

# 体育新闻整合与检索系统设计文档

励永辉

## 一. 系统介绍

体育新闻整合与检索系统主要包含球队热度榜与搜索界面，可以满足用户搜索相关新闻以及浏览 NBA 球队详细信息的需求。

## 二. 功能展示

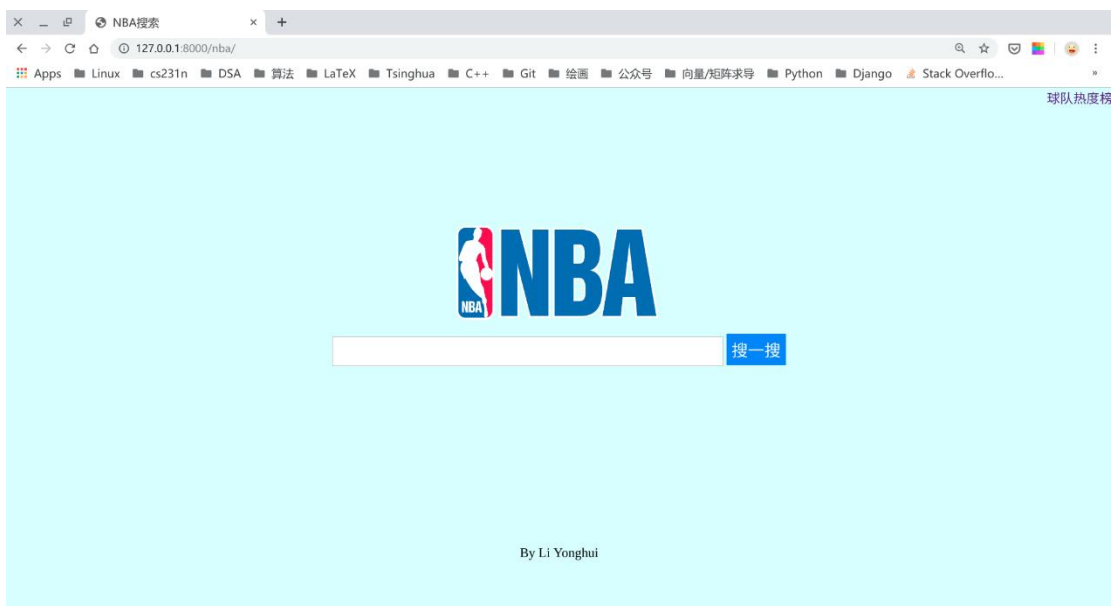
### a) 数据爬取

```
data.json
~/PycharmProjects/Python/easy-spider

{"title": "迈尔斯说道。\"这就是这家伙的竞争精神，他是一位竞争者。\"迈尔斯补充道。\"}，\n{\"title\": \"意气风发！戴维斯基家人出席联合国儿童基金会纪念晚宴\", \"name\": \"Instagram\", \"time\": \"2019-06-23 11:34:08\", \"content\": \"虎扑6月23日讯 湖人前锋安东尼-戴维斯今天更新Instagram分享他与家人参加联合国慈善晚宴的动态。\"感谢@联合国儿童基金会 今晚邀请我和我的家人出席他们的75周年纪念晚宴。同时祝贺@Toisalter，我们度过了美妙的时光。#uncfchicago\"戴维斯写道。值得一提的是，戴维斯发布的活动照中出现了其芝加哥老乡德里克-罗斯的身影。\"},\n{\"title\": \"魔术师：勒布朗世界最佳，AD会打出生涯最佳赛季之一\", \"name\": \"Twitter\", \"time\": \"2019-06-23 11:12:53\", \"content\": \"虎扑6月23日讯 湖人名宿魔术师约翰逊近日接受采访展望了湖人双星勒布朗-詹姆斯与安东尼-戴维斯的配合前景。\"勒布朗仍是世界最佳篮球运动员，\"魔术师说道，\"我认为当你在他身边配备安东尼-戴维斯这样的超级球星，他们二人都能尽情发挥并展现统治力。因为球场空间会被打开，勒布朗又是一名如此不可思议的传球手和突破手，他总能让队友变得更好。所以能够预见安东尼可能会打出其生涯最佳赛季之一。\"},\n{\"title\": \"丹尼-格林：希望继续帮球队冲击冠军，不知道科怀的选择\", \"name\": \"CBC\", \"time\": \"2019-06-23 11:03:43\", \"content\": \"虎扑6月23日讯 猛龙球员丹尼-格林在近日接受媒体采访时谈到了自己的休赛期选择。\"我只是想去一支能够保持竞争力的球队，能够赢球并且取得成功，希望这个地方会是这里（猛龙）。能够继续赢球，能够在每年都保持竞争力，这就是我想要做的事情。我想要继续赢球，我想要帮助球队冲击总冠军，我希望成为球队的粘合剂。\"格林说道。当被问及是否知悉队友科怀-伦纳德的决定去向，格林说：\"不，我一点都不知道，我实话实说。即便我知道了，我也不会告诉你，但我真的不知道。现在没有人知道，他将一切都藏在心里，即便是你认为你足够了解他，你依旧无法真正读懂他的内心。或许今天他想去这里，明天又想去那里，我们都不知道，或许只有他和他的家人知道。\"},\n{\"title\": \"缺前锋了？巴特勒自告奋勇：巴西队，美洲杯，我来了\", \"name\": \"Instagram\", \"time\": \"2019-06-23 09:55:30\", \"content\": \"虎扑6月23日讯 自由球员吉米-巴特勒近日发布Instagram Story动态表示自己正在启程前往观看美洲杯足球赛。\"美洲杯，我来了！既然历史最佳缺席了，我知道你们需要另一名前锋，所以我在路上！