

信息检索大作业

——新闻搜索引擎

| 小组成员 | | |
|------|----|----|
| 单位 | 姓名 | 学号 |
| ** | *然 | |
| | *舒 | |
| | *路 | |
| | *军 | |
| 组长 | 江* | |

1 系统功能需求

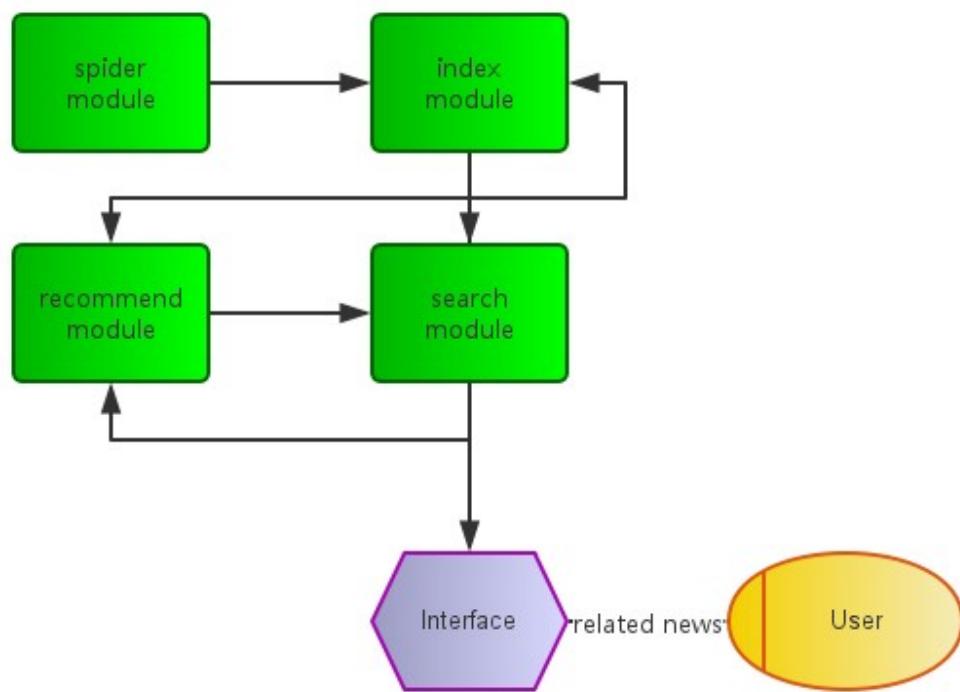
1.1 需求描述

新闻检索：爬虫定向采集 3-4 个网页，实现网页信息的抽取、检索和索引。网页个数不少于 10 个，能按时间、相关度、热度等属性进行排序，并实现相似主题的自动聚类。可以实现：有相关搜索推荐、snippet 生成、结果预览(鼠标移到相关结果，能预览)功能。

2 系统设计方案

2.1 系统总体结构框图

本系统总的实现思路如下所示：

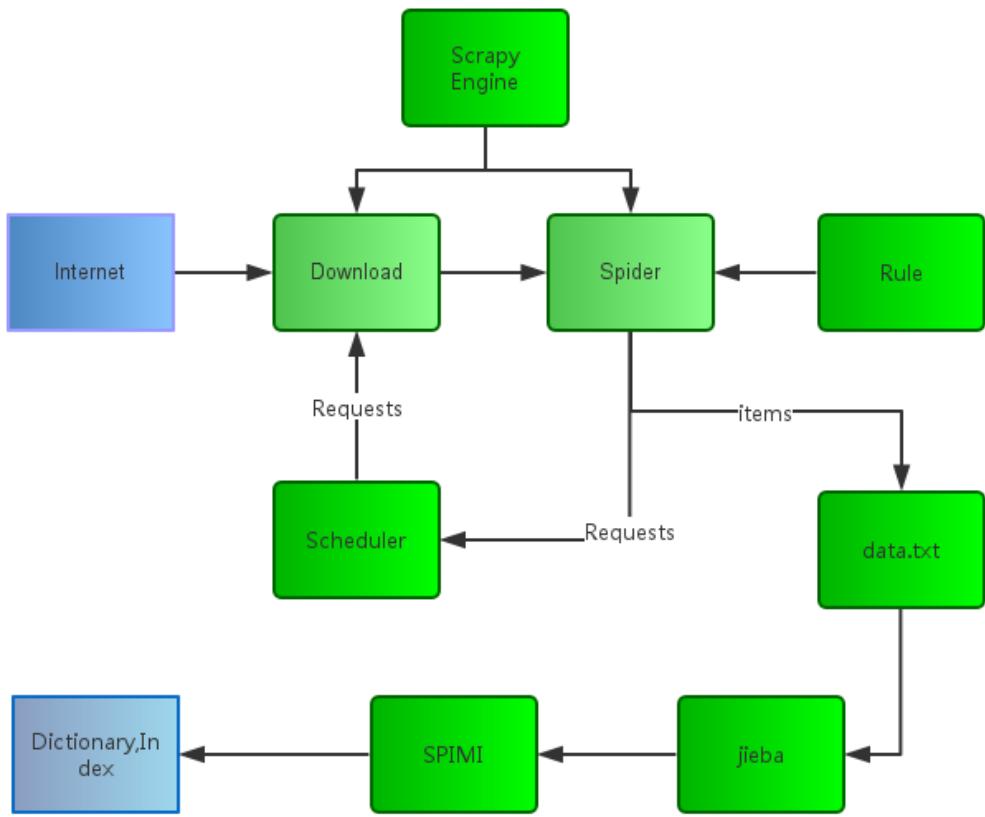


一个完整的搜索系统主要的步骤是：

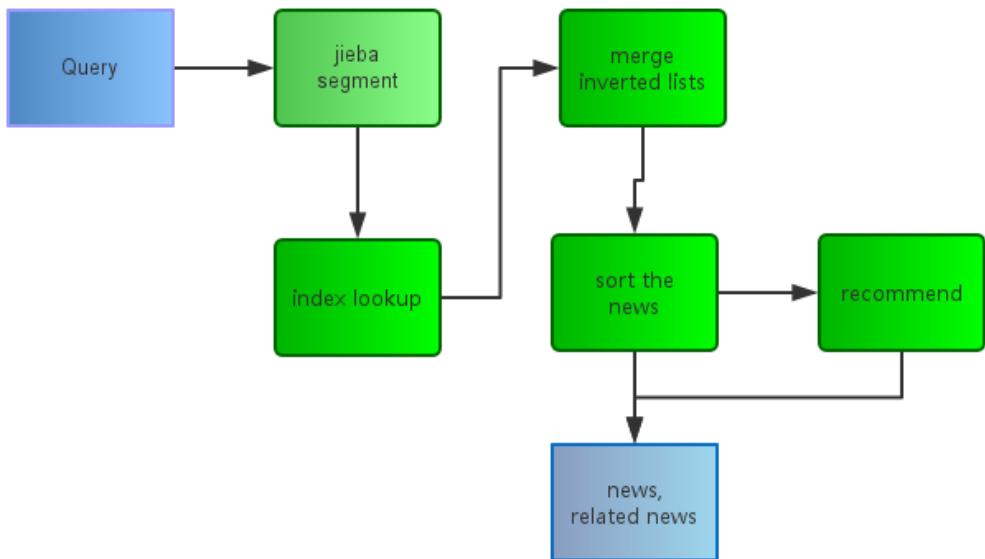
- 1、 对新闻网页进行爬虫得到语料库；
- 2、 根据所得到的语料库对文档进行分词，此步骤利用的是 `jieba`；
- 3、 对语料库进行词典和倒排索引的建立，此步骤使用的是 `SPIMI` 算法；
- 4、 用户输入查询，得到相关文档返回给用户。

以下是最关键的两个步骤的流程图。

第一步—建立索引的流程图：



第二步一对查询语句进行搜索：



2.2 网络爬虫&数据存储

2.2.1 实现方式

基于 Python，利用爬虫框架 Scrapy 爬取网页，按格式存储为文本文件。

2.2.2 Scrapy 简介

Scrapy 中采用 Xpath 从 Html 网页中提取数据。

XPath 即为 XML 路径语言，它是一种用来确定 XML（标准通用标记语言的子集）文档中某部分位置的语言。XPath 基于 XML 的树状结构，提供在数据结构树中找寻节点的能力。

常用的路径表达式包括：

| 表达式 | 描述 |
|----------|-------------------------------|
| nodename | 选取此节点的所有子节点。 |
| / | 从根节点选取。 |
| // | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置。 |
| . | 选取当前节点。 |
| .. | 选取当前节点的父节点。 |
| @ | 选取属性。 |

如获取网页中的“keywords”： Xpath 为 //meta[@name = "keywords"]/@content'

Xpath 的获取：采用火狐浏览器中的 FireBug 插件。



2.2.3 数据存储

从 Scrapy 数据流中利用 xpath 规则提取需要的数据, 提取规则如下:

