

NBA新闻整合与检索系统设计文档

程序介绍

设计目的

身处信息时代，数据的搜集、整合和检索是大数据技术的第一步。由于Python这个解释型语言的特性和库的丰富性，很适合用来进行爬虫的编写和作Web服务器的后端。使用不限制第三方库的Python编写数据搜集、整合和检索系统，有助于增加对Python的熟练度，并初步了解爬虫与Web服务器的搭建方法。

本程序的设计分为以下几个模块：(1)数据爬取

从虎扑爬取NBA板块（voice.hupu.com/nba）的**新闻信息**，内容包括新闻的标题、来源、发布时间、正文文本。

从虎扑NBA的球队主页爬取**球队信息**，包括球队名称、所在城市、球员。

(2)Web前端

对爬取的数据进行预处理并储存，以利于在其前端的Web网页上合理展现。

建立Web前端合理的地址和跳转，包含以下页面：

①**球队主页**，包括**球队基本信息**和与**球队的相关新闻**，相关新闻以列表的形式展现标题和部分正文，标题是可点击的**超链接**形式，点击后跳转到**新闻详情页**。

②**新闻详情页**，包含新闻的标题、正文、来源和发布时间，其中**正文**中的**球队**和**球员名称**标注上**超链接**，链接指向球队或球员所在**球队主页**。

③**球队热度榜**，显示**球队名**和**相关文章数量**，按相关文章数从高到低**排名**，**球队名**标注**超链接**调到对应**球队主页**。

新闻列表合理**分页**。

使用CSS等对页面布局进行**美化**。

(3)简易搜索引擎

仿照常规的搜索引擎界面，实现输入**单关键词**，筛选并分页显示标题或正文含有关键词的新闻列表，页面中需显示**查询结果数量**和**查询时间**，搜索结果列表中新闻标题和正文中关键词**高亮**显示，从搜索结果列表中也能**链接到新闻详情页**；在单关键词的基础上在同一页面加入**高级搜索**，允许用户输入**多个关键词**或**含多个关键词的句子**，查询与输入内容**相近**的若干条结果，多关键词搜索的文本**不一定要与新闻文本完全匹配**。应当计算查询文本和新闻文本的**匹配程度**，**优先展示匹配程度高的新闻**，匹配程度应当考虑到**不同词的重要程度不同**。

(4)爬虫的Web控制

将爬虫和Web系统集成成为一个系统，并增加管理页面，包含**按钮控制系统在后台开始/暂停爬虫与数据预处理工作**。在爬虫进行时，不影响前台**Web页面的访问**，并且新爬取的新闻在**预处理结束后能够更新到Web系统中显示**（更新球队的相关新闻、新闻能被搜索到、更新球队热度榜）。

程序使用方法与功能

程序已在zhengkw.com域名上部署，为了使外来ip能够通过此域名访问Web前端，需要将newsweb/newsweb下settings.py中ALLOWED_HOSTS = [] 改为ALLOWED_HOSTS = ['zhengkw.com', '127.0.0.1'] 以设定允许的域名来源。

在newsweb根目录下运行命令行指令 `python manage.py runserver (ip:)port` 为指定的ip地址和端口开放Web前端访问，若ip留空默认对本机地址127.0.0.1开放，port默认为8080。这里设为0.0.0.0:80，对所有主机开放默认Http的80端口。

在newsweb/newsscrapy目录下运行 `python run.py` 以开启爬虫后台。

访问zhengkw.com/index，访问Web主页

HomeSearch

NBA新闻（共 6164 条）

白曼巴：当年詹姆斯想错了，他本应关注勇士而非交易乐福

虎扑9月9日讯 近日，前NBA球员“白曼巴”布莱恩·斯卡拉布莱恩接受采访谈到了他在勇士的执教经历以及对于骑士和勒布朗·詹姆斯的一些看法。斯卡拉布莱恩曾于2013-14赛季担任勇士助教。勇士在那个赛季取得51胜的常规赛战绩，随后在季后赛首轮中被快船淘汰（总比分3-4落败）。记者询问他在执教勇士期间是...

