

云数据管理II 《检索结果分类/聚类》

2015080121 软工52 李在弦

2015013161 软工51 魏佳辰

实验环境

操作系统: Windows

IDE: PyCharm

系统结构: b/s 结构

Web框架: Django 2.1.2

后端编程语言: Python 3.6

数据库: MongoDB

数据库

• MongoDB 介绍

MongoDB 是一个基于分布式文件存储的数据库，它是一个面向文档存储的数据库，操作起来比较简单和容易。如果负载的增加（需要更多的存储空间和更强的处理能力），它可以分布在计算机网络中的其他节点上这就是所谓的分片。

• 分布式存储

当 MongoDB 存储海量的数据时，一台机器可能不足以存储数据，也可能不足以提供可接受的读写吞吐量。这时，我们就可以通过多台机器上分割数据，使得数据库系统能存储和处理更多的数据。分片结构端口分布如下：

```
shards:
  { "_id" : "shard0000", "host" : "localhost:27020" }
  { "_id" : "shard0001", "host" : "localhost:27021" }
  { "_id" : "shard0002", "host" : "localhost:27022" }
  { "_id" : "shard0003", "host" : "localhost:27023" }
```

Config server: 27100

Route Process: 40000

启动shard server, config server和route process之后，我们对分片进行配置

首先通过db.runCommand({addShard:" localhost:27020" })等命令添加分片

然后通过db.runCommand({enablesharding:" docdb" })来设置分片存储的数据库

再通过db.runCommand({shardcollection:" docdb.info" ,key:{time:1}})来分片info这集合

然后我们通过mongoimport导入处理好的json文件来插入数据，最后数据库自动分片结果如下

```

mongos> db.info.count()
36231
mongos> db.info.getShardDistribution()

Shard shard0000 at localhost:27020
data : 128.93MiB docs : 12835 chunks : 1
estimated data per chunk : 128.93MiB
estimated docs per chunk : 12835

Shard shard0001 at localhost:27021
data : 0B docs : 0 chunks : 1
estimated data per chunk : 0B
estimated docs per chunk : 0

Shard shard0002 at localhost:27022
data : 135.51MiB docs : 10120 chunks : 1
estimated data per chunk : 135.51MiB
estimated docs per chunk : 10120

Shard shard0003 at localhost:27023
data : 51.01MiB docs : 3155 chunks : 2
estimated data per chunk : 25.5MiB
estimated docs per chunk : 1577

Totals
data : 315.46MiB docs : 26110 chunks : 5
Shard shard0000 contains 40.87% data, 49.15% docs in cluster, avg obj size on shard : 10KiB
Shard shard0001 contains 0% data, 0% docs in cluster, avg obj size on shard : NaNKiB
Shard shard0002 contains 42.95% data, 38.75% docs in cluster, avg obj size on shard : 13KiB
Shard shard0003 contains 16.17% data, 12.08% docs in cluster, avg obj size on shard : 16KiB

```

算法说明

- **网络爬虫**

从 10 万多个单词词库中选取 1700 多个单词，通过 request 库访问百度首页，输入单词，爬取前三个搜索结果的 URL，标题等信息。

访问每个 URL，通过 BeautifulSoup 库分析 HTML 信息，排除不需要的标签，获取文本信息。

最后把 URL，标题，文本内容等信息保存到 txt 文件。

最终爬取到 5222 个文档信息。

- **分词，提取关键词，获取词干，建立“单词-文档权值矩阵”和倒排索引**

读取刚刚保存的所有文档的文本信息，提取关键词，进行分词（只对关键词进行分词）。

进行分词的时候，为了更好地对英文单词进行分词，用 nltk 开源库获取每个关键词的词干，然后用这些词干建立“单词-文档权值矩阵”并保存关键词在文档中出现的次数。

同时，为了以后实现命中词突出显示的功能，把每个关键词在文档中出现的位置存储到倒排索引。

- **筛选关键词**

删除关键词中的不需要的字符

```
del_keyword = '!\"#$%^&*()_+=~`|[]{}~:;/?>,<, . , ' " ' : , | } [ { - ) ( ~ ~ ~ ! ~ ~ ~ ~ ~ ? ~ ~ ~ ~ ~
```