新闻编号: `web_id`

新闻主题: `title=response.xpath('//head/title/text()').extract()`

新闻内容: `content=response.xpath('//div/p/text()').extract()`

新闻地址: `link=response.url`

按照 `web_id##### title##### content##### link#####` 格式以 utf-8 编码存储, 考虑到 windows 系统下, 换行符'\n'占用两个字节, linux 系统下占用一个字节, 为保证系统兼容性, 这里牺牲文档阅读便利性, 相邻文档间无换行。

磁盘读写是一种耗时操作, 为提高系统速度, 读写采用内存缓存机制, 将 Scrapy 数据流中数据存储至内存, 数据量达到设定的阈值后, 进行一次写内

```
#####
# 网易新闻 #####
# 全球各国央行何以都在搞负利率, 其实是本国政府都靠印钞为生。两岸都这么说, 此番军演的“年度例行”性质也就完全坐实。
# 对于这些网络里检举, 相信举报者也是冒着很大的风险。随着中国海上力量的增长, 会削弱美国在南海的统治性地位。韩国直播圈爆出卖淫丑闻, 美女主播利用网>
# 络直播进行性交易。设“暗门”表面上是追求享乐的表现, 背后暗藏的却是权力黑洞。去年10月蔡英文访问日本时, 专门参观了日产制作所车间。当前中国的这个通胀大部分的构成要素是制造业的产品。扫描二维码下载网易新闻客户端####http://news.163.com/####发现者专访####最新专题专访伦敦马拉松主办方: >
# 赛事如何不扰民艾滋病人隐瞒病情致医生感染可避免: 实现普遍防护 艾滋病已由一种绝症逐渐变成可控慢性病, 关键要积飞机高空飞行遇鸟击非常罕见, 若真发生>
# , 单引擎往期回顾TED演讲风靡全球, 它成功的经验是: 好的内容加好的传播。飞机备降, 是飞行过程中时常会遇到的情况。当飞行目的地的天气状况不有部分像
# “小龙虾”这样的外来入侵物种如今却是风靡全国的美食。有人近来日本现行国歌《君之代》被热议。实际上, 它的形成经历了漫长的过度抑郁症是一种需要专业治疗>
# 的大脑疾病。但市场上出现的五花八门抑郁症不同于人们通常的理解的情绪波动, 它是一种大脑疾病, 是社会。随着现在医患关系的紧张, 医院很多操作都>
# 要征求患者同意。有担当的直升机是否能在灾区投受到一系列因素的影响, 仅从客观条件上看, 气最近在非洲再次爆发的埃博拉病毒病是世界上最凶猛疾病之一>
# 病死率高7月21日, 习近平在委内瑞拉总统马杜罗陪同下抵达国家公署。委内瑞拉近日, 有专家提出, 通过修建城市风道, 将郊外的风引进主城区, 将需要最近科技界热议一个事件: 俄罗斯计算机软件首次成功通过图灵测试, 日本在垃圾处理方面有着世界一流的技术和丰富的经验。垃圾处理在日本世界卫生组织接受>
# 网易专访表示, 疫苗是可使用的最安全的医用产品, 但没船长在一艘船上拥有最高权力和责任, 发生海难后, 船长必须根据当时的目前马航失联客机黑匣子电量或已>
# 耗光, 无法发出信号。搜救团队只能后####http://news.163.com/special/interviewdiscoverer1/####统计: 71年来共133架飞机消失 1837人下落不明_网易
# 早报。> 中国驻纽约总领馆公使馆及有关方面负责人就“萨德”风波向有关方面表示严正交涉, 要求有关方面立即停止部署“萨德”反导系统的有关部署工
# 作, 重申中国对有关问题的立场, 表示将密切关注事态发展, 为维护地区和平稳定作出贡献。>
```

存数据至磁盘操作。

2.2.4 数据抓取关键问题

2.2.4.1 链接自动跟进

利用 Scrapy 爬取网页需要指定网页的链接, 本系统通过提取新闻网页中存在的链接并自动跟进:

```
links=response.xpath('//a/@href).extract()
```

注意到网页中包含的链接中有如广告等链接，如果爬虫跟进全部链接，会造成最终的新闻数据质量差，爬取时间长问题。采用域名过滤机制可以有效避免

```
if re.match(self.domains,url):#判断链接前缀是否为新闻域名  
    #跟进链接  
else:  
    pass
```

self.domains 为自定义的新闻网站域名，以网易新闻为例，域名为：

```
self.domains = http://news.163.com/
```

对符合域名前缀的链接，需要进行重复性检测，若爬虫已经跟进过此项链接，则应舍弃此项链接。本系统通过将爬取过的网页链接以字典方式存储，判断词典 key 中是否有网页链接的方式进行重复性检测。

爬取运行结束后，保存词典 key 至文件，再次运行爬虫时，通过重新加载词典文件，判断新闻网站中哪些新闻是新增，更新存储数据。

2.2.4.2 遵循礼貌性

采集器必须做到不要高频率采集某个网站，遵循礼貌性原则，系统采用请求速度根据网页响应速度动态调整。网页延迟较大时，表明采集的新闻网站负载较大，应降低爬虫的请求速度。

2.3 分词器

2.3.1 实现方式

采用结巴中文分词工具 `jieba`，其支持三种分词模式：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。本系统采用搜索引

擎模式。Jieba 分词同时支持繁体分词和支持自定义词典，使用 jieba 分词只需要遵循 MIT 授权协议即可。

2.3.2 分词结果示例

分词示例：# 全模式

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/".join(seg_list))
```

输出：我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

分词示例：# 精确模式

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/".join(seg_list))
```

输出：我/ 来到/ 北京/ 清华大学

分词示例：# 搜索引擎模式

```
seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")
print("".join(seg_list))
```

输出：小明, 硕士, 毕业, 于, 中国, 科学, 学院, 科学院, 中国科学院, 计算, 计算所, 后, 在, 日本, 京都, 大学, 日本京都大学, 深造

2.4 索引模块

2.4.1 SPIMI 构建倒排索引

倒排表维护了词项和该词项所出现的文档集合之间的一一映射关系，也就是说该词出现哪些文档中，以及文档集频率 $df(idf)$ 均由倒排表来进行维护。表结构用词典进行维护，文档集合用数组进行维护。

倒排索引建立采用 SPIMI 算法，其伪代码为下：

```
SPIMI-Invert(Token_stream)
output.file=NEWFILE()
dictionary = NEWHASH()
```

```

while (free memory available)
    do token <- next(token_stream) //逐一处理每个词项-文档 ID 对
        if term(token) != dictionary
            /*如果词项是第一次出现，那么加入 hash 词典，同时，建立一个新的倒
            排索引表*/
            then postings_list = AddToDictionary(dictionary,term(token))
            /*如果不是第一次出现，那么直接返回其倒排记录表，在下面添加其后*/
            else postings_list = GetPostingList(dictionary,term(token))
        if full(postings_list)
            then postings_list = DoublePostingList(dictionary,term(token))
        /*SPIMI 与 BSBI 的区别就在于此，前者直接在倒排记录表中增加此项新纪录
        */
        AddToPostingsList (postings_list,docID(token))
        sorted_terms <- SortTerms(dictionary)
        WriteBlockToDisk(sorted_terms,dictionary,output_file)
        return output_file

```

本系统封装 SPIMI 类，具有词典属性(用于存储词项)和缓存块属性(用于存储词性对应的文档 id 记录)，支持 push 操作(用于构建倒排记录)。Push 操作实现 在倒排记录表中增加词项、文档号对。

系统构建全部文档的倒排索引流程如下：

- (1) 按块读取磁盘中的网页文件至内存块，若文件读完，退出
- (2) 按格式解析内存中数据，返回一项网页记录，若内存数据解析完，
调至步骤 1
- (3) 对记录分词并去除标点符号、数字和单字母，调用 SPIMI 类 push 操
作。
- (4) 跳至步骤 2

2.4.2 合并倒排索引

由于爬取数据量大，无法一次性在内存中构建全部文档的倒排索引，SPIMI 算法完成后需要合并生成的多个倒排记录。

合并时，同时打开所有块对应的文件，内存中维护了为每个块准备的读缓冲区和一个为最终合并索引准备的写缓冲区。每次迭代中，利用优先级队列（如堆结构）选择最小的未处理的词项 ID 进行处理。分块索引，分块排序，最终全部合并。

2.4.3 索引本地化

由于爬取的数据量较大，因此每一次建立索引都需要较长的时间。在本系统中对于 220MB 的数据（10 万个网页），建立索引所需时间较长。因此将索引存储到本地磁盘，并在下一次系统启动时读入以提高系统运行效率是非常必要的。本系统中主要对倒排记录表建立了本地索引。