科比：和佩林卡有所共鸣，自己不参与处理湖人球队事务

虎扑9月9日讯 近日，前NBA球员科比·布莱恩特在接受媒体采访时，被问及是否会协助湖人总经理罗勃·佩林卡处理球队事务。“不，我远离球队事务。很显然，我和佩林卡关系很好，我们经常交流，我和他有着共鸣。”科比说道。“佩林卡有足够的能力来处理好他的工作，当然这份工作本身也有一定压力。”科比补充道。 科...

帅气牛仔！独行侠官方晒诺维茨基观战达拉斯牛仔照片

虎扑9月9日讯 本周，美国四大运动联盟之一的橄榄球赛场正式拉开了比赛帷幕。前独行侠球员德克·诺维茨基和妻子等人一起前往现场，观看了达拉斯牛仔队在本赛季的开幕战。今日，独行侠官方更新了社交媒体Instagram，晒出了一张诺维茨基身着西装的现场照片。“[举手鼓掌emoji]巨星来了。”独行侠写道。 ...

掘金官方更新社媒祝贺博比-琼斯入选名人堂

球队热度榜

相关文章数目

湖人	1900
凯尔特人	1330
奇才	1283
火箭	1281
雄鹿	1182
快船	1178
太阳	1172
活塞	1145
勇士	1130
国王	1050
爵士	1044
开拓者	1030

左侧为新闻列表（按发布时间倒序，每页显示十条新闻），右侧则显示了球队热度榜，最上方有菜单栏。将页面拖至最下方

活塞官方晒格里分照片：李刚赛就在30大后

虎扑9月9日讯 昨日，活塞官方更新了社交媒体Instagram，晒出了一张球队前锋布蕾克·格里芬手持篮球的照片。“第一场季前赛还有30天就要来啦。” 2018-19赛季常规赛，格里芬场均上场35.0分钟，得到24.5分7.5篮板5.4助攻；季后赛场均上场29.0分钟，得到24.5分6.0篮板6.0...

乔-拉科布以2910万美元价格购置马里布海滩一处豪宅

虎扑9月9日讯 根据美国媒体Variety的报道，勇士老板乔-拉科布以2910万美元的价格在加利福尼亚马里布地区购买了一处豪宅。这座面积为5512平方英尺（约512平方米）的豪宅坐落在著名的Carbon海滩旁，室内和室外生活环境融为一体，拥有五间卧室和六间浴室，配套家庭影院、私人电梯和健身房。该...

新司机上路！开拓者新秀利特尔更新社媒晒奔驰车

虎扑9月9日讯 今日，开拓者新秀纳西尔·利特尔更新个人社媒Instagram，晒出一张自己与一辆奔驰汽车以及三辆自行车的合影。“美满生活，美好行程。@梅赛德斯-奔驰”利特尔写道。利特尔今年19岁，在今年选秀大会上首轮第25顺位被开拓者选中。在北卡罗来纳大学的一个赛季中，他场均上场18.2分钟，得...

千锤百炼！训练师发布拉塞尔训练时的照片

虎扑9月9日讯 今日，勇士球员丹吉洛-拉塞尔的训练师Chris Brickley发布了一组拉塞尔在训练时的照片。2018-19赛季常规赛，拉塞尔场均出场30.2分钟，得到21.1分3.9篮板7.0助攻1.23抢断。

←1234567891011→

点击数字或上一页/下一页可跳转分页，不再演示。点击第一条新闻

HomeSearch

白曼巴：当年詹姆斯想错了，他本应关注勇士而非交易乐福

发布时间：2019-09-09 17:04:30来源：HoopsHype

虎扑9月9日讯 近日，前NBA球员“白曼巴”布莱恩·斯卡拉布莱恩接受采访谈到了他在勇士的执教经历以及对于骑士和勒布朗·詹姆斯的一些看法。

斯卡拉布莱恩曾于2013-14赛季担任勇士助教。勇士在那个赛季取得51胜的常规赛战绩，随后在季后赛首轮中被快船淘汰（总比分3-4落败）。记者询问他在执教勇士期间是否预见到勇士接下来连续5年进入总决赛并3次夺冠的王朝前景，斯卡拉布莱恩也给出了自己的想法。

