

# 2015 春季学期总结

韩喆

iampkuhz@gmail.com

PIE 组

2015 年 12 月 10 日

- 工作

- 研究生课程
- 实验室工作
  - 新华社新闻类别标注
  - 中文维基百科谓词归一化
  - 杂项

- 总结、展望

## 研究生课程

- 面向对象分析与设计
- 高等计算机体系结构
- 中国特色社会主义理论与实践研究
- 自然语言处理高级专题
- 高体不挂的话，就完成授课类课程要求的学分了
- 认真程度不及上学期
- OO 课件比较漂亮，和冯老师的课件有的拼
- 高体比较像考历史（必考 + 生僻知识点）
- 政治课都差不多，按时交作业就好
- NLP 讲的比较难，但作业比较简单

# 实验室工作

— 新华社新闻类别标注

## 新华社新闻类别标注

- 和曾颖 (网站搭建) 负责
- 上学期完成样例新闻标注
- 迭代增加功能

- ① 完成一级类别、二级类别标注任务
- ② 增加了多种新闻来源 新浪、中国新闻网、人民网、新华网、网易新闻、腾讯
- ③ 自动生成标注任务评价
  - 准确率: 标注语料随机混合部分已标注语料
  - 一致性: 不同标注者含有部分相同新闻, 与其他人比较
- ④ 二级类别根据个人意愿半自动生成
  - 手工记录标注者擅长类别优先级, 自动从已标注一级类别的新闻中选择给二级类别给标注者, 控制不同人的重合新闻数量
- ⑤ (?) 总体觉得比新华社的样例数据更好 (一级类别新闻分布类似长尾)
  - 标注了 5.7w 个一级类别, 1.9w 个 2 级类别

## 新华社新闻类别标注

### ICST-新闻聚合网站-供标注

● 登陆



登录

用户名

qingchunligood@sina.com

密码

\*\*\*\*\*

提交

Copyright: ICST @ PKU

Chrome 10 / Safari 5 / Opera 11 or higher version, with 1024x768 or higher resolution for best views.

## 新华社新闻类别标注

ICST-新闻聚合网站

Hi, 青春利! 注销

### ICST-新闻聚合网站-供标注

● 登陆

● 主页

●

标注任务 反馈 公告

任务名称	完成进度	标注入口
edit20141225	1000 / 1000	已下线
edit20150107	1000 / 1000	已下线
edit20150121	1000 / 1000	已下线
edit20150126	900 / 900	已下线
edit20150131	900 / 900	已下线
edit20150205	1350 / 1350	已下线
edit20150213	4500 / 4500	已下线
edit20150220	453 / 10000	已下线
editSec20150329	1470 / 1500	Go>
editSec20150413	1569 / 1600	Go>

## 新华社新闻类别标注

ICST-新闻聚合网站

Hi, 青春利! 注销

### ICST-新闻聚合网站-供标注

- 登陆
- 主页
- 公告

标注任务 反馈 公告

任务名称	完成进度	标注入口
edit20141225	1000 / 1000	已下线
edit20150107	1000 / 1000	已下线
edit20150121	1000 / 1000	已下线
edit20150126	900 / 900	已下线
edit20150131	900 / 900	已下线

标注任务 反馈 公告

- 20150427 请大家按任务顺序进行标注。过两天可能会暂停标注一段时间（持续几天），之后会安排新的数据，请大家注意安排任务
- 从第四次标注开始，每位同学的标注的新闻数量不一定为1000篇，敬请留意
- 第五次标注任务（edit20150131）已经生成，第四次标注结束的同学可以开始第五次标注了；
- 第四次标注任务（edit20150126）已经生成，第三次标注结束的同学可以开始第四次标注了；
- 请大家登录时留意用户名是否为自己的邮箱或姓名，以免错标成别人的任务；