索引本地化包含两种操作：索引的存储和索引的导入。

2.4.3.1 索引存储

在每一次更新索引后，按照“词项:文档号#tf:文档号#tf|”格式，系统将其保存在本地磁盘中。

以下为倒排记录表索引文件，其保存了倒排记录表中各个词项所在文档的信息：

所有词项在倒排记录表中字节位置的集合，按照词项的顺序进行存储，用于快速读取某项词项的倒排记录表信息，格式为“词项:文档集频率:在倒排记录字节位置:在倒排记录字节占用字节数|”。

与民同乐:6:14690928:47|与江:4:14690975:31|与沃:1:14691014:7|与治超:2:14691032:15|与泛:1:14691055:7|与浩英:1:14691091:7|与湛:1:14691106:7|与狼共舞:5:14691127:39|与现:2:14691174:15|与琶洲:3:14691260:23|与生俱来:4:2:14691237:32|与示:2:14691577:15|与秀秀:1:14691663:7|与种:1:14691618:7|与程:5:14691633:39|与立:1:14691680:7|与秉:1:14691695:7|与纪:2:14691710:15|与续:1:14691733:7|与翰:1:14691748:7|与艳翠:1:14691766:7|与芳芳:1:14691784:7|与茂桥:1:14691802:7|与茨尔文:1:14691823:7|与莉碧露:1:14691844:7|与蕃类:1:14691862:6|与莹:1:14691876:7|与萍萍:2:14691894:15|与虎谋皮:4:14691923:31|与行:1:14691962:7|与赛:2:14691977:15|与速:2:14692000:15|与那国岛:3:14692029:22|与量:1:14692059:7|与锐:1:14692074:7|与陆:1:14692089:6|与院:1:14692103:7|与雀鲷:1:14692121:7|与霆锋:2:14692139:15|与非:94:14692162:747|与腾父:1:14692190:7|与鹤北:1:14692193:7|与丐:1:14692195:7|与兀儿:1:14692196:7|与丐头:3:14692198:23|与妇:1:14692199:23|与丐有:1:14692208:7|与丑:74:14692208:588|丑丑:3:14692209:23|丑中:1:14692209:6|与院:1:14692210:7|与丑事:43:14692211:127|丑妇:1:14692212:127|丑八怪:3:14692213:15|丑剧:10:14692214:23|丑化:45:14692215:357|丑变:1:14692216:357|丑怪:1:14692217:1|丑名:1:14692218:7|丑女:16:14692219:127|丑妇:1:14692220:127|丑小鸭:23:14692221:183|丑态:16:14692222:127|丑态百出:6:14692223:47|丑眉:1:14692224:7|丑眉:1:14692225:7|丑嘴脸:3:14692226:23|丑时:1:14692227:7|丑样:1:14692228:7|丑死:3:14692229:23|丑男:3:14692230:23|丑相:2:14692231:23|丑眉:1:14692232:7|丑眼:881:14692233:6997|丑陋:64:14704036:504|专:235:14704539:1858|专一:22:14706405:175|专一性:2:14706591:15|专不精:1:14706617:77|专业:5011:14706632:15|专业课:1:14706642:7|专业书:13:14706643:63|专业书籍:8:14746513:63|专业人士:310:14746590:2466|专业人才:124:14746970:987|专业分工:10:14750071:78|专业化:291:14750160:2295|专业厂:1:14752466:7|专业名词:9:14752487:70|专业型:11:14752568:85|专业培训:72:14752667:570|专业学位:19:14753251:148|专业对口:16:14753413:126|专业展:4:14753558:30|专业性:145:14753591:1145|专业户:64:14754747:508|专业技能:77:14755269:613|专业本科:9:14755896:70|专业术语:16:14755986:78|专业村:10:14756069:781|专业版:4:14756158:30|专业班:2:14756199:15|专业界:13:14756225:103|专业知识:209:14756342:1656|专业科目:11:14758812:86|专业级:2:14758189:15|专业组:7:14758135:55|专业翻译:4:14758204:31|专业英语:6:14758249:46|专业训练:50:14758309:397|专业课:79:14758717:627|专业课程:35:14759358:277|专业部:1:14759646:7|专业银行:4:14759667:30|专业队:44:14759708:351|专业队伍:26:14760073:287|专:85:14760288:673|专习:1:14760969:71|专书:3:14760984:22|专深:1:14761014:71|专重:28:14761029:223|专产:2:14761029:15|专革:10:14761029:69|专工:501:14761360:39691|专川:1:14761360:23

2.4.3.2 索引导入

索引导入根据上述中索引存储的格式进行，由于索引是按照数据本身结构进行存储的，因此在导入的时候，根据读入的顺序构建其相应的数据结构。

系统运行时，先以字典格式读取索引文件至内存，其中词项为字典的 key 值，文档集频率、词项字节位置和占用字节数为字典 value。需要读取某个词项倒排记录表时先从字典中查找词项在文件中字节位置，判断内存中是否有对应的倒排记录表，若有，返回。否则，通过文件指针偏移方式读一块磁盘倒排索引文件，更新内存块。

2.5 查询&相关新闻推荐模块

查询即根据用户输入的关键字，返回其相关的新闻，相关新闻推荐则是根据用户查看的网页，推荐与此网页相似的新闻。

本系统中查询和相关新闻推荐模块均采用了余弦相似度算法，其伪代码如下所示：

```

COSINESCORE( $q$ )
1 float Scores[ $N$ ] = 0
2 float Length[ $N$ ]
3 for each query term  $t$ 
4 do calculate  $w_{t,q}$  and fetch postings list for  $t$ 
5 for each pair( $d, tf_{t,d}$ ) in postings list
6 do  $Scores[d] += w_{t,d} \times w_{t,q}$ 
7 Read the array Length
8 for each  $d$ 
9 do  $Scores[d] = Scores[d] / Length[d]$ 
10 return Top  $K$  components of Scores[]

```

查询词项一般较短，系统可以快速返回相关文档，但文档一般较长，所含词项较多，计算文档间相似度速度慢，本系统采用了一些加速技巧。

(1) 索引去除(Index elimination): 一般检索方法中，通常只考虑至少包含一个查询词项的文档。可以进一步拓展这种思路，只考虑那些包含高 *idf* 查询词项的文档，只考虑那些包含多个查询词项的文档(比如达到一定比例，3个词项至少出现2个，4个中至少出现3个等等)。

(2) 采用了最大值堆和最小值堆的优先队列的方法来提高查询的效率。

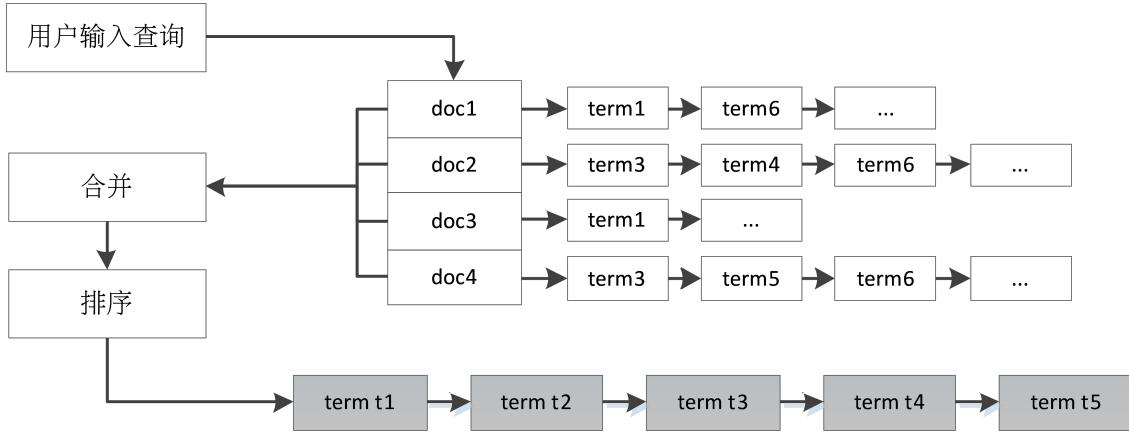
这部分我们主要构建了四个类：

| | |
|------------------------|---------------|
| class max_queue: | 最大值堆实现的优先队列 |
| class min_queue: | 最小值堆实现的优先队列 |
| class CosineScore: | 余弦相似度计算 |
| class FastCosineScore: | 加速优化后的余弦相似度计算 |

本系统查询具体实现实现如下：

- (1) 对用户输入进行分词
- (2) 去除停用词，采用 *tf-idf* 算法计算所有文档得分
- (3) 返回查询结果

其原理及过程图如下所示：



2.6 相关主题推荐模块

本系统实现了对查询词的相关词汇推荐。相关主题推荐模块采用了 word2Vec 算法，word2Vec 是 Google 在 2013 年年中开源的一款将词表征为实数值向量的高效工具，采用的模型有 CBOW（Continuous Bag-Of-Words，即连续的词袋模型）和 Skip-Gram 两种，是一个 Deep Learning（深度学习）的模型。word2vec 通过训练，可以把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。word2Vec 非常高效，一个优化的单机版本一天可训练上千万词。