“我提前想到了！我在实现梦想。我做的就是我曾向我妻子承诺我会做的事。我知道自己将成为波士顿凯尔特人队（斯卡拉布莱恩2008年代表绿衫军夺冠）的比赛解说员，我也清楚解说工作时我余生想做的事。然后，在我前往那里就职前，我观看了前一年勇士的季后赛比赛。我看到了格雷格·波波维奇的脸色，就是那一刻改变了我的想法。勇士对决马刺，斯蒂芬·库里和克莱·汤普森就在场上，而格雷格·波波维奇的脸上出现了这种神情，‘我所做的一切都无比正确，但我正落后18分。我们的防守策略完全按照计划执行，但这些家伙正在做着我从没见过的事。’那个神态出现的时间很短，摄像头只对准了他几秒钟，但当我看到那种脸色，我可以认为他在冥思苦想，‘我需要采取跟我学到的所有篮球理念完全相反的战术思路吗？因为这两个后卫现在正手感逆天地命中投篮，他们做到的事情是前所未有的。’他们还拥有安德鲁·博古特，我非常欣赏博古特。当时他在常规赛打得不怎么好，但由于规则不再允许使用hand-checking的防守方式，我觉得良好的护筐能力是必要的。那时候我对德雷蒙德·格林略知一二，但不像现在这般了解；我并不知道他将改变我们打球的方式。我和鲍勃·迈尔斯（勇士总经理）也有交情，马克·杰克逊（时任勇士主帅）和我非常熟悉，我在公牛打球时，Pete Myers（时任勇士助教）也在那边工作，所以我对他也有所了解。那就有点像是雪崩效应。我知道我想加入他们的教练组。我说服自己别加入电视转播圈，我想追逐总冠军，而且我认为那个赛季我们就将赢得冠军，然后我可能会去从事别的行业。我永远不会忘记波波维奇的那个脸色。最终，马刺赢得了那轮系列赛（总比分4-2取胜）。如果你还记得，克莱本来防死了托尼·帕克，然后他犯满离场，马刺随后逆转取胜。”斯卡拉布莱恩说道。

显示了新闻详情，其中的队名和球员们标注了超链接，点击可跳转到球队主页，不再演示。点击菜单栏的Home返回，点击右端热度榜上排名第一的湖人

HomeSearch

湖人队

相关新闻（共1900条）

白曼巴：当年詹姆斯想错了，他本应关注勇士而非交易乐福

虎扑9月9日讯 近日，前NBA球员“白曼巴”布莱恩·斯卡拉布莱恩接受采访谈到了他在勇士的执教经历以及对于骑士和勒布朗·詹姆斯的一些看法。斯卡拉布莱恩曾于2013-14赛季担任勇士助教。勇士在那个赛季取得51胜的常规赛战绩，随后在季后赛首轮中被快船淘汰（总比分3-4落败）。记者询问他在执教勇士期间是...

科比：和佩林卡有所共鸣，自己不参与处理湖人球队事务

虎扑9月9日讯 近日，前NBA球员科比·布莱恩特在接受媒体采访时，被问及是否会协助湖人总经理罗伯·佩林卡处理球队事务。“不，我远离球队事务。很显然，我和佩林卡关系很好，我们经常交流，我和他有着共鸣。”科比说道。“佩林卡有足够的能力来处理他的工作，当然这份工作本身也有一定压力。”科比补充道。科...

对霍华德有何建议？罗斯：我想说的就是保持耐心

虎扑9月9日讯 活塞球员德里克·罗斯在今日接受电话采访时谈到了湖人中锋德怀特·霍华德。当被问及对霍华德有何建议时，罗斯表示：“我想说的就是保持耐心，每个人的情况都有所不同，我给出的建议是从我个人角度出发。我并不知道自己要经历什么，但上天对我最大的眷顾就是我冷静的性格。我总是告诉自己，我是一名职业...

球队基本信息

所在城市

洛杉矶

球员

贾维尔·麦基
安东尼·戴维斯
肯塔维厄斯·考德威尔·波普
埃弗里·布拉德利
特洛伊·丹尼尔斯
丹尼·格林
塔伦·霍顿-塔克
贾里德·杜德利
德马库斯·考辛斯
亚历克斯·卡鲁索
凯尔·库兹马
奎因·库克
科斯塔斯·阿德托昆博
拉简·隆多
小扎克·诺维尔
勒布朗·詹姆斯
德怀特·霍华德

左侧显示相关新闻。右侧则是球队基本信息，新闻标题同样可点击跳转。再测试搜索功能，点击上方菜单栏上的Search

HomeSearch

NBA

Search...