巴西队，我来了！美洲杯，我来了！\"巴特勒在自拍视频中说道。值得一提的是，巴特勒在视频中头戴一顶洛杉矶道奇队的棒球帽。此外，巴特勒还分享了自己作画创作巴西队10号球员形象的动态。 \"},\n{\"title\": \"伯德：80年代三分线就有了，直到最近这些年才被重视\", \"name\": \"美联社\", \"time\": \"2019-06-23 10:34:38\", \"content\": \"虎扑6月23日讯 NBA名宿拉里-伯德在近日接受媒体采访时谈到了当今联盟的比赛风格。\"在我看来，15年前联盟中盛行的是魔术师约翰逊这种2.06米的高大后卫，小个后卫的生存空间狭小，现如今的情况恰恰相反\"伯德说道，\"大个子球员逐渐成为边缘人物，小个球员正在接管联盟。上世纪80年代三分线就有了（NBA1979-80赛季正式引进三分线），但是没有人重视它，直到最近的15-17年。我还记得肯尼迪大学在里克-皮蒂诺教练的执教下不停地出手三分球，我当时的想法是，天哪，你这样是无法赢球的。而现如今，如果一支球队缺乏三分投射能力，你就无法赢球。\"\"事实上，当年我打球的时候，我们根本不会在三分线外防守。\"伯德说道。\"},\n{\"title\": \"火药味十足！Big3昔日罗伊斯-怀特与约什-史密斯冲突\", \"name\": \"Twitter\", \"time\": \"2019-06-23 10:57:10\", \"content\": \"虎扑6月23日讯 \"},\n{\"title\": \"ESPY最佳NBA运动员奖候选人出炉，哈登等球员入选\", \"name\": \"Twitter\", \"time\": \"2019-06-23 09:27:00\", \"content\": \"虎扑6月23日讯 根据The Athletic记者Alykhan Bijani的报道，詹姆斯-哈登入选2019年体育大奖ESPY最佳NBA运动员奖候选人名单。其他候选人还包括凯文-杜兰特、保罗-乔治和扬尼斯-阿德托昆博。ESPY大奖是对过去一年在体育场上表现最好球员的奖励，由ESPN在1993年创办。\"},\n{\"title\": \"马库斯-艾伦和科迪-米勒-麦金太尔将为湖人打夏季联赛\", \"name\": \"Twitter\", \"time\": \"2019-06-23 08:25:00\", \"content\": \"虎扑6月23日讯 根据湖人记者Harrison Foigen的报道，马库斯-艾伦和科迪-米勒-麦金太尔将为湖人出战夏季联赛。麦金太尔身高1.91米，体重93公斤，司职后卫，毕业于维克森林大学，是2016年的落选秀。艾伦身高1.88米，体重87公斤，司职后卫，毕业于斯坦福大学，是2017年的落选秀。2018-19赛季，艾伦在发展联盟的南湾湖人队场均出战20.9分钟，可以拿到8.5分2.5篮板1助攻。\"},\n{\"title\": \"不卑不亢！卡鲁索鼓励落选秀：落选了，别着急\", \"name\": \"Instagram\", \"time\": \"2019-06-23 08:44:57\", \"content\": \"虎扑6月23日讯 NBA年度选秀大会刚刚结束，湖人后卫瓦克斯-卡鲁索通过Instagram发布图文鼓励今年的落选秀们。\"落选了，别着急。\"卡鲁索写道。队友凯尔-库兹马则留言道：\"尊重。\"卡鲁索是2017年的落选秀，此前两个赛季与
```

