

# MLB Search Engine

计 86 周恩贤 2018011438

2019 年 9 月 8 日

## 1 程序介绍

本程序是基于 `django` 的美国职棒大联盟 (MLB) 搜索引擎,从自由时报中爬取了 20000 篇新闻作为数据库,允许使用者搜索相关新闻并查阅相关内容。同时,本程序为每个球队做了一个介绍页,并加入了球队热搜榜,统计出了每个球队的相关新闻篇数。在实现方面,前端使用了 `CSS`、`jquery` 等进行美化,而后端使用了 `jieba` 分词、`TF-IDF` 算法等统计新闻匹配程度。欲使用此程序,请确保安装了 `django(version ≥ 2.2)` 以及 `python(version ≥ 3.5.6)`,并于 `SportsNews\SportsNewsSearchEngine` 目录下输入 `python manage.py runserver` (端口号),即可使用浏览器进行相关操作。

## 2 界面介绍

### 2.1 首页

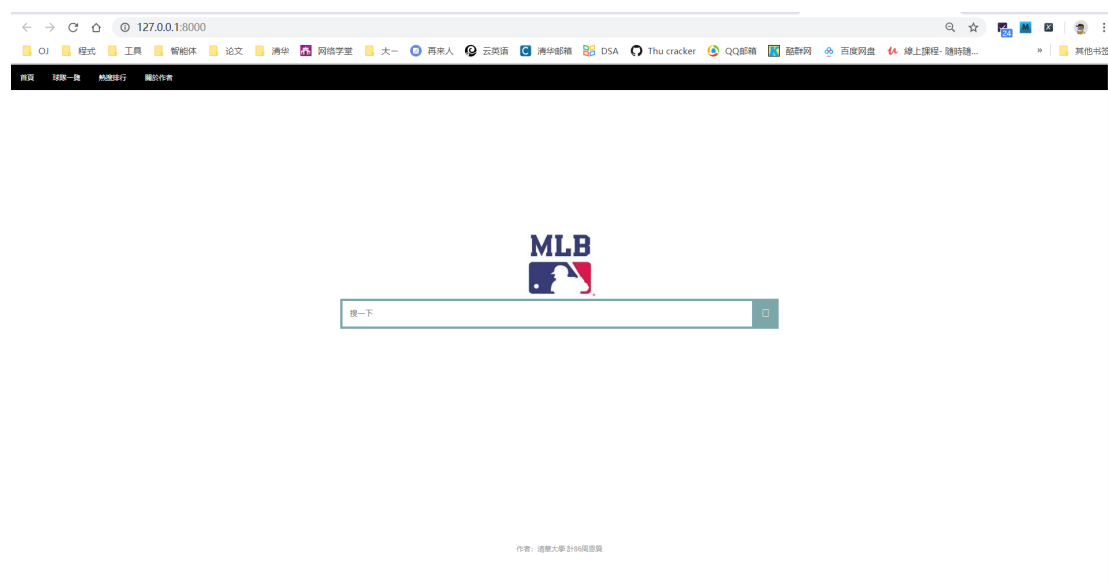


图 1: 搜索引擎的首页

图 1 是本搜索引擎的首页,利用正则表达式匹配 '\$'(除 IP 以及端口号外没有任何后缀的网址)。网页上方有工具栏,中间有个搜索框,最下方显示了作者。

2.2 工具栏



图 2: 工具栏

图 2 是浮动在页面上方的工具栏，点击" 首页 " 即可跳转到节 2.1 的页面。

2.2.1 球队一览



图 3: 球队一览表，网址为 */teamlist*

图 3 是点击工具栏" 球队一览 " 后所跳转到的页面，显示出 MLB 三十个球队。用户可点击超连接跳转至图 10 的球队介绍页面。