## 新华社新闻类别标注

ICST-新闻聚合网站

Hi, 青春! 注销

### ICST-新闻聚合网站-供标注

- 登陆
- 主页
- 公告
- 反馈

标注任务 反馈 公告

任务名称

edit20141225

edit20150107

edit20150121

edit20150126

edit20150131

edit20150205

edit20150210

edit20150215

edit20150220

edit20150225

edit20150228

标注任务 反馈 公告

任务

edit20150126

edit20150121

edit20150107

内容

评定：最优注：评定等级由计算机自动完成。抽样后按照和真实新闻类别的重合度打分。等级候选：达标类（最优、优、良、差）、不达标

评定：优注：评定等级由计算机自动完成。抽样后按照和真实新闻类别的重合度打分。等级候选：达标类（最优、优、良、差）、不达标

评定：优注：评定等级由计算机自动完成。抽样后按照和真实新闻类别的重合度打分。等级候选：最优、优、良、差

时间

2015-03-09 22:05:35

2015-03-09 17:25:38

2015-03-09 17:22:29

标注任务 反馈 公告

- 20150427 请大家按任务顺序进行标注。过两天可能会暂停标注一段时间（持续几天），之后会安排新的数据，请大家注意安排任务
- 从第四次标注开始，每位同学的标注的新闻数量不一定为1000篇，敬请留意
- 第五次标注任务（edit20150131）已经生成，第四次标注结束的同学可以开始第五次标注了；
- 第四次标注任务（edit20150126）已经生成，第三次标注结束的同学可以开始第四次标注了；
- 请大家登录时留意用户名是否为自己的邮箱或姓名，以免错标成别人的任务；

## 新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务

ICST-新闻聚合网站

个人主页

当前标注任务: editSec20150329

更换任务

Hi, 青春利!

公告

注销

新闻列表

◀

●

●

●

●

●

●

●

●

●

▶

输入页码

GO

请输入要查找的文章编号

查询

标注进度: 已完成0.0%

编号	题目	一级类别	修改的一级类别	二级类别	内容
1401	大涨后市场步入喘息期 私募调仓剑指白马股	财政、金融	未修改	股票市场	<a href="#">查看</a>
1402	福布斯:小米正考虑以超400亿美元估值进行融资	财政、金融	未修改	投资、融资	<a href="#">查看</a>
1403	程赤兵:老北京小吃比露天烧烤更环保	环境、气象	未修改	环境保护	<a href="#">查看</a>
1404	(电视通稿 国内) 贵州:“十一”期间将动态发布交通气象等信息	环境、气象	未修改	气象	<a href="#">查看</a>
1405	澳大利亚:确认疑似失联客机残骸还需两三天	灾难、事故	未修改	交通事故	<a href="#">查看</a>
1406	东莞一家餐厅发生严重爆炸 24人受伤(图)	灾难、事故	未修改	公共安全事故	<a href="#">查看</a>
1407	农妇因家庭矛盾菜坛内放农药毒死家中3名幼儿	法律、司法	未修改	刑事犯罪与案件	<a href="#">查看</a>
1408	内蒙古书记讲段子答记者:雾霾因内蒙古树多挡风	环境、气象	未修改	环境	<a href="#">查看</a>
1409	(服务·房产快报) 武汉房地产二级市场出现抛盘现象	财政、金融	基本建设、建筑业、房地产		<a href="#">标注</a>
1410	美国连遭多场龙卷风袭击已致29人死亡	灾难、事故	未修改	气象灾害	<a href="#">查看</a>

## 新华社新闻类别标注

ICST-新闻聚合网站

个人主页

当前标注任务: editSec20150329

更换任务

Hi, 周春利!

公告

注销

### ICST-新闻聚合网站-供标注

#### (服务·房产快报) 武汉房地产二级市场出现抛售现象

(服务·房产快报) 武汉房地产二级市场出现抛售现象

新华社武汉5月17日专电 据武汉晚报报道:6月1日,新的营业税政策就要实施。16日,来自武汉各房屋中介的消息,一些刚刚办证的房子想避开营业税新政策,已开始抛售了。

记者从顺驰、华明达、百屋易等中介获悉,现在武汉市场上抛售的“次新房”的量很大,这些房子都要交营业税的。华明达总经理叶正明说,这些人想抢在6月1日前进行交易,但现在一下子很难找到买主;无法抛出。据悉,这部分人以有多处房产的投资者和投机者为主。

顺驰中介的张俊峰经理说,很多房子委托给中介卖,周期是一个月,交易成本一增加,很多卖主有可能会跳单,最麻烦的是由于跳屋交易有一个周期,即使现在找到买主,也不一定能在6月1日前实现交易,跳单就麻烦。据几家中介介绍,这部分人以那些不急着想钱用,可卖可不卖的房子为主,跳单后就转为出租。

据了解,现在的二手房交易中,费用一般都是买房人承担。百居易总经理谢娟说,6月1日后,买房成本将增加。现在卖房者在中介登记交易时,多有“所有费用不管”这样的一条,营业税增加后,买房者要承担更多的费用。(完)