用 scrapy 总计爬取新闻 5998 条，保存在.json 文件中。

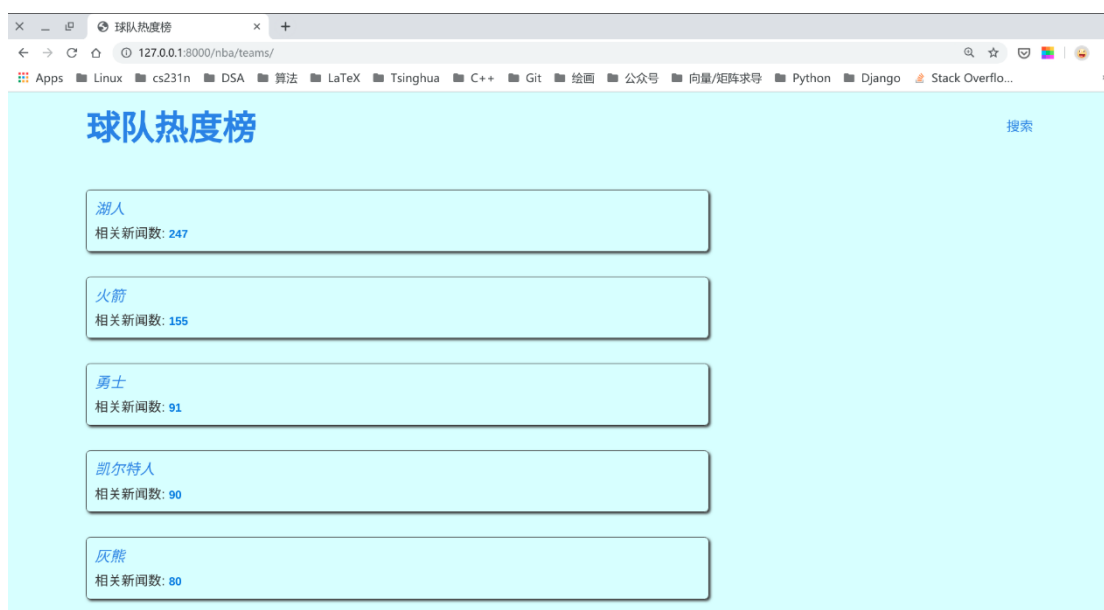
### b) 搜索





在搜索栏输入目标词汇，可获取相关新闻。支持多关键词搜索。匹配的词汇会高亮显示，便于查找。

### c) 球队热度版



页面包含一个列表，显示球队名称以及球队相关新闻数量等信息，并按照球队的相关新闻数量对球队排序。点击页面中的球队名称可跳转到球队主页。

#### d) 球队主页



球队主页包含两部分内容：基本信息与相关新闻。新闻数量较多时，进行分页等操作以方便用户交互

## e) 新闻详情页



每条新闻有一个新闻详情页，新闻详情页中显示新闻标题与全文。为新闻正文中包含的球队名与运动员姓名添加超链接，跳转到本系统的球队主页。

## 三. 性能统计信息

总共用爬虫爬取新闻总数 5998 条。但由于建立倒排索引的时候过于耗时，抽取其中 1500 条新闻建立倒排索引，所以搜索功能只能查询到 1500 条新闻中的内容。

高级搜索查询时间稳定在毫秒量级。

建立了 30 个球队的个人主页，记录了 554 名球员信息。

## 四. 设计思路

利用 Scrapy 爬取虎扑新闻, 过滤其中域名开头为 <https://voice.hupu.com/nba/> 的, 得到 NBA 新闻, 保存在 JSON 文件中。

Django 读取 JSON 中的文件, 保存在 Sqlite3 数据库中。

Django 有三个模型 Team、Person、News, 分别记录队伍信息、球员信息、新闻信息。

```
class Team(models.Model):
    name_text = models.CharField(max_length=200)
    time_text = models.CharField(max_length=200)
    city_text = models.CharField(max_length=200)
    news_int = models.IntegerField(default=0)
    def __str__(self):
        return self.name_text

class Person(models.Model):
    question = models.ForeignKey(Team, on_delete=models.CASCADE)
    firstname_text = models.CharField(max_length=200, null=True)
    lastname_text = models.CharField(max_length=200, null=True)
    def __str__(self):
        return self.lastname_text

class News(models.Model):
    team = models.ForeignKey(Team, on_delete=models.CASCADE)
    title_text = models.TextField(max_length=200) # Field name made lowercase.
    posttime_text = models.TextField(max_length=200) # Field name made lowercase.
    author_text = models.TextField(max_length=200) # Field name made lowercase.
    content_text = models.TextField(max_length=200000) # Field name made lowercase.
    def __str__(self):
        return self.title_text
```

## 五. 具体技术说明

### a) 分词

利用 jieba 库的 lcut 中对正文和标题进行分词, 建立正排表, 统计出文本中出现过的所有词汇。

```
all_words = []
for i in News.objects.all():
    cut = lcut(i.title_text+i.content_text)
    all_words.extend(cut)

set_all_words = set(all_words)
```

b) 索引

再遍历一遍建立倒排索引，把倒排索引的表储存在 JSON 文件中。

```
invert_index = dict()
for b in set_all_words:

    temp = []
    for j in range(1, News.objects.count()):

        field = News.objects.get(pk=j)

        split_field = lcut(field.title_text+field.content_text)

        if b in split_field:
            temp.append(j)
    invert_index[b] = temp
```

c) 搜索匹配程度

首先建立 TF-IDF，用来分析字词的重要性指标，评价多关键词搜索时文档与用户查询之间的相关程度。

GET 到用户的查询字符串后，利用 jieba 分词，再读取 TF-IDF 中列表中每个词的权重。然后对于每一篇文章，计算其所包含所有被搜索字符的总权重。对文章的总权重排序，降序输出。