**体育新闻整合与检索系统——说明文档**

20180211359 计84 乐阳

1. **功能展示**
2. 信息爬取与数据导入

* 将工作目录切换到spider文件夹中，执行 便可爬取虎扑体育NBA板块的100页共6000条滚动新闻，结果存储在test.json中。

python news.py

* 将工作目录切换到web文件夹中，执行初始化数据库（本程序链接的是本地mysql，数据库名为webServer）

python manage.py makemigrations

python manage.py migrate

接着执行以下语句可将爬取的新闻从test.json导入数据库，并建立索引。

python manage.py update

再执行以下语句可以从虎扑体育爬取并存储所有NBA球队信息，并计算热度

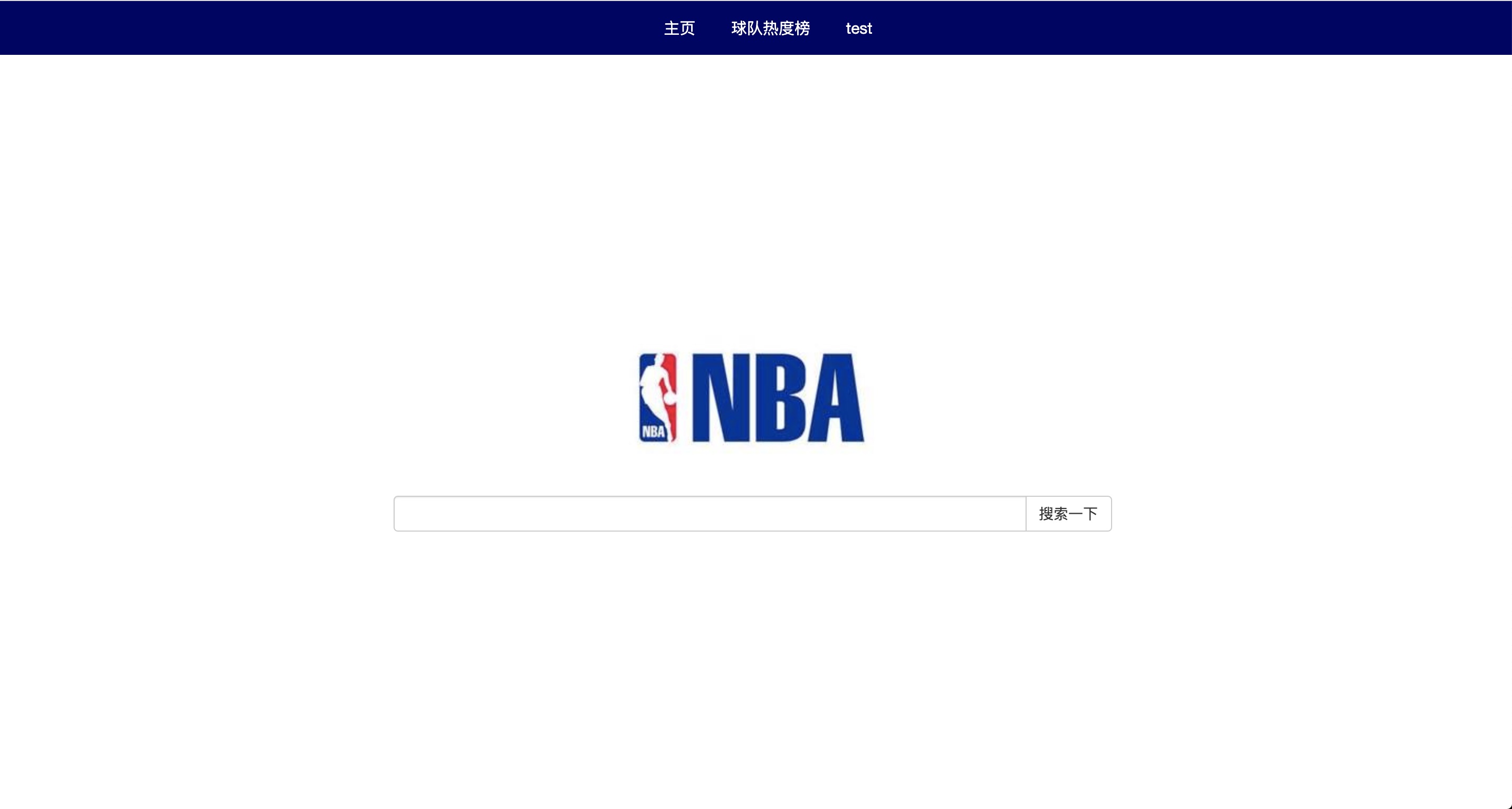
python manage.py get\_info

1. 网站界面

* 执行如下语句开启web服务器

python manage.py runserver 0:8000

在浏览器访问127.0.0.1:8000进入**主页界面**



* 主页界面由上方导航栏和主体的搜索框组成。

导航栏元素在所有网页上都有，可以返回主页、查看球队热度榜。

鼠标悬停于链接上方时，有选中效果。（下方黄色实线）



在搜索框中键入关键词，点击搜索按钮进入搜索结果展示页面。



* 搜索结果展示页面顶部仍有搜索框，以便重复搜索。

标题和正文节选中匹配关键词的部分予以标红高亮显示，

点击标题可以链接进入新闻全文页面。



多页信息支持分页显示。



* 新闻全文页面中包含新闻标题、时间、来源，以及正文的文字和图片



在新闻的正文中的球队名和球员名（全名）上添加有超链接，链接到相关的球队主页。

* 球队主页界面分为基本信息、球员名单以及有关新闻三部分。

球队基本信息部分包含球队名和城市，球队logo，进入NBA的时间以及简短的介绍。