搜索

此时是搜索内容为空的状态。测试单关键词搜索，输入“詹姆斯”，点击搜索



詹姆斯

搜索

相关新闻数量：681条

搜索用时：0.00421863秒

搜索结果

Taco Tuesday 2.0：詹姆斯邀浓眉来到家里共进晚餐

虎扑7月17日讯 今日，湖人球员勒布朗·詹姆斯Instagram晒出在家中与队友安东尼·戴维斯共进晚餐的视频。

多诺万·米切尔将继续参加美国男篮训练营

詹姆斯·哈登、埃里克·戈登、CJ·麦科勒姆、布拉德利·比尔、达米安·利拉德、德马尔·德罗赞和托拜厄斯·哈里斯等球员相继宣布不会代表美国队参加今夏的男篮世界杯。据此前的相关报道，谈到收到美国男篮的邀请，米切尔说：“我在高中和大学的时候一直都想要代表国家出战，但一直没有获得机会。所有事情的发生都是有原因的，这只是时间问题，能够收到邀请我心怀感激。现在我需要努力训练，努力成为最终大名单的一员。”20...

享受生活！詹姆斯社媒动态展示自家种植番茄和上球鞋

显示相关新闻数量、搜索用时和分页的搜索结果列表。再测试多关键词搜索，输入“勇士与火箭”

勇士与火箭

搜索

相关新闻数量：1168条

搜索用时：0.011420159999999999秒

搜索结果

疯狂休赛期！上赛季最佳阵容15人中6人休赛期换队

与休斯顿火箭队达成交易，雷霆送出拉塞尔·威斯布鲁克从火箭得到克里斯·保罗以及多个首轮选秀权以及选秀互换权。在加上威斯布鲁克之后，上赛季最佳阵容中的15名球员中已经有6人在休赛期已经更换球队，分别是：保罗·乔治（雷霆至快船）、凯文·杜兰特（勇士至篮网）、科怀·伦纳德（猛龙至快船）、凯里·欧文（凯尔特人至篮网）、肯巴·沃克（黄蜂至凯尔特人）以及拉塞尔·威斯布鲁克（雷霆至火箭）。2018-19赛季最...

ESPY颁奖典礼明日在洛杉矶举行，字母哥哈登皆获提名

火箭凯文·杜兰特，金州勇士保罗·乔治，俄克拉荷马雷霆布莱安娜·斯图尔特，西雅图风暴戴安娜·陶莱西，菲尼克斯水星艾琳娜·唐尼，华盛顿神秘人坎黛丝·帕克，洛杉矶火花密尔沃基雄鹿队球员扬尼斯·阿德托昆博获得提名 亚特兰大老鹰队球员特雷·杨获得提名 杜克大学蔡恩·威廉森获得提名 克莱汤普森14记三分打破NBA三分纪录获得提名 科怀·伦纳德东部半决赛G7压哨绝杀76人获得提名 达米安·利拉德赛季...

鹈鹕有意追求凯文·卢尼，而勇士方面希望尽快留下他

火箭外，鹈鹕也是有意追求凯文·卢尼的球队之一。Deveney还表示，他被告知勇士希望能尽快留下卢尼。2018-19

前几条为“勇士”“与”“火箭”三个关键词均包含的新闻。我们到最后一页（第117页）

道：“我能完美融入。球队喜欢拉开空间，我能更好地攻击对手和突分。在防守端，我能够换防和抢...

莫雷：哈登保罗关系紧张报道不是真的，我已与保罗会面

火箭总经理达雷尔·莫雷在当地时间本周一与克里斯·保罗完成了会面，他在昨日NBA颁奖典礼上接受了采访。谈到关于保罗和詹姆斯·哈登之间关系紧张报道，莫雷说：“这肯定不是真的。在我职业生涯里，媒体总是最奇怪的一部分，之前有一段时间我们理应得到这样糟糕的报道，但是现在，现在的讨论应该是我们是下赛季最被看好的球队，所以这样的报道很奇怪，之前出现了这样的报道，然后一直都在出现这样的报道。”“交流很通...

