

黄东川

tel: 15580901455

国防科大·海量数据处理实验室

email: huangdongchuan@outlook.com



基本信息

生日: 1992.09.25	籍贯: 四川达州	政治面貌: 中共党员
研究领域: 大数据处理、知识图谱、自然语言处理、网络爬虫		
学校及专业: 计算机科学与技术 (硕士, 国防科大); 计算机科学与技术 (学士, 电子科大)		
实验室: 国防科大计算机学院681实验室		导师: 贾焰

Github: <https://github.com/hplover>

个人博客: www.hplover.xyz

专业技能

1. 英语水平: 国家六级 (优秀)
2. 精通 C / Python / Java / Maven, 熟悉 Linux / MySQL / MongoDB
3. 精通 Python / BeautifulSoup / Selenium, Java / JSoup 爬虫框架
4. 熟悉 大数据处理、机器学习、自然语言处理 等相关算法和技术
5. 熟悉 HTML / CSS / Bootstrap、Django 等网站开发语言及框架
6. 熟悉 PhotoShop、Word / PowerPoint / Excel 等图形处理工具及办公软件
7. 精通 Eclipse / PyCharm / Visual Studio / IntelliJ IDEA 等IDE开发工具

项目经历

1. 2015.06 – 2015.08: NIST TREC 2015 Microblog task

成绩: RANK 1st.

介绍: 该比赛由 NIST (美国国家标准与技术研究院) 主办, 已持续近 20 年, 每年都吸引着来自 CMU、牛津大学等全球优秀的参赛者。该比赛由许多小的任务组成, 我们参与的是微博任务, 根据用户的兴趣实时推送相关的Tweeter推文。

担任角色: 共4人参赛。我负责文本扩展、特征抽取以及生成候选项模块的代码编写。

相关技术: 文本预处理、网络爬虫、TFIDF特征抽取、词向量技术、文本相似度计算。

文本预处理: 利用Stanford的NLP工具处理描述用户兴趣的文本, 包括分词及去停用词

网络爬虫: 利用Python+Urllib2+BeautifulSoup编写的网络爬虫爬取关键词的google搜索结果

TFIDF特征词抽取: 获取google搜索的前20个页面, 利用TFIDF计算与该关键词相关的词汇作为该关键词的特征词

词向量技术: 利用Python+Gensim的方式处理英文版的Wikipedia, 获得每个词汇的向量表示

文本相似度计算: 推文经过文本预处理操作后与用户的兴趣文本做Word2vec相似度计算, 获得推文与每个用户兴趣的相关度

2. 2016.03 – 2016.06: 681实验室“大搜索与情报搜索”项目

介绍: 该项目为国家973项目, 实现一个针对情报的搜索引擎。该项目由三个部分组成, 用户意图理解模块、知识图谱模块以及答案质量评估模块。

担任角色：用户意图理解模块中命名实体链接任务的完成者，知识图谱模块参与者。

任务介绍：一个名称很可能对应多个概念，一个概念也有可能拥有多个名称；比如：国防科大的学生允许使用苹果产品吗？中的“国防科大”是“国防科学技术大学”的同义词，“苹果”可以指“苹果公司”、“苹果水果”、“苹果电影”等概念。实体链接的目的就是明确文本中的名称到底指向哪个概念。知识谱图是一种管理知识的数据库。

相关技术：网络爬虫、实体字典构建、文本预处理、词向量技术、文本相似度计算。

网络爬虫：使用Java+Jsoup的方法编写网络爬虫多线程地爬取约两百万的百度百科页面

实体字典构建：利用爬取得数据构建包含同义词、多义词关系的实体字典，同时写入知识图谱

文本预处理：利用中文自然语言处理工具HanLP对文本进行分词和词性标注获得名称的上下文

词向量技术：利用Python+Gensim工具构建基于搜狐新闻语料库的词向量

文本相似度计算：利用Jaccard以及Word2vec计算名称与概念的相似度

3. 2016.05 – 2016.07: Baidu Cup 2016 NLP 挑战赛

成绩：RANK 3rd.

介绍：该比赛由CCF和百度公司主办。比赛任务是实现一个推荐系统，用户输入要求，计算机自动推荐满足要求的实体（电影/电视剧/菜馆/名人）。

担任角色：共2人参赛。我负责特征抽取、词汇扩展部分。

相关技术：网络爬虫、文本特征工程、词向量。

社会活动及个人经历

2015.10 – 2015.11: 国防科大首届乒乓球比赛季军

2015.03 – 2015.07: 计算机学院五队14级10班班长

2013.11 – 2014.01: 获成都市高校棋牌联赛五子棋个人及团队冠军

2012.09 – 2013.09: 获电子科技大学“国家励志奖学金”

2012.09 – 2013.03: 任电子科技大学三星俱乐部主席

2012.03 – 2012.05: 获三星手机游戏“优秀开发者”称号

2011.09 – 2012.09: 获电子科技大学“人民二等奖学金”

2010.09 – 2011.09: 获电子科技大学“人民三等奖学金”

2010.11 – 2011.06: 任电子科技大学计算机学院新闻中心宣传组长

2010.09 – 2010.10: 获电子科技大学计算机学院新生乒乓球比赛亚军

教育经历

2014.09 – 至今: 国防科学技术大学 | 计算机科学与技术 | 硕士

主修课程：数据库管理系统、高级计算机体系结构、计算复杂性、数理逻辑、UNIX 高级编程、人工智能、高级人工智能技术等

2010.09 – 2014.06: 电子科技大学 | 计算机科学与技术 | 学士

综合成绩：84.7/100

排名：23/250

主修课程：电路原理、模拟电子技术、数字逻辑、计算机组成原理、微型计算机技术、计算机系统结构、计算机网络、高级语言、汇编语言、数据结构、操作系统、数据库原理及应用、编译原理、图形学、离散数学、概率统计、线性代数、C 语言程序设计、C++程序设计、Java 程序设计、数据结构及算法设计等