2.2.2 热搜排行表



图 4: 热搜排行表，网址为 */rank*

图 5 是点击工具栏" 热搜排行 " 后所跳转的页面，显示每个球队在数据库中的相关新闻数量以及排行。用户可点击超连接跳转至图 10 的球队介绍页面。

### 2.2.3 关于作者



图 5: 个人介绍页, 网址为 `/zex`

图 5 是点击工具栏"关于作者"后所跳转的页面, 这是我预习 WEB 编程所编写的个人介绍页, 之后会再继续美化维护。

## 3 功能介绍

### 3.1 搜索功能

用户可在图 1 或图 6 的搜索框输入关键词句, 并点击右侧绿色小按钮进行新闻搜索。

#### 3.1.1 搜索结果页

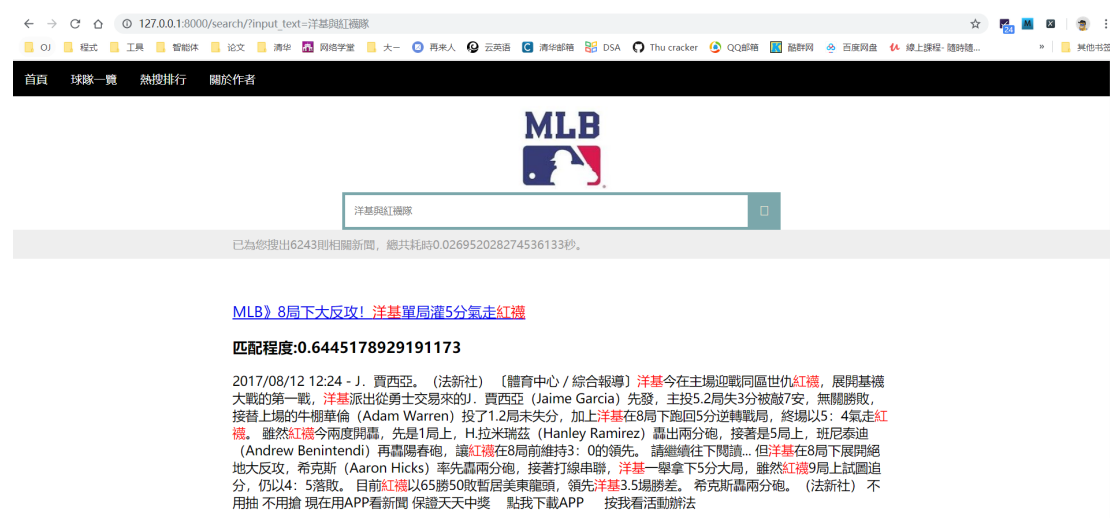


图 6: 搜索结果页, 网址为 `/search/?input_text = xxx`

图 6 为用户搜索" 洋基与红袜队 " 后点击按钮后所跳转到的搜索结果页。前端利用表单设置跳转网址, 后端利用了 GET 方法获取了用户输入的搜索词句并进行搜索。搜索出的相关新闻依序显示于页面中标题加入了超连接并高亮显示关键词汇。若没有相关新闻则会输出" 已无更多搜索结果 " 的提示。

### 3.1.2 分页展示



图 7: 页码框, 点击以跳转至不同页

图 7 为每个搜索结果页面底下的页码框, 当前页面仅显示 10 条新闻, 用户点击数字  $i$  后可跳转至 `/search/?input_text=xxx&page_num=i` 的搜索结果页, 后端并利用 GET 方法获取目标页码以展示正确页面。