训练语料采用搜狗实验室语料集，语料大小共 3.1GB。对原始语料进行格式解析，并去除相应的“<content>”、“</content>”标签后进行分词，作为 word2Vec 的训练数据。中间文件格式分别如下所示：

```

<content>1分为最差，5分为最好。您的评分对别人的钱包至关重要，请如实评分。清爽度：1分 2分 3分 4分 5分吸收度：1分 2分 3分
4分 5分效果：1分 2分 3分 4分 5分性价比：1分 2分 3分 4分 5分推荐度：1分 2分 3分 4分 5分</content>
<content></content>
<content></content>
<content></content>
<content></content>
<content></content>
<content>s l i d e c h a n g e s e v e r y 1 0 s e c o n d s 没有姚麦火箭也能胜掘金 5 7 3 5 5 4 9 6 _8 . j p g</content>
<content>组图：西南赛区次轮比赛精彩瞬间幻灯片每 1 0 秒变换一次 2 1 . j p g 张烨 艾轩 摄</content>
<content></content>
<content>2 0 0 8 第七届青岛国际车展于 1 5 日在青岛国际会展中心盛大开幕。本次车展将持续到本月 1 9 日。今年青岛国际车展是历年岛城车展规模最大的一次，使用了青岛国际会展中心的全部室内外展馆。以下为现场模特图片。</content>
<content>开奖结果 1 0 0 4 0 8 0 1 3 3 3 [详细] 一等奖 5 注 5 0 0 万 二等奖 3 2 注 8 5 9 2 9 8 元 三等奖 4 2 6 注 3 0 0 0 元 四等奖 2 2 0 0 1 注 2 0 0 元 五等奖 4 5 8 6 0 5 注 1 0 0 元 六等奖 6 6 3 2 6 9 8 注 5 元</content>
<content>开奖结果 8 6 9 5 6 5 0 [详细] 特等奖 0 注 0 元 一等奖 1 4 注 5 6 6 7 6 元</content>
<content>建邦华府全景效果图</content>
<content>作为美国的著名体育杂志，《体育画报》泳装秀网罗了一大批世界著名体育明星及模特，图为网球巨星格拉芙，拍摄时间 1 9 9 7 年</content>
<content>速度：（说明：点击自动播放）说明：点击该按钮，选择一论坛即可</content>
<content></content>
<content>图为中华骏捷 F R V 钥匙。</content>
<content>葡萄牙男子骑车支持奥运万里行发车仪式 精彩图集</content>
<content>/ 1 张网易公司版权所有 (C) 1 9 9 7 - 2 0 0 8 标题</content>
<content></content>
<content>爱 . 喜欢 . . . 高兴 . . . 兴奋 . . . 幸福 . . . 期待 . . . 想念 . . . 生气 . . . 愤怒 . . . 恨 . . . 讨厌 . . . 嫉妒 . . .
. 羡慕 . . . 紧张 . . . 悲伤 . . . 难过 . . . 忧郁 . . . 烦恼 . . . 害怕 . . . 担心 . . . 压力 . . . 害羞 . . . 快乐 . . . 大笑 . . . 微笑 . . .
. 笑容 . . . 加载中 . . . 上一篇：下一篇：评论 loading . . . 加载中 . . . 最近更新最受欢迎新闻榜加载中 . . . 加载中 . . . 加载中 . . . 加载中 . . . 精彩推荐加载中 . . . 图片说话加载中 . . . 推荐新闻加载中 . . . 加载中 . . . 恒指现报 2 5 2 9 6 . 1 7 点，跌 4 4 6 . 6 0 点，成交金额 1 6 7 . 5 1 亿元；认沽证成交金额为 3 7 . 6 7 亿元，占大市成交 2 2 . 4 9 %。认购证总成交额为 3 1 . 7 1 亿元，占认股证成交的 8 4 . 1 7 %；认沽证总成交额为 5 . 9 7 亿元，占认股证成交的 1 5 . 8 3 %。***以上资讯与实际发放时间延迟十五分钟，并供用户参考。网页</content>
<content>新浪提示：本文属于研究报告栏目，仅为分析人士对一只股票的个人观点和看法，并非正式的新闻报道，新浪不保证其真实性和客观性，一切有关该股的有效信息，以沪深交易所的公告为准，敬请投资者注意风险。报告作者：张晗 撰写日期：2 0 0 8 - 0 5 - 0 5 预计国电电力权证的中枢价值为 3 . 9 0 2 元，对应波动率为 1 0 0 %，溢价率为 5 4 . 0 9 %，相应可分离债的价值中枢约在 1 1 . 6 3 4 元，区间估计为 1 1 . 2 9 ~ 1 1 . 9 6 元。这表明，当股价下跌超过 1 1 . 6 1 % = 1 6 . 3 4 % ÷ (股票成本 7 . 4 % ÷ 配售价值 0 . 7 3 )，持股配售者短期会出现亏损。而网下直接申购可获得 5 倍的杠杆 (2 0 % 的申购定金)，由于新股密集发行，会分流申购资金，所以我们预计中签率可能为 0 . 6 % 左右，则对应网下申购收益率为 0 . 4 9 %，年化收益率 1 2 . 6 1 %。但以上上市价假设基于正股现价，若在可分离转债上市时正股价格高于现价，则收益率有望提高，建议投资者积极申购。新浪声明：本版文章内容纯属作者个人观点，仅供投资者参考，并不构成投资建议。投资者据此操作，风险自担。
华龙网 - 重庆商报 本报讯 (记者 廖宇翔) 在昨天凌晨进行的西甲联赛第 3 7 轮的比赛中，主场作战的巴萨在两球领先的情况下惨遭大逆转，2 : 3 输给了马洛卡。由于这是巴萨本赛季最后一个主场比赛，西班牙媒体也称此战是里杰卡尔德的“告别赛”，“黑天鹅”将在赛季结束后下课。不过，巴萨球员并未用胜利为里杰卡尔德送行，反而惨遭逆转。而之前有消息称小罗将在本场比赛最后一次代表巴萨亮相，但最终他未能入选参赛名单。在昨天这场比赛中，亨利和埃托奥为巴萨先进两球，马洛卡随后连追三球。网页不支持 Flash，我们来到拉萨已经是第四天，再这几天里，我们的所看所闻所感都让我们深刻感受到了一个真实宁静的拉萨。官方网站讯 我们来到拉萨已经是第四天，再这几天里，我们的所看所闻所感都让我们深刻感受到了一个真实宁静的拉萨。奥运圣火即将抵达拉萨，无论是学校里的小学生还是街上偶遇的拉萨居民，每个人心中都期待着奥运圣火早点到来。沿街采访我们遇见了一位拉萨教师，他情绪高涨的表达了自己的和家人渴望奥运圣火来到拉萨的愿望。他说，这是整个中华民族的荣耀，作为拉萨市民的一份子，他会做好一切迎接准备，等到圣火到来的那一天，他和家人一定会上街为奥运圣火传递加油，为北京奥运加油。送走这位教师，我们打车前往西藏大学，司机大哥是本地居民，得知我们是搜狐奥运圣火传递报道团的记者很热情的

```

原始语料

新浪提示：本文属于研究报告栏目，仅为分析人士对一只股票的个人观点和看法，并非正式的新闻报道，新浪不保证其真实性和客观性，一切有关该股的有效信息，以沪深交易所的公告为准，敬请投资者注意风险。报告作者：张晗 撰写日期：2 0 0 8 - 0 5 - 0 5 预计国电电力权证的中枢价值为 3 . 9 0 2 元，对应波动率为 1 0 0 %，溢价率为 5 4 . 0 9 %，相应可分离债的价值中枢约在 1 1 . 6 3 4 元，区间估计为 1 1 . 2 9 ~ 1 1 . 9 6 元。这表明，当股价下跌超过 1 1 . 6 1 % = 1 6 . 3 4 % ÷ (股票成本 7 . 4 % ÷ 配售价值 0 . 7 3)，持股配售者短期会出现亏损。而网下直接申购可获得 5 倍的杠杆 (2 0 % 的申购定金)，由于新股密集发行，会分流申购资金，所以我们预计中签率可能为 0 . 6 % 左右，则对应网下申购收益率为 0 . 4 9 %，年化收益率 1 2 . 6 1 %。但以上上市价假设基于正股现价，若在可分离转债上市时正股价格高于现价，则收益率有望提高，建议投资者积极申购。新浪声明：本版文章内容纯属作者个人观点，仅供投资者参考，并不构成投资建议。投资者据此操作，风险自担。