- **停用词处理**

设置几十个常见的停用词，然后修改“单词-文档权值矩阵”中每个停用词出现次数。

```
stop_words.txt
1 the 100
2 of 100
3 is 100
4 and 100
5 to 100
6 in 100
7 that 100
8 we 100
9 for 100
10 an 100
11 are 100
12 by 100
13 be 100
14 as 100
15 on 100
16 with 100
17 can 100
18 if 100
19 from 100
20 which 100
21 you 100
22 it 100
23 this 100
24 then 100
25 at 100
26 have 100
27 all 100
28 not 100
29 one 100
30 has 100
31 or 100
32 that 100
33 的 1000
34 了 100
35 是 100
36 在 100
37 和 100
38 有 100
39 他 100
40 不 100
41 我 100
42 人 100
43 也 100
44 为 100
```

- **使用 Tf.idf 扩展布尔模型，计算 m_j , tf , d_j , n_k , idf , W_{kj}**

- **Term frequency (tf):**

$$tf_{ij} = f_{ij} / m_j \quad \text{when } f_{ij} > 0 \quad m_j = \max (f_{ij})$$

Inverse document frequency (idf):

$$idf_i = \log \left(\frac{n}{n_i} \right) + 1 \quad n_i > 0$$

- Suppose there are n documents and that the number of documents in which term k_i occurs is n_i .

$$w_{kj} = (f_{kj} / m_j) * (\log (n/n_k) + 1) \quad \text{when } f_{kj} > 0$$

- 检索, 相关性计算

收到查询语句之后对它进行分词, 记录每个单词出现次数, 计算 m_q , w_{kq} , d_q 。

$$w_{kq} = (f_{kq} / m_q) * (\log(n/n_k) + 1) \quad \text{when } f_{kq} > 0$$

$$w_{kq} = (0.5 + 0.5 * f_{kq} / m_q) * (\log(n/n_k) + 1) \quad \text{when } f_{kq} = 0$$

然后计算查询语句和每个文档之间的相关度,

$$\text{sim}(\mathbf{d}_q, \mathbf{d}_j) = \frac{\sum_{k=1}^n w_{kq} w_{kj}}{|\mathbf{d}_q| |\mathbf{d}_j|}$$

最后按相关度排序结果。

- 聚类

获取所有搜索结果文档的关键词, 从“Tf.idf 扩展布尔模型矩阵”获取各个文档和关键词相应的权值。用这些文档编号列表, 关键词列表和权值来建立新的“Tf.idf 扩展布尔模型矩阵”。

Ex) 假设下边图是原来的“Tf.idf 扩展布尔模型矩阵”

	ant	bee	cat	dog	eel	fox	gnu	hog
q	1.18	0.59	0.74	1.18	0.74	0.59	0.74	0.74
d_1	1.18	0.59	0	0	0	0	0	0
d_2	0.29	0.29	0	1.18	0	0.29	0	0.37
d_3	0	0	1.48	1.18	1.48	1.18	1.48	0

其中, 如果跟搜索语句有相关的文档是 d_2 和 d_3 , 它们两个的关键词并集是“ant, bee, cat”, 那么新的“Tf.idf 扩展布尔模型矩阵”是

	ant	bee	cat
d_2	0.29	0.29	0
d_3	0	0	1.48

用这些权值来进行动态聚类。

聚类部分用了动态聚类方法中的重心法。首先计算全部聚类文档权值的重心, 建立新凝聚点的最小临界距离 d , 依次逐个输入全部文档权值, 如果输入文档与已有凝聚点的距离大于 d , 则将它作为新的凝聚点; 否则不作为凝聚点;

最后，把每个类的关键词权值分别加，然后用权值最高的关键词来代表每个类

Ex) ant = 0.29 + 0 = 0.29

bee = 0.29 + 0 = 0.29

cat = 0 + 1.48 = 1.48

所以权值最高的关键词是“cat”，所以用它代表类名

可视化

IR

亚运会

Submit

2018年雅加达亚运会_新浪体育_新浪网

2018年11月14日·[亚运会](#) 飞艇印尼 与韩国组建3个项目队伍 赛艇中心焦点机师马云：全力配合女足发展 下届冠军咱们的 周继红：10金6银成双满意 为东京周期热身 奖牌榜优势无可撼动 五连胜点阵中国代表 五连胜继续雅加达 ...
<http://sports.sina.com.cn/yayun2018/>

LOL 亚运会表演赛详细赛程公布 中、韩将上演宿命对决_游迅网

2018年8月25日·[亚运会](#) 表演赛小组赛赛程：小组赛将进行六轮，为BO1。 小组赛结束后，每个小组的前两名出线，进行半决赛，半决赛为BO3。 半决赛败者只能进行季军战，争夺本次比赛的季军，胜者进行冠军赛，冠军赛为B ...
<http://www.yxdown.com/news/201808/415586.html>