### 3.1.3 多词汇搜索

本搜索引擎支持多词汇的搜索。多词汇搜索是指当用户输入多关键词或长句时, 会先利用节 4.1 的技术进行分词, 再按照节 4.3 所计算出的匹配程度排序新闻。

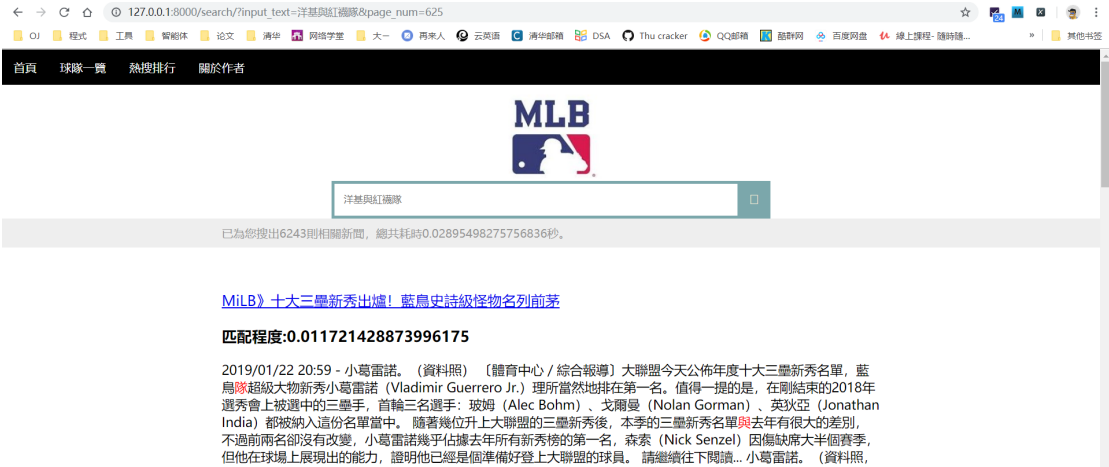


图 8: 匹配程度不高的例子

图 8 为同样搜索" 洋基与红袜队 " 的最后一页搜索结果页。相比于图 ?? , 可看出不仅在关键词的数量上两篇新闻有所落差, 且洋基、红袜等词汇相比于队这个词汇更重要, 因而一个在首页, 一个在末页。

## 3.2 新闻页面

当用户点击图 6 的新闻标题超连接后, 会跳转到该新闻页面。

## MLB》用終結者換先發？洋基可能不簽陳偉殷

日期：2015/11/19 14:00

【記者陳人齊／綜合報導】台灣左投陳偉殷轉戰紐約洋基可能曾添變數，據《每日新聞》報導指出，今年冬天洋基可能不會在自由市場上簽下任何先發投手，而是透過交易的方式補強，被點名送出的，就是最具賣相的洋基終結者米勒（Andrew Miller）。紐約媒體做這樣的猜測其實不難理解，因為去年洋基透過交易換來馬林魚潛力投手艾歐瓦迪（Nathan Eovaldi）獲得不錯的回報，艾歐瓦迪本季在洋基繳出14勝3敗，防禦率4.20的成績，站穩洋基先發輪值。請繼續往下閱讀。有了艾歐瓦迪的成功經驗，洋基今年可能不會在自由市場大採購，畢竟投手身價逐年上漲，且務實的建隊方式，已是近年來洋基的經營方針。本季投出36次救援成功、防禦率僅2.04的米勒，被點名可能會被洋基當作交易籌碼，用來換取一名輪值的先發投手。據傳已有多支球隊對米勒與洋基外野手賈納（Brett Gardner）感興趣，並與洋基展開交涉；而洋基未來守備陣一職，可能會由佈局投手聖坦西斯（Dellin Betances）接任，他本季後援表現相當理想，6勝4敗防禦率只有1.50，拿下9次救援成功。陳偉殷可能無緣加盟洋基。（資料照）洋基有可能將米勒送上談判桌。（法新社）