华龙网 - 重庆商报 本报讯 (记者 廖宇翔) 在昨天凌晨进行的西甲联赛第 3 7 轮的比赛中，主场作战的巴萨在两球领先的情况下惨遭大逆转，2 : 3 输给了马洛卡。由于这是巴萨本赛季最后一个主场比赛，西班牙媒体也称此战是里杰卡尔德的“告别赛”，“黑天鹅”将在赛季结束后下课。不过，巴萨球员并未用胜利为里杰卡尔德送行，反而惨遭逆转。而之前有消息称小罗将在本场比赛最后一次代表巴萨亮相，但最终他未能入选参赛名单。在昨天这场比赛中，亨利和埃托奥为巴萨先进两球，马洛卡随后连追三球。网页不支持 Flash，我们来到拉萨已经是第四天，再这几天里，我们的所看所闻所感都让我们深刻感受到了一个真实宁静的拉萨。官方网站讯 我们来到拉萨已经是第四天，再这几天里，我们的所看所闻所感都让我们深刻感受到了一个真实宁静的拉萨。奥运圣火即将抵达拉萨，无论是学校里的小学生还是街上偶遇的拉萨居民，每个人心中都期待着奥运圣火早点到来。沿街采访我们遇见了一位拉萨教师，他情绪高涨的表达了自己的和家人渴望奥运圣火来到拉萨的愿望。他说，这是整个中华民族的荣耀，作为拉萨市民的一份子，他会做好一切迎接准备，等到圣火到来的那一天，他和家人一定会上街为奥运圣火传递加油，为北京奥运加油。送走这位教师，我们打车前往西藏大学，司机大哥是本地居民，得知我们是搜狐奥运圣火传递报道团的记者很热情的

训练语料

训练后，分别用人名、名词、形容词、地点等做测试，结果分别如下：

| Enter word or sentence (EXIT to break): 中国 | |
|--|-----------------|
| Word | Cosine distance |
| 日本 | 0.566599 |
| 亚洲 | 0.539846 |
| 我国 | 0.537607 |
| 世界 | 0.515456 |
| 亚洲地区 | 0.513031 |
| 大国 | 0.502661 |
| 印度 | 0.496540 |
| 越南 | 0.494813 |
| 本国 | 0.494067 |
| 海外 | 0.475862 |
| 大陆 | 0.468649 |
| 外国 | 0.467237 |
| 西方 | 0.462890 |
| 欧美 | 0.457420 |
| 全世界 | 0.451915 |
| 中国政府 | 0.450379 |

Enter word or sentence (EXIT to break): 漂亮

Word: 漂亮 Position in vocabulary: 2041

| Word | Cosine distance |
|------|-----------------|
| 很漂亮 | 0.691262 |
| 好看 | 0.657226 |
| 可爱 | 0.614957 |
| 赏心悦目 | 0.614937 |
| 潇洒 | 0.612026 |
| 聪明 | 0.599426 |
| 绝妙 | 0.586788 |
| 华丽 | 0.585658 |
| 美妙 | 0.578707 |
| 秀气 | 0.574636 |
| 恰到好处 | 0.572583 |
| 轻巧 | 0.563818 |
| 精彩绝伦 | 0.558048 |
| 美丽 | 0.551078 |
| 灵巧 | 0.545716 |
| 惹眼 | 0.544055 |
| 网急 | 0.542498 |
| 养眼 | 0.539689 |
| 讨喜 | 0.538323 |
| 精妙 | 0.537025 |

Enter word or sentence (EXIT to break): 停止

Word: 停止 Position in vocabulary: 2392

| Word | Cosine distance |
|------|-----------------|
| 关闭 | 0.596192 |
| 中止 | 0.590682 |
| 取消 | 0.550663 |
| 禁令 | 0.521028 |
| 停掉 | 0.518421 |
| 停止使用 | 0.517587 |
| 恢复正常 | 0.505522 |
| 安迪纳矿 | 0.489993 |
| 无限期 | 0.477630 |
| 被迫 | 0.473339 |
| 采取措施 | 0.472726 |
| 暂停 | 0.472556 |
| 与斯通 | 0.470605 |
| 停业 | 0.469379 |
| 开始 | 0.469252 |
| 停止下来 | 0.466472 |
| 过唱多 | 0.456932 |
| 采取行动 | 0.451249 |
| 在此期间 | 0.449231 |
| 责令 | 0.448290 |
| 暂时中止 | 0.447781 |
| 中断 | 0.446916 |
| 放弃 | 0.446279 |
| 关门 | 0.443222 |
| 撤销 | 0.443035 |
| 宣布 | 0.442200 |
| 强制 | 0.441401 |
| 停歇 | 0.439220 |
| 停工 | 0.439096 |

2.7 前端模块

由于本项目前端功能简单，为了提高前后端交互效率及开发效率，在前端选型方面采用轻量级前端框架 webpy，能够快速高效的集成到现有项目中。

Webpy 框架由以下几部分构成：

- 1) web 服务，这部分可以自动创建，指定服务端口即可
- 2) Url 解析，这是主要部分，需要编写每一类的 URL 模式以及对应模式处理的类。
- 3) 前端渲染引擎，由每个类处理后会返回对应的内容，但是需要渲染后才能在首页显示。
- 4) 前端模板，需要自定义编写不同的前端页面的样式结构，渲染引擎会根据不同类名来选择模板，并把结构填充到模板中，返回给浏览器。

2.7.1 基本结构

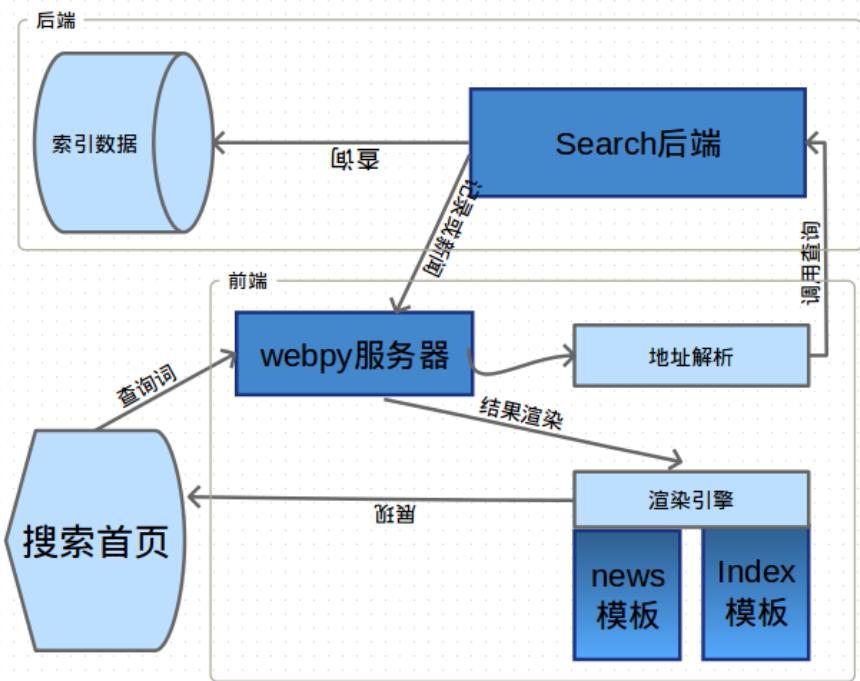
前端部分位于本项目中 PROJECT_ROOT/web 文件夹下.

主体结构包含：

- 1) Main.py，为整体项目入口，会启动搜索引擎后端及前端。
- 2) Templates/index.html 为搜索引擎首页样式模板
- 3) Templates/news.html 具体新闻页样式模板

2.7.2 前后端架构

整体交互架构如图：



- 1) 浏览器输入服务器地址, 如 `http://0.0.0.0:1111/`
- 2) 搜索首页默认返回一个无新闻列表的首页模板, 提供查询功能
- 3) 搜索框输入查询词, 如果为查询句子, 则会由 `jieba` 分词分割成不同的查询次, 默认按照布尔检索模式, 最后返回‘与操作’的新闻列表
- 4) 输入搜索词点击查询, 会跳转到如下格式的地址: `http://0.0.0.0:1111/?searchword=中国`, 该 URL 会传递给 web 服务器, 然后 web 服务器对 url 进行解析, 该地址会调用查询新闻列表的类, 并最终调用后端的查询引擎 `Search.py`
- 5) 查询引擎把结构返回给 web 服务器, 此时 web 服务器调用渲染引擎, 选择 `index` 模板, 把查询列表填充到该引擎中。其中每个列表结果包含: 新闻标题, 时间, 摘要, 原始新闻 url, 新闻在搜索引擎中的 id 以及 具体新闻页地址 (该地址会在渲染过程中组装好)
- 6) 点击查询列表中的“查看全文”, 会跳转如下格式的地址: `http://0.0.0.0:1111/news?id=18`, 然后该地址以同样的方式传递给搜索引擎, 并获得全部的新闻正文内容, 由渲染引擎渲染 `news` 模板, 返回到浏览器中

3 系统测试情况

3.1 依赖项

在运行本系统前，需要安装以下依赖项：

1、scrapy:

安装方法为： pip install scrapy， Ubuntu 系统可直接使用命令 sudo apt-get install python-scrapy。

2、webpy:

安装方法为： sudo easy_install web.py。官方网站为：
<http://webpy.org/>。

3、jieba:

安装方法为：pip install jieba

本系统已采集了 10 万余条网易新闻、头条新闻等网站的网页、倒排索引、word2Vec 等数据可供下载，地址：
<https://pan.baidu.com/s/1pLvOrR9>，提取码为 h7x6。如果不准备重新爬取数据、建立倒排索引、训练 word2Vec 模型，可下载后，将 data 文件夹数据放在 Information_retrieva_Projectl-/data 目录。

3.2 爬虫测试

在源代码所在的目录下运行命令行窗口，输入：“scrapy crawl netese”，可以看到(要求先在命令行下切换至配置文件 scrapy.cfg 所在目录)：

```
de24685..e2409d0 master -> master
jt@jt:~/Information_retrieva_Projectl-$ scrapy crawl netese
2016-05-29 20:08:58 [scrapy] INFO: Scrapy 1.0.5 started (bot: crawl)
2016-05-29 20:08:58 [scrapy] INFO: Optional features available: ssl, http11
2016-05-29 20:08:58 [scrapy] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'crawl.spiders', 'SPIDER_MODULES': ['crawl.spiders'], 'BOT_NAME': 'crawl'}
2016-05-29 20:08:58 [scrapy] INFO: Enabled extensions: CloseSpider, TelnetConsole, LogStats, CoreStats, SpiderState, AutoThrottle
2016-05-29 20:08:58 [scrapy] INFO: Enabled downloader middlewares: HttpAuthMiddleware, DownloadTimeoutMiddleware, UserAgentMiddleware, RetryMiddleware, DefaultHeadersMiddleware, MetaRefreshMiddleware, HttpCompressionMiddleware, RedirectMiddleware, CookiesMiddleware, ChunkedTransferMiddleware, DownloaderStats
2016-05-29 20:08:58 [scrapy] INFO: Enabled spider middlewares: HttpErrorMiddleware, OffsiteMiddleware, RefererMiddleware, UrlLengthMiddleware, DepthMiddleware
2016-05-29 20:08:58 [scrapy] INFO: Enabled item pipelines:
2016-05-29 20:08:58 [scrapy] INFO: Spider opened
2016-05-29 20:08:58 [scrapy] DEBUG: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2016-05-29 20:08:58 [scrapy] DEBUG: Telnet console listening on 127.0.0.1:6023
2016-05-29 20:08:58 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/> (referer: None)
2016-05-29 20:09:03 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/jnews/> (referer: http://news.163.com/)
2016-05-29 20:09:07 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/air/16/0529/10/B07P2PMI00014P42.html> (referer: http://news.163.com/)
2016-05-29 20:09:09 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/air/16/0529/13/B0854SK700014P42.html> (referer: http://news.163.com/)
2016-05-29 20:09:10 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/air/16/0529/14/B086EDR200014P42.html> (referer: http://news.163.com/)
2016-05-29 20:09:10 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/air/16/0529/09/B07LP90000014P42.html> (referer: http://news.163.com/)
2016-05-29 20:09:10 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/echo_of_radiation/> (referer: http://news.163.com/)
2016-05-29 20:09:10 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/000113C4/163kanke.html> (referer: http://news.163.com/)
2016-05-29 20:09:11 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/16/0525/20/BNUK94RU00014PRF.html> (referer: http://news.163.com/)
2016-05-29 20:09:11 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/00011269/gdmore.html> (referer: http://news.163.com/)
2016-05-29 20:09:11 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/16/0525/16/BNU30SHM000156PO.html> (referer: http://news.163.com/)
2016-05-29 20:09:11 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/singledog_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/heiqiang_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/leguan_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/hatefulight_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/fuguoji_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/special/cheating_pc/> (referer: http://news.163.com/)
2016-05-29 20:09:12 [scrapy] DEBUG: Redirecting (301) to <GET http://view.163.com/> from <GET http://news.163.com/review/>
2016-05-29 20:09:13 [scrapy] DEBUG: Crawled (200) <GET http://news.163.com/16/BKII1RF5000012QEA.html> (referer: http://news.163.com/)
```

抓取完成后，完成后在 `data` 目录下得到格式如下的文件 `netease_data.txt`

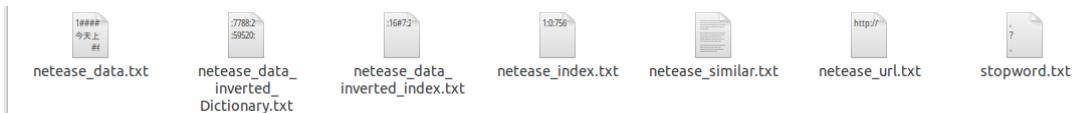
```
1#####网易新闻#####
    全球各国央行何以都在搞负利率，其实是本国政府都靠印钞为生。两岸都这么说，此番军演的“年度例行”性质也就完全坐实。对于这些网络里检举、相信举报者也是冒着很大的风险。随着中国海上力量的增长，会削弱美国在南海的统治性地位。韩国直播圈爆出赤裸丑闻，美女主播利用网络直播进行性交易。说“暗门”表面上是追求享乐的表现，背后暗藏的却是权力黑洞。去年10月蔡英文访问日本时，专门参观了日立制作所车间。当前中国的这个通胀大部分的构成要素是制造业的产品。扫描二维码下载网易新闻客户端####http://news.163.com/#####发现者专访####最新专题专访伦敦马拉松主办方：赛事如何不扰民艾滋病患者隐瞒病情致医生感染可避免；实现普遍防护，艾滋病已由一种绝症逐渐变成可控慢性病，关键要积飞机高空飞行遇鸟击非常罕见。若真发生，单引擎也往期回顾TED演讲风靡全球，它成功的经验是：好的内容加好的传播。飞机备降，是飞行过程中时常会遇到的情况。当飞行目的地的天气状况不部分像“小龙虾”这样的外来入侵物种如今却是风靡全国的美食，有人近来日本现行国歌《君之代》被热议。实际上，它的形成经历了漫长的过抑郁症是一种需要专业治疗的大脑疾病。但市场上出现的五花八门抑郁症不同于人们通常的理解的情绪波动，它是一种大脑疾病，是社会。随着现在医患关系的紧张，医院很多操作都要征求患者同意。有担当的医直升机会否能在灾区空投受到一系列因素的影响，仅从客观条件上看，气最近在非洲再次爆发的埃博拉病毒病是世界上最凶猛疾病之一，病死率高7月21日，习近平在委内瑞拉总统马杜罗陪同下抵达国家公墓，委内瑞拉。近日，有专家提出，通过修建城市风道，把郊外的风引进主城区，将雾霾最近科技界热议一个事件：俄罗斯计算机软件首次成功通过图灵测试，媒日本在垃圾处理方面有着世界一流的技术和丰富的经验。垃圾处理在日本世界卫生组织接受网易专访表示，疫苗是可使用的最安全的医用品，但没船长在一艘船上拥有最高权力和责任，发生海难后，船长必须根据当时的目前马航失联客机黑匣子电量或已耗光，无法发出信号。搜救团队只能启####http://news.163.com/special/interviewdiscoverer1/#####统计：71年来共133架飞机消失，1837人下落不明。网
```

3.3 索引构建测试

```
#part：从原始爬取数据中 处理得到 倒排索引和词典
...
buff_size=1024*1024*10
output_record_size=10000
filename="data/netease_data.txt"
inverted_files.make_inverted_index(filename,buff_size,output_record_size,100
000)
similar_doc.establish_document_index("data/netease_data.txt",buff_size,"data
/netease_index.txt")
...

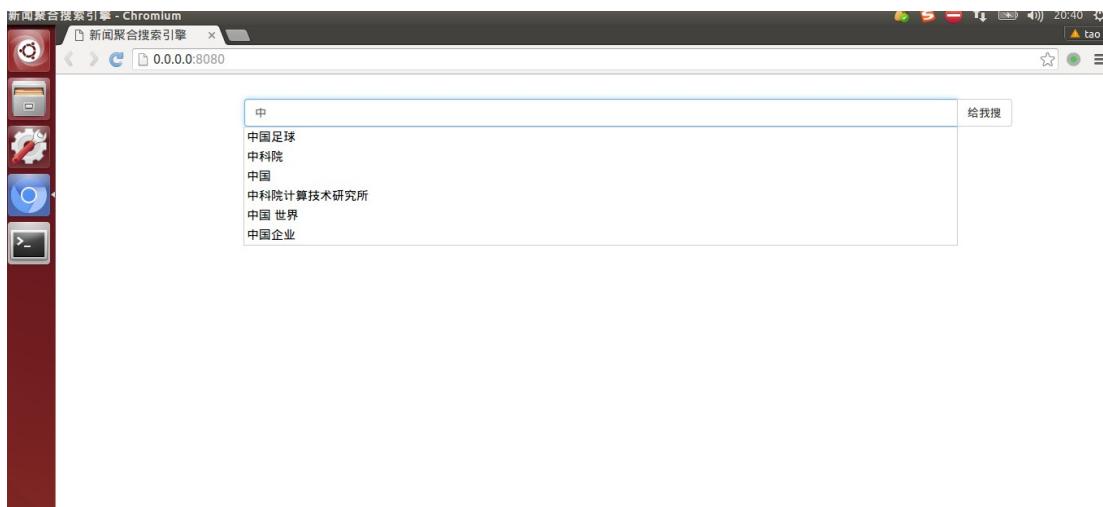
```