哈尔滕施泰因：得知伤势没那么糟后让我松了口气

火箭中锋以赛亚·哈尔滕施泰因被诊断为右脚腓一二级扭伤，很可能将缺席2-3周的时间。在火箭与国王的拉斯维加斯夏季联赛中，哈尔滕施泰因在一次落地时右脚受伤，倒在火箭板凳席前，随后被队友搀扶下场。根据《休斯顿纪事报》记者Jonathan Feigen的报道，目前哈尔滕施泰因右脚已经穿上了保护靴，他称得知伤势情况没有那么糟后让他松了口气，在夏季联赛结束后，会有伤病防范理疗师在德国帮助哈尔滕施泰因恢复，他...

塔克谈美国队训练营：年轻球员更要注意场上的小细节

火箭球员PJ·塔克和爵士球员多诺万·米切尔在美国队训练营期间接受了采访。谈到训练营几天对抗赛内取得的进步，塔克说道：“最大的进步是大家都在变得更自如，开始知道波波维奇想要什么，知道他想让我们如何打球。现在大家在防守端都更有侵略性了，更注意挡拆的配合，避免犯规，这些小细节。对于年轻的球员来说这可不是小事，因为他们运动能力太强，打得很快。”多诺万·米切尔谈到了在训练营跟随多位知名教练训练的感受，他...

← 107 108 109 110 111 112 113 114 115 116 117 →

这时的新闻从含有“火箭”和“与”到只含有“火箭”，相关度达到最低（只含有“与”的新闻由于相关度过低，没有被列入）。我们再搜索“湖人的詹姆斯和勇士球队打篮球”，前几条新闻为

湖人的詹姆斯和勇士球队打篮球

搜索

相关新闻数量: 3526条

搜索用时: 0.02468747秒

搜索结果

考辛斯：今年夏天有很多意想不到的事情发生，这令人兴奋

...湖人，与肯塔基大学校友安东尼-戴维斯还有一个叫勒布朗-詹姆斯的家伙并肩作战。谈到自由球员市场的疯狂，考辛斯说：“有很多意想不到的事情发生，我从未经历过像今年这样的夏天，我很确信你们整个夏天都看得津津有味，很多新面孔和新球队，这令人感到兴奋。”关于TBT球队的建队理念，考辛斯说他喜欢身高体长和技术全面的球员，他的首发球员身高都达到6尺10寸，其中包括他的兄弟Jaleel和在肯塔基大学效力时的队友...

...詹姆斯给我发来很长的欢迎短信

...的23号球衣。我说，‘哇，谢谢你，Bron。’然后就是规则和其他方面发生了许多事，所以我无法得到23号。就是耐克和勒布朗球衣的事情，他们的态度就是，‘我不管你们是谁，牵涉的资金损失太大了。’我有点伤心。”谈到自己挑选3号作为球衣号码的过程，戴维斯说：“那是我小学时的号码，也是我打篮球以来使用的第一个号码。不过我选择3号的过程是这样的。我打开了NBA 2K游戏，我没法选择一个号码。因为23号不能...

...詹姆斯和加内特

...湖人前锋安东尼-戴维斯近日接受专访谈到了队友勒布朗-詹姆斯。谈到他与詹姆斯的友谊，戴维斯说：“自从2012年起我就和勒布朗有私交了，自从第一次打篮球开始我就是他的粉丝。从没看过迈克尔-乔丹打球，所以我一直以来的崇拜对象就是勒布朗和KG（凯文-加内特）。现在可以和勒布朗在整个赛季内携手作战，我们的友谊会得到加强，希望我们能在洛杉矶成就特殊的事业。”谈到自己为球队带来的补充，戴维斯说：“我自认为是...

包含了尽量多的关键词。而最后一页（第353页）

攻。

官方：奇才签约以赛亚-托马斯

...球队和以赛亚，他渴望展示出他已经找回了之前的样子，那种之前让他成为了这个联盟最高效和最特别球员之一的样子，我们可以为他提供这样的机会，让他在我们球队理念的框架之内这么做，我们看重他的领导力和经验。”奇才篮球运营高级副总裁Tommy Sheppard说道。根据之前NBA记者Adrian Wojnarowski的报道，小托马斯与奇才达成了一年的协议，小托马斯来自华盛顿的塔科马。2018-19赛季常...