图 9: 新闻页面，网址为 `/news/xxx`

图 9 为新闻页面的展示。显示出了新闻标题、新闻时间、新闻内文。内文中关于球队名、球员名等关键字进行了超连结，点击后会跳转至节3.3 的球队介绍页

### 3.3 球队介绍页

当用户点击图 5 或图 9 的超连接后，会跳转相应的球队介绍页。

#### 洋基

加入時間：1901年

所在位置：紐約

紐約洋基(New York Yankees)，是一支位於紐約布朗克斯區的職業棒球隊。隸屬於美國職棒大聯盟美國聯盟東區，是美國聯盟八支創始球隊之一。與另一支屬於國家聯盟的紐約大都會是紐約主要的兩支球隊，球隊主場為新洋基體育場。在1901年球隊初創時，主場位於馬里蘭州巴爾的摩，當時名為巴爾的摩金鶯（與現在的巴爾的摩金鶯無關）。1903年球隊遷至紐約市，改名為紐約高地人（New York Highlanders）。在1913年起，球隊官方正式更名為洋基（Yankees）。紐約洋基為史坦布瑞納家族所掌控的一間有限公司洋基全球企業集團（Yankee Global Enterprises）所擁有，該家族在1973年買下球隊。布萊恩·凱許曼是球隊經理，亞倫·布恩是球隊總教練。1923年到1973年及1976年到2008年，洋基的主場為老洋基體育場。1974及1975年，因為紐約噴射機與紐約巨人的關係，洋基與大都會共用謝亞球場。2009年，因為舊球場已經關閉並拆除，洋基搬到新的同名球場，並在同年奪得世界大賽冠軍。而洋基一直是大聯盟最多觀眾的球隊之一（47次大聯盟第一），2011年球隊的總觀眾人數（3,653,680人次）、主場平均觀眾人數（45,107人次）、整年平均觀眾人數（33,228人次）皆排名大聯盟第一。身為世界上最成功的運動俱樂部之一，紐約洋基贏得19次分區冠軍和40次聯盟冠軍，還有27次世界大賽冠軍，皆為大聯盟紀錄。在北美主要之體育項目中，洋基在1999年獲得世界大賽冠軍以後，超越國家冰球聯盟之蒙特婁加拿大人隊的所獲得的24次史坦利盃冠軍，成為美國職業運動裡面拿下最多冠軍的隊伍。值在球隊的歷史裡面，有44位的棒球名人堂選手，並且共有23個球隊號碼退休，這裡面有名的球員包括貝比·魯斯、盧·賈里格、喬·狄馬喬、米奇·曼托、尤吉·貝拉和懷堤·福特等選手，而洋基也是在所有的球隊中，唯一每個守備位置皆有球員獲選登錄棒球名人堂中的球隊。為了贏得冠軍，球隊利用高額的薪資來吸引人才，特別是在史坦布瑞納時代。根據2017年美國《富比士》雜誌的報導，紐約洋基是美國第二以及世界上第二最具價值的運動球隊，估計價值約為37億美元。洋基是一個受到大家喜愛和擁有許多球迷的球隊，他們的競爭對手是波士頓紅襪，彼此間的競爭關係延續了一個多世紀，在體育競技的範疇中堪稱世仇。這種情緒和關係在美國職業體育的歷史上非常著名也異常激烈。另外，YES頻道（YES Network）在2002年成立以後，負責報導洋基的相關新聞。

图 10: 球队介绍页上半部份，网址为 `/team/xxx`

图 10 为球队介绍页的上半部份的展示，显示队名字、创立时间、地点及简介。

图 11 为球队介绍页的下半部份，显示出了该球队的球员列表以及相关新闻。

## 投手(P)

Domingo Germán

J. A. Happ

James Paxton

Masahiro Tanaka

Chance Adams

Zack Britton

Luis Cessa

Nestor Cortes Jr.

Ryan Dull

Cory Gearrin

Chad Green

Tommy Kahnle

Jonathan Loáisiga

Tyler Lyons

Adam Ottavino

Aroldis Chapman

## 相關新聞

[MLB》達比修有終於簽了！小熊砸36.6億肥約網羅](#)

2018/02/11 07:20 - 達比修有將改披小熊戰袍。(資料照, 美聯社) [體育中心 / 綜合報導] 自由市場最大咖達比修有總算簽了! 如不少外媒臆測, 達比修有約定小熊, 雙方談妥6年1.26億美元 (.....

[MLB》後生可畏！托瑞斯8月猛敲13轟 比肩三大名人堂傳奇（影音）](#)

2019/08/27 11:37 - [體育中心 / 綜合報導] 搶在「法官」賈吉 (Aaron Judge) 生涯百轟之前, 22歲托瑞斯 (Gleyber Torres) 先用全壘打寫紀錄, 今天對水手擊出8月份.....

[今日MLB戰績 陳偉殷6局6K惜敗](#)

2016/05/18 15:06 - (Marlins.com) 紅人 1: 13 印地安人 勇士 9: 12 海盜 馬林魚 1: 3 費城人 水手 10: 0 金鶯 光芒 12: 2 藍鳥 雙城 2: 7 老.....

[MLB》曾是建仔投捕搭檔 塞維里10億續留海盜](#)

2016/05/18 11:46 - [記者陳人齊 / 綜合外電報導] 海盜今天與主戰捕手塞維里 (Francisco Cervelli) 達成績約協議, 雙方簽下3年3100萬美元 (約新台幣10.1億元) .....

