《综合设计》大作业展示

2019201409 于倬浩

KoalaGo Search About 赵鑫 Go

Results for 赵鑫

赵鑫 - 师资科研

Time not available.

赵鑫 赵鑫，信息学院副教授（HTTPS://WEIBO.COM/BATMANFLY)、博士生导师。博士师从北京大学李晓明教授，专注于研究面向文本内容的社交用户话题兴趣建模，曾获得2011年谷歌博士奖研金（该年大陆高校共4人获奖）和2012年微软亚洲博士奖研金（该年亚太高校共10人获奖）。近五年内在.....

我院赵鑫老师在社交媒体数据挖掘方面取得系列成果 - 学院新闻

2016-04-19 09:57

我院 赵鑫 老师在社交媒体数据挖掘方面取得系列成果来源：大数据分析实验室我院 赵鑫 老师最近在社交媒体数据挖掘方面取得一系列成果，自2015年底至今，共发表3篇CCF A类论文，其中两篇期刊论文分别被数据知识工程领域核心期刊IEEE TRANSACTIONS ON KNOWLEDGE AND DATA EN.....

我院赵鑫老师在信息检索领域和数据挖掘方面取得系列成果 - 学院新闻

2015-04-24 15:27

我院 赵鑫 老师在信息检索领域和数据挖掘方面取得系列成果来源：大数据实验室我院大数据分析与智能实验室 赵鑫 老师最近在信息检索领域和数据挖掘方面取得一系列成果，作为第一作者共发表3篇CCF A类论文、1篇CCF B类论文，其中两篇期刊论文分别被信息系统领域顶级期刊 ACM TRANSACTIONS ON I.....

我院2019年班级建设工作成绩优异 - 学院新闻

简洁的搜索结果界面

简单介绍

- 实现了多线程Python爬虫，实测测试最快**62s**爬完整个网站(16线程)。
- 后端采用Python实现TF-IDF算法，实测API查询100次 ~0.5s
- 切词器采用百度LAC，并加载一个由HanLP, pkuseg和jieba共同生成的切词词典，提高准确率，稍后详细介绍。
- 前端采用Flask和BootStrap4框架实现了简洁的响应式用户界面，移动设备友好。

思路&技巧

- 后端系统大致分为三个部分，网页爬虫, 页面切词和规范化，索引&搜索。
- Crawler: 爬取网页，瓶颈主要在网络响应延迟。为了保证爬取数据的效率，采用Python多线程实现并行爬取。技术上遇到的问题主要是Python没有原生支持的atomic ++操作符，如果使用连续整数对爬取文档进行标号，需要在生成编号时，手动加进程锁，防止多个文档共用一个编号。爬取网页时通过简单的字典去除非HTML格式的文件(例如pdf)，提高效率。

思路&技巧

- Parser: 将HTML转化为纯文本的主要挑战是提取正文部分。直接查找正文不太容易。为了保证做法的通用性，采用Python的BeautifulSoup库，识别并删除HTML body中的所有链接标签(“a”)、脚本标签(“script”)以及HTML注释。如果网页遵守HTML设计规范，这种方法的效果非常好，经测试在我校多个学院的官网均可正常运行。在实现中我仍添加了一个小型词典去除无关词语(如“版权所有”)，进一步优化提取结果。

思路&技巧

- 切词器：在选择不同切词器时，我遇到了不少困难。
- 首先，实测pkuseg和百度LAC表现最好，能在保证效率的情况下兼顾准确度。HanLP可以做到更高的精确度，以及基于BERT模型的实体识别，但是缺点是在本机运行效率过低，不可能对全文进行识别，且只支持短句。而Jieba支持“搜索引擎模式”，可以返回更多可能的结果，缺点是精确性不够高。几种切词器在不同数据下各有优缺点。
- 综合上述，使用pkuseg和jieba的搜索引擎模式，分别对语料进行切词，并对词库取并集。下一步使用HanLP的实体识别，判断词库中每个词是否为实体，并将实体输出到词典中。使用百度LAC加载词典后，再用于parser和搜索引擎的切词，提高切词精确度，

构建索引的问题

- 对于索引部分，我们可以发现其实朴素TF-IDF算法在实际情况下表现并不优秀，精确度不高。例如老师的主页，老师的名字(中文)出现频次并不如其他介绍老师成果的文章高，所以排名靠后。
- 而且在文章外链结构相似(采用同一套网页模板)的情况下，即使使用PageRank算法可能效果依然不佳。
- 不易判断用户输入关键词是否在搜索结果中连续出现(码量较大)。
- 某些关键词更为重要，用户更加关注(**2019年**优秀大学生夏令营获奖名单)

简单解决

- TF-IDF算法，实际上是基于词频的统计分析，来实现对文档的评价和排序。
- 那么比较简单的优化是，手动增加同时出现在文章和标题中的词语，在计算TF值时的权重。简单理解为标题中的词“一个顶俩”(重要性更大)。
- 观察网页结构，给URL中包含”academic”的页面更大权重(体现在词频计算上)。
- 因不支持模糊匹配，将所有的英文字母转为大写可以简单的优化。
- 给包含年份或者量词的词语更大权重，提高精确度(正则表达式匹配)，例如”xxx年优秀夏令营”，匹配到用户更关注的结果(精确年份)。
- 以上几个优化均在构建索引过程中实现，不涉及写规则特判query。

查询范例

- 陈红 杜小勇(一类页面权重较大)
- 2013sigmod(数字、年份权重较大)
- 数据工程与知识工程教育部重点实验室学术委员会(切词器字典)
- 第五届百度(标题词权重大)
- 2019年优秀大学生夏令营获奖名单(数字、标题权重都大)

使用方法

- $ git pull https://github.com/zhuohaoyu/KoalaGo.git
- $ pip install –r requirements.txt
- $ cd crawler
- $ python crawler.py
- $ python parser.py
- $ cd ../frontend
- $ python app.py

- 或者直接访问https://bj.zhuohao.me:2333/
- ~~2333~~