...的TJ在同一支球队了，教练组会为此伤脑筋

...球队新赛季的期待等相关话题。“我在过去的四年一直作为对手和他们比赛，我能看到他们打球的方式、内特（麦克米兰）教导的方式，以及他们所拥有的球员，综合所有因素，当然这里的球迷基础也非常好。”我认为这是一支很好的球队，年轻、有天赋。我认为有了现有的这些成员我们可以一起做一些很特别的事。维克托（奥拉迪波）、道格（麦克德莫特）、迈尔斯（特纳），更衣室所有的球员和他们之间形成的友好的气氛，我很期待成为他们...

联盟高管：若雷霆明夏交易威少乔治，回报可能会非常少

...的首轮签、1个受保护的首轮签和2个选秀权互换的权利。随后雷霆将拉塞尔-威斯布鲁克交易到火箭，得到克里斯-保罗，火箭的2024年和2026年的首轮签和2021年和2025年首轮签互换权。根据Bleacher Report记者Ken Berger的报道，一位对手球队高管透露，如果乔治没有提出交易请求，雷霆留下乔治威少再战一个赛季，然后尝试在明夏交易他们两人，那么雷霆得到的回报也许只有他们现在得到回...

←

343

344

345

346

347

348

349

350

351

352

353

→

只包含了“球队”“的”“和”这样匹配度很低的新闻。

最后时爬虫控制页面，访问zhengkw.com/spider

爬虫当前状态：运行

暂停爬虫

页面为简单的指示和按钮，点击可控制后台爬虫运行状态。

性能统计

性能状态在上一页已展示完毕。总新闻数量越6000，单关键词搜索时用时为0.001s数量级，多关键词搜索则达到0.01s数量级，关键词越多、句子越长则用时越长。

进入Django的SQLite3数据库后台（zhengkw.com/admin），存储分词的数据库内约27500记录



程序设计

技术基础

本程序在PyCharm中编写测试，共约1000行代码（.py/.html，含自动生成），Python版本为3.7.4。

- Semantic UI
 - Semantic UI是一个优秀的前端开发框架，本程序主要使用了其CSS样式文件和图标。
 - 顶部菜单栏——ui menu,item
 - 页面左右布局——ui grid,XX wide column
 - 主页、球队主页使用10:2(空白):4的比例设置
 - 搜索页面、新闻详情页使用3(空白):10:3(空白)的比例设置
 - 新闻列表的每一框——ui container segment
 - 水平分割线——ui horizontal divider
 - 分页菜单栏——ui pagination menu,icon item,item
 - 右方表格——ui table
 - 搜索表单——ui action input
- Django
 - 本程序使用了Web框架Django最基础的MTV模式和其自带的SQLite3数据库
 - Model
 - Team包含自增的id字段，name、city、players字段分布储存球队名称、所在城市、球员基本信息，relatedids字段储存相关文章的id
 - Player包含name、team，分布是球员名字和所在球队名称
 - Article包含自增的id，以及title、time、source、content存储文章所有信息
 - Word用于存储倒排索引，word字段是词名，relatedids记录含有此词的文章id
 - Spider用于进行爬虫的Web控制，完成Scrapy和Django两个进程间的通信
 - Template
 - 共5个网页模板，存储在newsweb/newsapp/templates文件夹下，在newsweb/newsweb/Settings.py中添加

```
TEMPLATES = [  
    {  
        'BACKEND':  
        'django.template.backends.django.DjangoTemplates',
```



```

        'DIRS': [os.path.join(BASE_DIR,
'templates').replace('\\', '/')],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',

'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]

```

- 完成模板加载。之后在View中使用模板templates/xx.html可直接调用 `return render(request, "xx.html", context)` 完成渲染。
- 在newsweb/newssapp/static文件夹下存储了模板要用到的CSS和图片资源文件。在newsweb/newsweb/Settings.py中添加

```

STATIC_URL = '/static/'
STATICFILES_DIRS = (
    # os.path.join(BASE_DIR, 'static')
    os.path.join(os.path.dirname(__file__),
'../static/').replace('\\', '/'),
)