取消 `main.py` 中如上注释，运行 `main.py`，完成索引构建，`data` 文件夹下得到：



3.4 系统主界面

完成以上步骤后或从网盘中下载文件至 `data` 目录下，在 `linux` 系统下进入 `web/` 文件夹，在终端中运行 `python main.py`，在浏览器中打开网址：`http://0.0.0.0:8080/`。



3.4 查询测试

启动系统，在浏览器中输入查询字符串“信息检索”，得到的结果如下。

因为此文大量出现信息检索，因此相关度排在了第一名，是符合搜索准则的，因此通过测试。

在正文的左边，会出现关键词推荐，“机器翻译、在线翻译、图像处理、网络软件”均是和“信息检索”一样是领域相关，排在第一的为“考试试题”则是搜索相关。

新闻聚合搜索引擎 - Chromium

新闻聚合搜索引擎

0.0.0.0:8080

给我搜

中

中国足球
中科院
中国
中科院计算技术研究所
中国 世界
中国企业

新闻聚合搜索引擎 x 0.0.0.0:8080/?searchword=信息检索

给我搜

信息检索

关键词推荐:

- 历年试题
- 机器翻译
- 在线翻译
- 图像处理
- 网络软件
- 课程管理
- 计算机管理
- 静力学
- 外贸英语
- 搜索引擎

网易新闻中心--方正通网络入侵检测系统让网络充满阳光

入侵检测技术起源于对入侵行为以及相应防范手段的研究，但是权威数据表明，70%的损失是由于组织内部的行为造成的，而且很多内部员工滥用网络资源，带来很多安全隐患和工作缺乏效率。所以，内网安全防护和网络行为监控是网络安全市场的重中之重。从技术手段上，为了防范内网的安全必须对内部网络各类异常行为的进行监控、分析、记录、预测、响应。上海方正科技软件有限公司研制的方正通网络入侵检测系统就具备上述内网安全防护和网络行为监控的手段。[查看全文>>](#)

原始网页链接:http://news.163.com/2004w04/12524/2004w04_1082107336386.html

网易新闻中心--日本出版业现状：出的多卖的少面临崩溃

中国日报网站消息：本书对上世纪末开始的日本图书业的崩溃作了详实的介绍和评述。其中有些情形与中国颇不相同，如作为出版社与书店之间的经济中介的“图书交易公司”，在中国就很少见，许多出版泡沫的出现恰恰与之相关。但也有很多情况与我们相似，如出书过滥，退货堆积如山，大书店挤垮小书店，等等。他山之石，可以攻玉，日本的出版危机应当引起我们的警觉。今天摘介的，主要是出版社与书店所陷入的经营怪圈。下期介绍处于“革命”中的日本出版业如何挣扎和自救。[出得多，卖不动第二次世界大战后的日本出版业，在发展的过程 ...](#) [查看全文>>](#)

原始网页链接:http://news.163.com/2004w04/12536/2004w04_1083119347839.html

网易新闻中心--全面集成整体应用——用友全力打造制造业信息化应用价值链

随着世界制造中心向中国转移，大大小小的中国制造企业立志做大做强，不仅想要赢得国内市场的竞争，还要挺进国际市场，扩展企业的经营疆域。在这一背景下，“国际管理，中国制造——用友ERP创造制造业最佳管理实践高峰会”将于2004年3月在全国东、西、南、北四大区域陆续举行。用

点击“原始网页链接”，跳转到原始网页：



新闻中心 NEWS.163.COM WEB NEWS

您目前的位置 国易首页 > 新闻中心 >

方正通网络入侵检测系统让网络充满阳光

新闻中心 http://news.163.com
2004-04-16 17:22:16 来源:

发布评论 查看评论

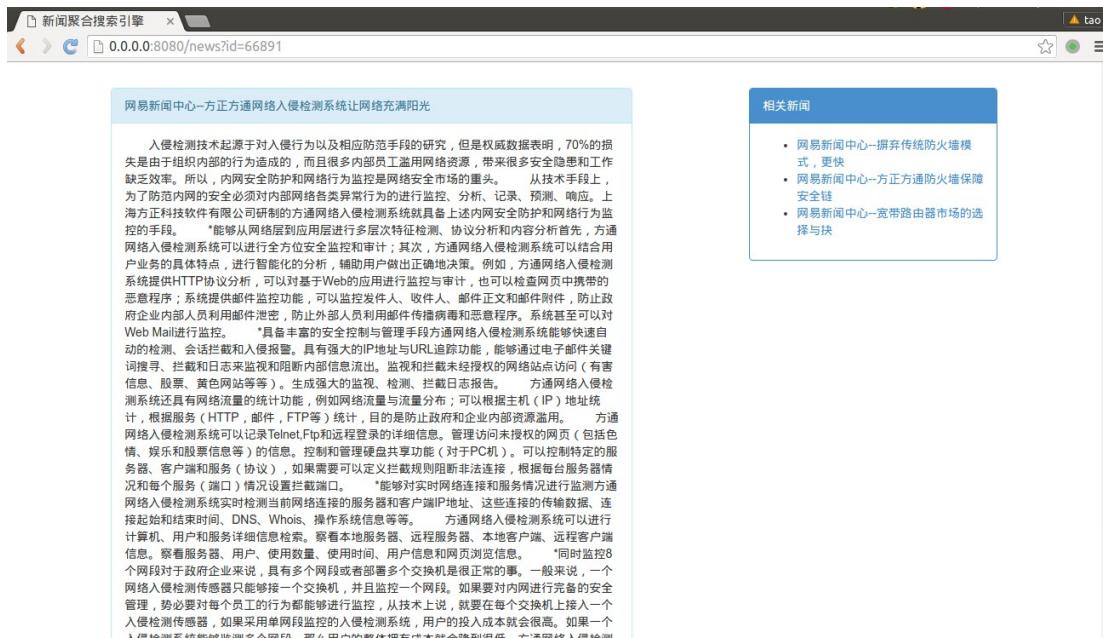
入侵检测技术起源于对入侵行为以及相应防范手段的研究，但是权威数据表明，70%的损失是由于组织内部的行为造成的，而且很多内部员工滥用网络资源，带来很多安全隐患和工作效率低下。所以，内网安全防护和网络行为监控是网络安全市场的重点。

从技术手段上，为了防范内网的安全必须对内部网络各类异常行为的进行监控、分析、记录、预测、响应。上海方正科技软件有限公司研制的方通网络入侵检测系统就具备上述内网安全防护和网络行为监控的手段。

*能够从网络层到应用层进行多层次特征检测、协议分析和内容分析首先，方通网络入侵检测系统可以进行全方位安全监控和审计；其次，方通网络入侵检测系统可以结合用户业务的具体特点，进行智能化的分析，辅助用户做出正确地决策。例如，方通网络入侵检测系统提供HTTP协议分析，可以对基于Web的应用进行监控与审计，也可以检查网页中携带的恶意程序；系统提供邮件监控功能，可以监控发件人、收件人、邮件正文和邮件附件，防止政府企业内部人员利用邮件泄密，防止外部人员利用邮件传播病毒和恶意程序。系统甚至可以对Web Mail进行监控。

*具备丰富的安全控制与管理手段方通网络入侵检测系统能够快速自动的检测、会话拦截

点击“查看原文”：



相关新闻

- 网易新闻中心--摒弃传统防火墙模式，更快
- 网易新闻中心--方正通防火墙保障安全链
- 网易新闻中心--宽带路由器市场的选择与抉

3.5 结果排序&关键字推荐测试

在浏览器中输入不同查询，系统返回结果分别如下：

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=蒋介石

蒋介石

给我搜

关键词推荐:

- 林彪
- 李宗仁
- 彭德怀
- 白崇禧
- 袁世凯
- 粟裕
- 何应钦
- 国民政府
- 斯大林
- 胡宗南

标志作为大国首次参与国际外交 蒋介石访印内幕_网易新闻中心

何平 蒋介石访印前的远东战局 1941年12月7日，日军偷袭珍珠港。12月8日，日军在香港、马来亚和泰国登陆，并轰炸在菲律宾的美军及其设施。与此同时，日本海军在马来亚海岸附近击沉英国远东舰队主力舰“威尔士亲王”号；12月25日，驻香港英军投降。1942年1月2日，日军攻占马尼拉，驻菲律宾的美军撤往山区。2月7日，日军在新加坡登陆，西方强国在远东的殖民大厦岌岌可危。在中国战场，1941年底，日军打到湘北和鄂西，离抗日大后方四川和陪都重庆已不远，中国抗日战争进入关键时期。令蒋介石高兴 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/41125/1/161QQR9T00011247.html>