2018年雅加达亚运会_百度百科

2018年10月24日·[亚运会](#) 国际V百科驻期员胡希统计浏览次数：本编辑次数：142次历史版本最近更新：最新计划图G（2018-10-24）突出贡献者zhongguo5526周的天说会下个转角来意510雪地红雪1231重次 ...
<https://baike.baidu.com/item/2018%E5%B9%B4%E9%98%B...>

2022年杭州亚运会_百度百科

2018年10月16日·[亚运会](#) 第7影就意义8社会评论赞你关注进入词典清除历史记录关闭编辑收藏赞数分享 ...
<https://baike.baidu.com/item/2022%E5%B9%B4%E6%90%A...>

ALL

1

亚运会

4

1

《输入“亚运会”的搜索结果》

IR

软件学院

Submit

南京点明 软件 科技有限公司_点明论坛

软件 论坛 点明2018云记陈点明通晓联系购买点明市场帮助维修机原查看更多荣耀9青春版 4+64 全网通尊享版 更新时间: 2018年03月26日 14:37 安卓版系统8.0CPU核心:八核运 ...
<http://www.dmjky.com/>

董兆(生命科学 学院)老师 - 西北大学 - 教师点评网

2014年10月8日·[学院](#) Processed in 0.092269 second(s). Total 4, Slave 4 queries, Gzip On, MemCache On. 男 女 ...
<https://teacher.zjut.cc/782163/>

合众国诉微软案_百度百科

2018年7月1日·[软件](#) 通过购买的、与之相关的一些而概括包括：微软是否通过改变或操纵应用程序接口（API）来拉拢使用Internet Explorer更具优势的目的；微软与原始设备制造商（OEM）之间是否达成过相关的限制性 ...
<https://baike.baidu.com/item/%E5%90%88%E4%B8%B7%E5...>

一秒钟变萌电脑版下载|一秒钟变萌app安卓版 v3.6.1_手机天堂

软件 人气下载1斗鱼直播社交下载2奇热小说阅读下载3小皮助手工具下载4聊客社交下载5免费小说阅读下载6微信多开助手系统下载7乐乐抢红包社交下载8360浏览器2017工具下载9微信2018社交下载10 ...
<https://www.xpgod.com/shougousoft/11656.html>

画图精彩实例官方版-独木成林

2018年9月14日·[软件](#) V1.32绿色中文版网页E图语音电话管理软件及驱动V139官方版普通网络计划图学习版官方下载V1.0.0官方版曙光文字游戏制作工具官方版通国总部 ...
<https://www.3kxiazai.com/soft/tool/6681.html>

行政管理专业_百度百科

2018年7月4日·[学院](#) - 广东第二师范学院 - 中国社会科学院大学 - 山东大学威海分校 - 参考资料来源：[3] 2013-2014年行政管理专业排名榜 院校名称 办学 办学专业 院校在1中山大学5★2562中国人民 ...
<https://baike.baidu.com/item/%E8%A1%8C%E6%94%B%E7...>

ALL

29

下载

1

学院

2

教师

2

小事儿

2

安卓版

2

点明

2

软件

1

同德

1

短基

1

微软

1

《输入“软件学院”的搜索结果》

清华大学

Submit

南台科技 **大学** - 南臺科技大学首頁

南臺科技大学首頁 Your browser does not support JavaScript!Your browser does not support JavaScript!Your brows ...

<https://www.stust.edu.tw/>

陈龙飞_工程热物理系_北京航空航天大学 **大学** 能源动力与工程学院

2016年11月18日 • **清华大学**、中国原子能所以及东风汽车公司等7家单位成功应用，以第一完成人，获法中委员会颁发的中法校企合作创新研 ...

[http://sepe.buaa.edu.cn/cpy/jscy/tjmsy/gzwtgsp/...](http://sepe.buaa.edu.cn/cpy/jscy/tjmsy/gzwtgsp/)

在渤海 **大学** 读书是一种怎样的体验? - 知乎

2016年1月19日 • **大学**，我爱你，甚名校，我爱你，物理系，我爱你，怀念你们。继续学习，以上。（此处应该有一张表达热内心热爱的情 ...