```

- 之后在模板文件头部添加 `{% load staticfiles %}`，便可使用 `{% static 'filename' %}` 调用static文件夹下的资源文件。

■ View

- Viem层设置了6种URL路由
- /index——重定向到/index/1
- /index/page_num——访问主页面，其中新闻列表为page_num页
- /team/nid/page_num——编号为nid的球队，第page_num页
- /news/nid——第nid条新闻
- /search——搜索页面，使用GET获取搜索内容和页码
- /spider——爬虫控制页面

• Scrapy

◦ Scrapy是一个方便的爬虫框架

- 本程序使用了其Spider-Pipeline结构爬取球员数据和新闻数据。
- 在Spider中可以使用xpath结点路径定位，本程序使用了css标签属性定位的方法。
- 在Pipeline中对数据进行入库和预处理操作（分词后入库）。
- 在Scrapy中可直接使用Django的Model层，方法是在newsweb/newsscrapy/Setting.py中添加

```

import django
sys.path.append(os.path.dirname(os.path.abspath('.')))
os.environ['DJANGO_SETTINGS_MODULE'] = 'newsweb.settings'    # 项目
名.settings
django.setup()

```

- 之后在Pipeline中from newsapp(Django app名称).models import xx即可
- 在Scapy的Web控制上，本程序使用Django的Model层数据实现进程间通信。
 - newswb/newssrapy/newscrapy/spiders文件夹下存储了用于新闻爬取的新sspider。
 - 在newswb/newssrapy目录下的命令行中键入scrapy crawl newsspider可启动newsspider，每个newsspider运行的一个轮次只顺序爬取1-10页的新闻一遍。而newsspider内部根据Django数据库中的标识，判断爬虫运行情况，若运行则进行文章的解析预处理入库操作。
 - newswb/newssrapy/run.py是爬虫的后台控制端，它会等待每一个轮次运行完成，等待一分钟运行下一个轮次，循环往复。
- jieba
 - jieba是优秀的中文分词库，是本程序搜索功能的顶梁柱。本程序主要使用了其中的两个功能
 - 分词。在文章的预处理中，先对其标题和正文分词。再建立倒排索引（将与某词相关的文章id列表使用JSON保存至数据库）。其中分词使用了jieba.cut_for_search()函数，这是一种专为搜索引擎服务的较为细致的分词。本程序的匹配程度没有考虑词在文章中的出现次数，如“詹姆斯与篮球”，即使“与”在文章中出现100次，仍没有一个“詹姆斯”或“篮球”有效力。
 - TF-IDF。TF-IDF是一种用于信息检索与数据挖掘的常用加权技术，可用于句段中每个词语的重要性评估。使用jieba.analyse.extract_tags()函数可调用jieba内置的TF-IDF权重打分系统。本程序的对于每篇文章的匹配程度评分是简单的每个在文章中出现的关键词的权重相加。

细节处理

- 列表排序
 - 有时列表的元素不是简单的可比较对象，可能是元组，可能是Model。要根据元组的某个分量，或根据Model的某个字段排序，可结合sorted()函数与lambda表达式
 - 根据文章的发布时间倒序排列（结合Python的列表推导）

```
articles = sorted([ars[0] for ars in
[Article.objects.all().filter(id=id) for id in ids]], key=lambda x:
x.time, reverse=True)
```

- 根据元组列表中元组的第一个分量倒序排列

```
relist = sorted(relist, key=lambda x: x[1], reverse=True)
```

- 元组列表的遍历
 - 在Template中的语句只支持简单的if/for遍历Model.field等，并不能进行运算和[]等符号的支持。对于元组列表的遍历，如

```
rank = []
for team in Team.objects.all():
    idjson = team.relatedids
    ids = json.loads(idjson)
    rank.append((len(ids), team.name, team.id))
rank = sorted(rank, reverse=True)
```

- 将rank通过context传入模板后可以使用如下方法遍历


```
{% for num,name,id in rank %}
...
{% endfor %}
```

