数据挖掘平台(财经杂志已部署使用), 主要负责:
- 数据挖掘部分可视化展示(D3.js)
- 实体以及词语的关联分析以及趋势分析算法设计

2012–2013 **单词喵喵喵(某科学的背单词软件)** 开源项目 – goo.gl/KAfAVs
Android应用，用于记忆考研英语、CET、托福、雅思等常用单词。目前Google Play中1万–5万安装次数, 4.5分平均评分, 在中文区教育类应用中排名153。

2012–2013 **ICTCLAS安卓移植** 实验室项目
实验室与华为南京研究院合作项目, 分词软件ICTCLAS向安卓端的移植, 主要负责:

- 编译环境搭建以及参数的设置
- Java代码部分的编写

编程技能

编程语言
熟悉常用编程语言, 了解OOP、FP等常见编程范式

- ♥ Python, 熟悉Python以及会写Python C语言扩展
- 熟悉C++常用语法, 有良好的代码习惯, 对C++内部原理和优化技术略有了解
- 了解Java和C#, 有Android编程经验
- 其他熟悉或者尝试过的语言包括Javascript, Go, Scala, Ruby, Lua等.

Unix环境
熟悉Unix/Linux环境编程, 包括编辑器、编译器以及shell编程

- 熟悉常用的gcc/clang的参数以及gdb/lldb命令
- 习惯使用vim编辑器
- 了解shell编程以及常用工具sed、awk等的使用
- 有关于Makefile以及automake构建工程的知识

数据处理和数据库
了解关系型数据库相关知识, 有简单使用MongoDB以及Hadoop的项目经验

- 了解MapReduce基本原理, 写过简单的Hadoop Streaming脚本
- 熟悉MongoDB以及其shell的常用语法以及常用的Driver(C, Python, Java, Node.js)
- 了解常用的SQL查询语句, 在实际项目中使用过MySQL和SQLite数据库

专业特长

中文自然语言处理: 熟悉中文自然语言处理相关模型以及算法, 包括依存句法分析, 中文分词, 词性标注, 命名实体识别, 新词发现、关键词抽取、文本信息抽取以及. 了解相关领域内的论文, 开源工具以及库文件, 并且实现过相关论文中的算法和模型.

文本挖掘: 了解文本聚类、分类, 主题模型, 短文本处理, 情感计算等等相关的任务.

机器学习: 熟悉常用机器学习以及数据挖掘模型和算法(最大熵、逻辑回归、感知器、SVM、HMM等), 了解相关模型的数学理论, 包括概率论, 矩阵分析, 最优化方法等. 会使用相应的开源库解决实际任务.