图 11: 球队介绍页下半部

## 4 信息检索相关技术与实现

### 4.1 分词

利用 `jieba` 库进行分词, 具体内容写于 `views.py` 中

### 4.2 倒排索引

将分词、TF-IDF 的结果加入全局变量 `InvertedIndexMap` 中, 一个从 `string`  $\rightarrow$  `set` 的 `dict`, 透过词可查询所有包含该词的新闻编号集合

### 4.3 TF-IDF 值

利用 `sklearn` 库进行 TF-IDF 的计算, 具体内容写于 `tfidf.py` 中, 结果存于 `tfidf.txt`、`tfidf1.txt` 以及 `tfidf2.txt`

## 5 爬虫相关技术与实现

利用 `scrapy` 进行新闻爬取以及球队爬取, 具体内容写于 `getNews` 目录中, 结果储存在 `data.json` 以及 `team.json`

## 6 性能统计

### 6.1 总新闻量

本次作业从自由时报 MLB 专栏中 (<https://news.ltn.com.tw/topic/MLB>) 提取, 总共提取 23035 条。为方便后续匹配以及管理, 我使用了前两万条作为新闻库。

### 6.2 查询时间

平均查询时间落在 0.01 - 0.02 秒之间。

### 6.3 精确率和召回率

依照用户输入的关键词, 我们可以将每一条新闻分为两类: 正类 (与关键词相关) 以及负类 (与关键词无关), 而搜索结果也有两种可能。因此对于每条新闻有四种情况: 相关且有搜索到 (TP)、不相关且没被搜索到 (TN)、相关但没被搜索到 (FN)、不相关但被搜索到 (FP)。而在信息检索领域有两个关键指标:

- 精确率 (precision) =  $\frac{\text{检索出的相关信息量}}{\text{检索出的信息总量}} = \frac{TP}{TP+FP}$
- 召回率 (recall) =  $\frac{\text{检索出的相关信息量}}{\text{系统中的相关信息总量}} = \frac{TP}{TP+FN}$

为了加快搜索速度, 我采取的搜索方式是: 预处理将每篇新闻标题、内容分词并计算完 TF-IDF 值, 把所有值非 0 的关键词以及其子词加入倒排索引, 因此搜索出的新闻必定相关, 在单个词语的查询情况下,  $precision \approx 100\%$ 。

但鱼与熊掌不可兼得, 此方法的缺点会遗漏掉部份因为 TF-IDF 值过小而被筛漏的新闻。以搜索 "红袜" 为例, 直接在搜索框中搜出来的结果为 3002 条, 但实际包含 "红袜" 的总新闻共 3460 条, 可说明  $recall \leq \frac{3002}{3460} = 86\%$ 。

高级搜索时, 因为长句子分词时会存在歧义, 指标会再低一些; 且每个新闻没有一个确定的指标, 比较难去判断与计算。但至少在我多次反复试验下, 单词搜索能准确且涵盖多数相关新闻, 高级搜索能按匹配程度降序排序, 已符合简易搜索引擎的特性。

## 7 心得

第三周学习的内容主要为 python 以及 django 编程, 但要完成一个大作业要学习的内容就更多了: Web 前端部份, 除了课上的基础内容 HTML 与 CSS 外, 为了实现与交互功能, 我还尝试着学习 JS 及 jquery 以完整完善用户交互功能; Web 后端部份, 除了掌握 MVT 思想并学习正则表达式匹配 URL 外, 更要完整理解 GET 方法与 POST 方法以完善搜索以及分页功能; python 的部份除了基本语法外, 更要查询并学习 scrapy、jieba、sklearn 等第三方库的用法以完善整个爬虫与信息检索的功能。因此, 这个大作业是我三周投入最多的一个, 但也是收获最多的!

这周最震撼的部份就是了解到了 python 的强大 (还有题目整整一页的斗地主考试题)。相对前两周使用 C++ 写 Qt, 这周的 python 方便、简洁、第三方库多且实用 (短短二

十行的 `scrapy` 就能完成爬虫任务!），也难怪现在 `python` 如此火热且广泛应用于各领域中。

不过很可惜因为时间缘故，我还没来得及完善整个程序，包括但不限于动态爬虫功能、利用 `bootstrap` 美化页面、学习信息检索的机器学习算法..... 但这周的学习(爆肝)让我掌握了基本 `Web Programming` 以及信息检索的基础，相信到未来我能继续精进！

奋战一星期，画个 `DMFB`；奋战一星期，写个国际象棋；奋战一星期，搭个搜索引擎。在三周的奋斗下小学期终于结束了！小学期结束了，但我不会忘记这三周以来坚持刻苦的学习过程！~~（毕竟马上开学了）~~ 再次感谢助教与老师这几周来的指导 >0<