标注任务

editSec20150329

文章id

1499

一级类别

基本建设、建筑业、房地产

二级类别

提交

修改一级类别

返回上一篇文章

查看下一篇文章

返回新闻列表

- 登陆
- 主页
- 公告
- 反馈
- 标注任务
- 标注新闻

## 新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务
- 标注新闻
- 管理员

ICST-新闻聚合网站

Hi, 管理员! 注销

### ICST-新闻聚合网站-供标注

用户列表

用户名	真实姓名	密码
summeryumin@pku.edu.cn	于敏	45202263
sjqs.s@163.com	苏嘉琦	48337590
zhangguodong.09@163.com	张国栋	97335022
1144651728@qq.com	程磊	70510211
1200013976@pku.edu.cn	关凌宇	04691827
huruiying_pku@163.com	胡瑞英	52858249
qingchunligood@sina.com	青春利	93083011
wshqin@gmail.com	祁鑫	43066428
zengying		123456
admin		Icst2015

edit20141225(已下线)

edit20150107(已下线)

## 新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务
- 标注新闻
- 管理员

ICST-新闻聚合网站

Hi, 管理员! 注销

### ICST-新闻聚合网站-供标注

用户列表

用户名
summer
sjqs.s@163.com
zhangguodong.09@163.com
1144651728@qq.com
1200013976@pku.edu.cn
huruiying_pku@163.com
qingchunligood@sina.com
wshqin@gmail.com
zengying
admin

edit20141225

edit20150107

edit20150107

ICST-新闻聚合网站

Hi, 管理员! 注销

### 我的反馈

用户名	真实姓名	标注任务	反馈
sjqs.s@163.com	苏嘉琦	edit20141225	<a href="#">查看</a>
zhangguodong.09@163.com	张国栋	edit20141225	<a href="#">查看</a>
1144651728@qq.com	程嘉	edit20141225	<a href="#">查看</a>
1200013976@pku.edu.cn	关凌宇	edit20141225	<a href="#">查看</a>
huruiying_pku@163.com	胡瑞英	edit20141225	<a href="#">查看</a>
qingchunligood@sina.com	青春利	edit20141225	<a href="#">查看</a>
wshqin@gmail.com	祁鑫	edit20150107	<a href="#">查看</a>
zhangguodong.09@163.com	张国栋	edit20150107	<a href="#">查看</a>

# 实验室工作

— 中文维基百科谓词归一化

## 中文维基百科谓词归一化



Subject	Predicate	Object
張家輝	罗马拼音	Cheung Ka Fai
張家輝	英文名	Nick Cheung
張家輝	国籍	中国(香港)
張家輝	籍貫	广东番禺
張家輝	出生	1967年12月2日 (46岁)
張家輝	出生	英属香港
張家輝	语言	粤语
張家輝	语言	英语
張家輝	语言	普通话
張家輝	配偶	关咏荷(2003年至今)
張家輝	儿女	张童(Brittany Cheung)
張家輝	活跃年代	1989年至今
張家輝	经纪公司	钟珍

- **背景/回顾**: 基于中文维基百科的知识库: 330w 三元组, 1.6w 个谓词

# 中文维基百科谓词归一化

### “ 邮政 ”

INSEE/邮政编码、ISO 3166-2  
邮政简写、美国邮政编号、美国  
邮政编码、邮政、邮政代  
码、邮政信箱、邮政分区、邮  
政区号

- Motivation: 消除语义重复的谓词
- 问题转化: 任给两个谓词, 判断是否为相同语义
- 实验效果: 在正负样本 1:1 的数据集上有监督训练, 将近 70% 的准确率



## 中文维基百科谓词归一化

### “ 邮政 ”

INSEE/邮政编码、ISO 3166-2  
邮政简写、美国邮政编号、美国  
邮政编码、邮政、邮政代  
码、邮政信箱、邮政分区、邮  
政区号

- Motivation: 消除语义重复的谓词
- 问题转化: 任给两个谓词, 判断是否为相同语义
- 实验效果: 在正负样本 1:1 的数据集上有监督训练, 将近 70% 的准确率
- 进展: 投了 NLPCC2015, 被拒。。

## 杂项

- ① 提供了 NLPCC2015 测评的知识库
  - NLPCC2015 task: Entity Recognition and Linking in Chinese Search Queries
- ② 大组 + 小组讨论班 (报告 4 次)

## 总结展望

# Summary

## 上学期计划

- 每日备忘录
- 每周两篇论文
- 考托福

## 本学期完成

- 每日备忘录
- 维基百科谓词的归一化

## 暑期计划

- 完善中文维基百科知识库的体系
  - 借鉴 dbpedia, 整合中英文维基百科、freebase
- 看一下有关的书: shell, Java
- 修改 NLPCC 的论文, 继续投稿

## 下学期计划

- 尽量在开学时完成知识库体系构建, 整理出一篇期刊
- 完成助教学分
- 百度百科的知识库提取和整合
- 动手做点小东西 (展示知识库)

谢谢大家！