<https://www.zhihu.com/question/53325104?sort=creat...>

一二九| 忆往昔 又今朝 搜狐

2017年12月9日 • **大学**八届社会学系的学生，由于一张《大众生活》杂志封面的演讲照片而成为“一二·九”运动的象征人物。当时他是学生救国 ...

http://www.sohu.com/a/209504916_99903699

“学长学姐去哪儿了”——2017届毕业生之星候选人风采展—**百度**

2017年6月19日 • **大学**之前一样，努力地追求自己的梦想，并坚持了四年，且四年里专业成绩优秀，在最后拿到推免名额，考取至上海交通 ...

http://www.sohu.com/a/150299123_694747

行政管理专业_**百度**百科

2018年7月4日 • **大学**，-山东大学威海分校-参考资料来源：[3] 2013-2014年行政管理专业排名 序学校名称水 平开此专业学校跟1中山大学 ...

<https://baike.baidu.com/item/%E8%A1%8C%E6%94%BF%E7...>

UNIVERSITY OF TORONTO

All

43

英语

1

学院

2

大学

2

课程

1

热

1

在

1

又

1

去

1

远程教育

1

朱恒基

1

《输入“清华大学”的搜索结果》

python

Submit

七月在线 - 国内领先的人工智能教育平台

Python · Django实战》，并在个人网站idiffer.com上刊登有关技术类课程。清华大学西安交大应用数学博士西安交大应用数学专业博三，擅长机 ...

<https://www.julyedu.com/>

十条| 开发者公众号大全

Python 开发量大数据就业前景怎么样？附11免费学习全套资源！1小时前 02新Linux学习教你三招快速文件批量重命名方法1小时前 02新Linux ...

<http://t10tao.com/>

滴滴的几桩罪 - 硬娘说 - CSDN博客

2018年9月17日 • **Python** 资料免费领会员任意学Java薪资多少怎样才能不被裁员 登录注册点赞取消点赞0评论目录收藏手机看上一篇 更多上一 ...

[https://blog.csdn.net/qg_29679573/article/details/...](https://blog.csdn.net/qg_29679573/article/details/)

1

All

1

滴滴

2

七月

1

《输入“python”的搜索结果》

使用说明

(先安装好 nltk库, pymongo库, jieba库, requests库, bs4库, hashlib库, MongoDB, django)

1. 执行src/preprocess目录下的pachong.py文件(爬虫)

修改pachong.py文件的188行就能调整爬虫文档数。

```
188         if countNum % 65 == 1: (修改65)
```

(因为爬虫需要两三个小时, 为了让助教老师更方便运行我们的程序, 我把5222个文档打包之后放在src/preprocess/pachong目录下。)

2. 启动MongoDB

打开cmd, 进入MongoDB目录,

이름	수정한 날짜	유형	크기
bin	2018-12-28 오후...	파일 폴더	
data	2018-12-28 오후...	파일 폴더	
log	2018-12-28 오후...	파일 폴더	
new_data	2019-01-04 오후...	파일 폴더	
GNU-AGPL-3.0	2018-10-03 오전...	0 파일	34KB
LICENSE-Community.txt	2018-10-03 오전...	텍스트 문서	3KB
mongo.config	2018-12-28 오후...	XML Configuratio...	1KB
mongod.cfg	2018-12-28 오후...	CFG 파일	1KB
MPL-2	2018-10-03 오전...	파일	17KB
README	2018-10-03 오전...	파일	3KB
THIRD-PARTY-NOTICES	2018-10-03 오전...	파일	56KB

创建新的文件夹(比如, new_data)

进入MongoDB/bin目录

然后运行 mongod --dbpath (new_data文件夹的路径)

```
C:\MongoDB\bin>mongod --dbpath c:\MongoDB\new_data
```

就可以启动MongoDB。

3. 执行src/preprocess目录下的fenci.py文件

(注意! 运行程序的时候必须从pachong文件夹删除打包文件, pachong.zip)

运行fenci.py文件做

“分词”, “提取关键词”, “获取词干”,

“建立单词-文档权值矩阵和倒排索引”, “筛选关键词”, “停用词处理”,

“使用 Tf.idf 扩展布尔模型, 计算 mj, tf, dj, nk, idf, Wkj”

这些工作。

做完这些工作之后会把数据保存到MongoDB。

4. 打开django服务器

打开cmd, 进入src目录, 执行 `python manage.py runserver 0:8000 --noreload`

每次打开服务器的时候需要从数据库获取 `tf.idf`扩展布尔模型矩阵, 所以可能需要2~3分钟的时间。

5. 进入网站

打开浏览器, 访问 `127.0.0.1:8000`