球员名单部分含有所有球员的全名、号码以及照片。



有关新闻部分显示与球队及其核心球员有关的所有新闻，支持分页浏览。



* 球队热度榜界面按照球队的相关新闻数量从多到少建立榜单。点击球队名可以链接到球队主页。



鼠标悬停于球队名之上时会在下方显示球队logo和简介。



1. 附加功能：高级搜索

支持多关键词搜索。后端能够从用户输入的文本中中提取出关键词及其权值，进行综合加权搜索，返回最相关的新闻。

本功能在同一个搜索框实现。



1. **性能统计**

* 本系统共收录了6000条新闻。
* 经过测试，打开服务器后第一次搜索时由于jieba库需要加载模型，搜索较慢，勇士大概在0.7秒左右。
* 随后每次搜索用时均在0.1秒以内。

1. **技术说明**
2. 建立索引

* 考虑到jieba默认语料库的通用性，本系统手动添加了所有球队名进入词典。
* 对每条新闻建立索引时，对­­新闻标题和正文使用jieba.cut\_for\_search搜索引擎分词。对于词典中的每个词汇，在新闻标题中出现一次加权值50，正文中出现一次加权值1。计算得到每个词汇对应每条新闻的权值并保存在IndexInfo模型中。
* 搜索单个词汇时，从索引库中找相应词汇以及其索引，按权值倒序排列返回结果。

1. 文本搜索流程

* 搜索过程中，先对用户的输入调用jieba.analyse.extract\_tags获取其关键词和关键指数(内部为TF-IDF机制)。随后对每个关键词进行搜索，将所有关键词的权值与关键指数之积相加。最后进行倒排并返回结果。
* 由于extract\_tags过程中会自动过滤掉某些词性的词汇，因此在极少特殊情况（比如单搜一个“的”字）下可能出现无结果的情况。若第一步搜索结果为0，则进入第二步搜索，直接对文本调用jieba.cut获取其分词段进行搜索（不提取关键词）。

1. 球队搜索流程

* 我们将每个球队的搜索关键词定为球队的简称以及前5名关键球员的名字(last name)。将球队简称的权值定为3，每位球员名的权值定为1。进行加权搜索，最后得到与球队最相关（权值和最高）的新闻。