

# 潘浩杰

+86 1781 6855 741 ◇ myscarletpan@gmail.com ◇ github ◇ blog

## 教育经历

### 工学学士 (计算机科学与技术)

2014 年 9 月 - 2018 年 6 月

浙江大学 (ZJU), 杭州, GPA: 3.94/4.0, 排名: 3/137

导师: 蔡登教授

浙大何志均荣誉班成员, 省级优秀毕业生。

### 研究性硕士 (Mphil, 两年制, 计算机科学与技术)

2018 年 8 月 - 现在

香港科学与技术大学 (HKUST), 香港

导师: 宋阳秋教授

获得为期两年的研究生全额奖学金 (postgraduate studentship award), 预计 2020 年毕业。

## 实习 & 经历

### 腾讯

2018 年 11 月 - 现在

算法实习生, 微信技术架构部数据中心

深圳, 中国

- 探索基于内容的微信公众号文章流推荐, 服务于微信应用“看一看”。
- 清理并构建微信公众号文章数据集, 包含超过**六千万**公众号文章和**四百万**用户。为了在单机环境中跑相应的实验, 我用了一系列数据分片和多核的思想来对文章进行关键词抽取, 生成用户的历史。并优化相应的数据结构, 权衡硬盘/内存与运行时间分别用数据库和内存字典存/读取数据。
- 在 30 天文章流点击预测中, 仅仅利用用户和朋友阅读文章文本特征达到 **65.17%** AUC (超过 baseline **8.4%**)。

### 阿里巴巴

2018 年 1 月 - 2018 年 6 月

算法实习生, 达摩院机器智能语言技术实验室

杭州, 中国

- 挖掘了超过 **10 亿**的天猫商品评论, 并且仅仅利用了成本较低的“Aspect”标注构建了一个弱监督的“评论集-摘要”的数据集, 为部门节省了很大一笔人工标注的支出。
- 构建了一个利用 Aspect 对齐的序列到序列模型, 大量实验后进行了人工评价, 评价结果我们的模型超出 baseline **7%**。这个评论摘要系统很有可能在淘宝平台进行部署。
- 参与了“司法大脑”项目, 通过构建不同的模型进行罪名、对应法条、刑期的预测, 并且在中国法研杯智能竞赛中取得了总分**第二名**的好成绩。

### 网易

2017 年 9 月 - 2018 年 1 月

算法实习生, 网易游戏伏羲人工智能实验室

杭州, 中国

- 构建了网易游戏第一个简历筛选系统, 并在每个招聘期间有稳定的 API 调用。
- 这个项目期间, 我参与了特征工程, 以及用 sklearn/xgboost 训练机器学习模型, 并通过决策树路径解释对简历每个方面的特征重要性进行评估并汇报。
- 与同事合作部署了相应的 API, 招聘期 HR 的反馈非常良好, 6,359 份简历中取得了 **85.7%** 的准确率。

### 浙江大学 CAD&CG 国家重点实验室

2016 年 10 月 - 2018 年 6 月

研究助理, 何晓飞 & 蔡登机器学习组

杭州, 中国

## 其他项目

---

### 低资源语言的情景结构抽取评测 (NIST 组织的 LoReHLT 评测)

2017 年 7 月 - 2017 年 9 月

- 我们团队 (UIUC-HKUST 团队) 协同构建了一个系统, 自动将一个在一些罕见语言 (如提格里尼亚语, 奥罗莫语, 维吾尔语等) 地区灾难新闻翻译成英文, 提取情景帧句子, 并训练分类器来判断是否这个新闻里有一些“issue”和“need”。
- 我的主要工作在于将网络中大量带 tag 的文章进行分类, 并将 tag 与目标“issue”和“need”建立映射, 在评测期间我用了一些启发式的关键词匹配使得我们在某一项评测中的结果得到大量提升。

## 技能 & 兴趣

---

编程语言	Python, C/C++, Java, Javascript, Matlab, SQL
机器学习	PyTorch, Sklearn, Pandas, Numpy, Xgboost, d3.js
兴趣	Text Mining, Natural Language Processing, Data Mining

## 专利 & 论文

---

**Haojie Pan** (First author), Manuscript, Long Paper, *WWW*, 2019 (submitted)

**Haojie Pan** (First author), Manuscript, Long Paper, *NAACL*, 2019 (submitted)

Stephen Mayhew, Chase Duncan, Mark Sammons, Mark Sammons, Chen-Tse Tsai, Xin Li, **Haojie Pan**, Sheng Zhou, Jennifer Zou, Yangqiu Song, Dan Roth. Illinois CCG LoReHLT 2017 Named Entity Recognition and Situation Frame Systems. *LoReHLT*. 2017 [pdf]

潘浩杰, 刘洋, 周君沛, 陆家林, 面向视觉障碍人士的深度视觉问答系统的构建方法, 专利号 201710128118.6 [link]

## 比赛

---

### 释义识别 (Paraphrase Identification)

2018 年

- AETC 蚂蚁开发者大赛/金融大脑-金融智能 NLP 服务[link] 团队, Top 5%, 21/449
- 拍拍贷 AI/第三届魔镜杯大赛 [link] 团队, Top 3%, 16/576

### 文本分类

2018 年

- 中国“法研杯”法律智能挑战赛 (CAIL) [link] 团队, Top 1%, 2/202

### 数据挖掘

2017 年

- Kaggle: Two Sigma Connect: Rental Listing Inquiries [link] [code][post(收藏量 >2k)] 单人, Top 5%, 105/2488