- 分页菜单栏的生成和分页页码范围

- 后端：使用Paginator结合简单的逻辑分支确定每页的内容和下方可选的页码范围。

```
paginator = Paginator(articles, 10) #每页10条新闻
page_num = int(pn)
try:
    current_list = paginator.page(page_num)
except:
    raise Http404("Page not found")
if paginator.num_pages > 12: # 如果分页的数目大于12
    if page_num - 5 < 1: # 你输入的值
        pageRange = range(1, 12) # 按钮数
    elif page_num + 5 > paginator.num_pages: # 按钮数加5大于分页数
        pageRange = range(page_num - 10, page_num + 1) # 显示的按钮数
    else:
        pageRange = range(page_num - 5, page_num + 6) # range求的是按钮数 如果你的按钮数小于分页数 那么就按照正常的分页数目来显示
else:
    pageRange = range(1, paginator.num_pages + 1) # 正常分配
```

- 前端：上一页和下一页，循环生成页面按钮，并使用Paginator分页得到的列表自带函数动态判断上一页/下一页的disable状态。

```
<div class="ui pagination menu">
{% if current_list.has_previous %}
    <a href="{% url 'index' current_list.previous_page_number %}"
class="icon item">
        <i class="left arrow icon"></i>
    </a>
{% else %}
    <a class="disabled icon item">
        <i class="left arrow icon"></i>
    </a>
{% endif %}
{% for i in pageRange %}
{% if current_num == i %}
    <a href="{% url 'index' i %}" class="active item">
        {{i}}
    </a>
{% else %}
    <a href="{% url 'index' i %}" class="item">
        {{i}}
    </a>
{% endif %}
{% endfor %}
{% if current_list.has_next %}
    <a href="{% url 'index' current_list.next_page_number %}"
class="icon item">
        <i class="right arrow icon"></i>
    </a>
{% else %}
```

```

<a class="disabled icon item">
  <i class="right arrow icon"></i>
</a>
{% endif %}
</div>

```

- 时间日期转换

- models.DateTimeField()与Python中的datetime库是对应的关系，存储日期和时间信息。使用这个数据类型方便进行排序等操作。
- 将表示日期时间的文本转化为datetime类型可使用datetime库中的 datetime.strptime() 函数

```

dt = datetime.datetime.strptime(response.css("a.time
span::text").extract_first("1000-01-01
00:00:00").replace("\r", "").replace("\n", "").strip(), "%Y-%m-%d
%H:%M:%S")

```

- 将datetime类型格式化为文本可使用datetime类型的 strftime() 函数

```

time = article.time.strftime("%Y-%m-%d %H:%M:%S")

```

- 正文截取

- Django自带stringfilter truncateword，如在模板中使用 {{ content|truncateword:'150' }} 会自动限制词数不超过150，并将之后的文本用"..."显示。但这个过滤器只能用于英文文本的截取。
- 为实现中文段落的截取，实现了文本过滤器truncatehanzi，在 newweb/newspapp/templatetags/filter.py中。文件结构为

```

# coding=utf8
from django.template import Library
from django.template.defaultfilters import stringfilter

register = Library()

@stringfilter
def truncatehanzi(value, arg):
    """
    Truncates a string after a certain number of words including
    alphanumeric and CJK characters.
    Argument: Number of words to truncate after.
    """
    ...
    register.filter('truncatehanzi', truncatehanzi)

```

- 这样便完成了过滤器标签的注册。在模板头部使用 {% load filter %}，正文部分使用变量 {{ content|truncatehanzi:'150' }} 会自动显示截取后的文本内容。

- 搜索结果关键字高亮

- 搜索结果的高亮同样由templatetags来完成，相关文件包括 newweb/newspapp/templatetags/highlight.py和newweb/newspapp/newshighlight.py。这两个文件主要借鉴了Haystack（一个全文搜索插件，之前本打算使用Whoosh+Haystack完成高级搜索功能，奈何并不能复现TF-IDF中对关键字重要性的评估）。

- 在模板头部使用 `{% load highlight%}`，并在需要高亮的部分使用 `{% highlight content with query %}`，其中content是期望被渲染的内容，query是搜索内容，可自动完成截取+高亮的工作。高亮实际是在关键词两侧加入了span标签，需要用CSS自定义样式，如 `<style>span.highlighted { color: red; }</style>`。