历史揭密：毛泽东和蒋介石鲜为人知的遗愿_网易新闻中心

70年代上半期，对于毛泽东、蒋介石来说，是他们人生的最后岁月。历史把他们的希望与遗憾、成功与失败、喜悦与忧伤交织到生命的最后。1972年3月，在台湾上空的浓重阴云中，蒋以86岁高龄出任第五届“总统”，悲壮宣誓：“只要毛共及其同党一日尚存，我们革命的任务不会终止，纵使我们必须遭受千百挫折与打击，亦在所不惜，决不气馁。”但“英雄”暮垂，他的精神已支撑不住他的宏志。健康每况愈下，多种疾病与车祸交相而至，最后三年，他只公开露面三次。毛泽东推动了乾坤的转移，但接连的胜利并未给这位巨人带来太多的喜悦 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/41116/0/15A6IMU400011246.html>

蒋介石官邸失火 外界猜测是反蒋极端分子所为_网易新闻中心

新华网北京3月20日电 中共中央最近发出《关于进一步繁荣发展哲学社会科学的意见》（以下简称《意见》）。《意见》强调指出，在全面建设小康社会、开创中国特色社会主义事业新局面、实现中华民族伟大复兴的历史进程中，哲学社会科学具有不可替代的作用。必须进一步提高对哲学社会科学重要性的认识，大力繁荣发展哲学社会科学。《意见》分七部分：一、繁荣发展哲学社会科学是建设中国特色社会主义的一项重大任务；二、繁荣发展哲学社会科学的指导方针；三、繁荣发展哲学社会科学的目标；四、实施马克思主义理论研究 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12497/2004w03_1079768564836.html

网易新闻中心--将“三个代表”重要思想载入宪法是时代的需要

陈奎元 内容摘要 “三个代表”重要思想同马克思列宁主义、毛泽东思想和邓小平理论是一脉相承而又与时俱进的科学体系，是面向21世纪的中国化的马克思主义，是新世纪新阶段全党全国人民实现全面建设小康社会宏伟目标的根本指针。将“三个代表”重要思想载入宪法，确立其宪法地位，反映了党的主张和人民意志的统一，体现了我们依法执政的坚定理念，实现了国家指导思想的又一次与时俱进。中国共产党是中国特色社会主义事业的领导核心，是以马克思主义理论为指导的中国工人阶级的先锋队和中国人民、中华民族的先锋队。中国共 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12507/2004w03_1080613432257.html

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=刘德华

刘德华

给我搜

关键词推荐:

- 张学友
- 谭咏麟
- 容祖儿
- 陈奕迅
- 曾志伟
- 汪明荃
- 古巨基
- 郑秀文
- 杨千嬅
- 何韵诗

小伙整容超刘德华成名 刘德华表示成功无捷径_网易新闻中心

本报讯(东亚记者张南)21岁的长春小伙子吴可，长得与香港明星刘德华有几分相似，上学期间有人管他叫小刘德华，他还曾参加过明星模仿秀节目模仿刘德华还获得了奖，现在他是一名业余歌手。他想成名而且还想超过刘德华，为此他选择了整容，以便让自己更像刘德华，以圆明星梦。12月8日，记者在长春某医疗美容机构看到了这位小伙子。吴可身高在1米73左右，他还有一个艺名叫吴珂帆。从外表看，他的整体形象还不错，仔细观察眉目肖的他，还真有什么地方分刘德华的神态。在与记者交谈时，吴可显得有些腼腆，还没 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/41209/0/174E9R4A00011228.html>

一见钟情 喻可欣：狗仔队助华仔离开我(组图)_网易新闻中心

□ □ 《神雕侠侣》这部连续剧在马来西亚极受欢迎，于是他应邀去作秀。他打电话给我，很兴奋地问我说：“我要去马来西亚作秀，你能不能来？我很希望你能来陪我。”感觉到他在电话中迫切的渴望，我答应了，并坚定地告诉他：“无论如何我都会去。”□ □ 跟他在电话中说好没多久，我就病了，一开始以为只是皮肤过敏，但是身体内阵阵剧痛袭来，令我痛苦不堪。妈妈带我去看医生，医生说，我得的病名叫做蛇缠身。疼痛像电击一样，无预警地间歇性袭来。妈妈真是担心极了，当然不让我出门，更别说出国了。万一没有照顾好会受到感染，而且会 ... [查看全文>>](#)

原始网页链接:http://news.163.com/50109/7/19KVF5EV0001122D_2.html

情海星空——我与刘德华(一)_网易新闻中心

我和刘德华偶遇在星空，却坠入情海中浮沉。记得是高三的那一年，有个朋友想要去报考演员训练班，非要我陪着他去不可，我们报了名，结果很顺利，两人都考上了。我还记得当时的主考官是知名导演徐进良先生，受了几个月的训练后，结业时，徐导演、邵氏的方逸华小姐，还有琼瑶

图 1: 查询人物

新闻聚合搜索引擎 > 0.0.0.0:8080/?searchword=台湾

台湾

给我搜

关键词推荐:

- 台湾地区
- 大陆
- 岛内
- 台北
- 来台
- 马英九
- 台湾当局
- 中国台湾
- 台当局
- 香港

网易新闻中心-《台湾问题与中国的统一》

一、台湾是中国不可分割的一部分 台湾地处中国大陆的东南缘，是中国第一大岛，同大陆是不可分割的整体。台湾自古即属于中国。台湾古称夷洲、流求。大量的史书和文献记载了中国人民早期开发台湾的情景。距今一千七百多年以前，三国时吴人沈莹的《临海水土志》等对此就有所著述，它们是世界上记述台湾最早的文字。公元三世纪和七世纪，三国孙吴政权和隋朝政府都曾先后派万余人去台。进入十七世纪之后，中国人民在台湾的开拓规模越来越大。十七世纪末，大陆赴台开拓者超过十万人。至公元一八九三年(清光绪十九年)时，总数达到 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12494/2004w03_1079510660891.html

网易新闻中心-台湾问题的由来与实质

第二次世界大战结束之后，台湾不仅在法律上而且在事实上已经归还中国。台湾问题的出现，是国民党发动反人民内战的结果，其本质是中国的内政问题。台湾问题之所以长期存在且迄今尚未解决的一个重要因素，是美国等西方反华势力插手台湾问题，干涉中国内政，阻碍中国统一。台湾问题是中美关系中最重要的、最敏感的核心问题。虽然台湾问题尚未最终解决，海峡两岸尚未统一，但世界上只有一个中国，台湾是中国的一部分，中国的领土和主权完整不容分割。台湾问题的出现是国民党发动反人民内战的结果 1945年抗日战争胜利后，中国人民 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12494/2004w03_1079510810418.html

新闻聚合搜索引擎 > 0.0.0.0:8080/?searchword=杭州

杭州

给我搜

关键词推荐:

- 南京
- 苏州
- 绍兴
- 无锡
- 福州
- 合肥
- 嘉兴
- 济南
- 温州
- 广州

网易新闻中心-西湖春风万般情——杭州以民为本求真务实纪事

“呀，这不是老沈吗？怎么这么巧。”浙江省委常委、杭州市委书记王国平一踏进商报热线值班室，就一眼认出杭州市“十佳”文明市民、八十高龄的沈伯欣。王国平一边紧握老人的手，一边热情地向周围的同志介绍，沈大爷是一位关心杭州城市建设的好市民，为市委、市政府的工作出了好多金点子，那关于中国茶叶博物馆周边环境的改造、西湖西进工程中恢复杨公堤六桥等都是很好的建议。沈大爷乐呵呵地说，关心自己所生活的城市，为杭州建言献策是应该的。他刚到热线值班室，送的是自己设计的一份新年礼物：一枚圆圆的印章，中有金猴捧桃的图 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w02/12449/2004w02_1075625303253.html

海宁临杭新区拍出新“地王” 每亩367.8万元_网易新闻中心

今年4月2日，嘉兴市委常委、海宁市委书记俞志宏在“海宁连杭经济区实施五大工程，再造一个海宁动员大会”上，提出了“将连杭经济区建设成为杭州城市副中心”的宏大设想。如今，这一设想正在被一个个“杭州元素”证实。5月27日，杭州市第一条跨市域城市道路——人民大道海宁段开通；7月28日，浙江财经学院东方学院一期工程开工.....10月30日，诞生于临杭新区人民大道边上的海宁新“地王”，进一步证实“杭州元素”在默默“托”起海宁临杭地价。10月30日，海宁市临杭新区人民大道北侧一块面积38.47亩的商办用地，经过 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/09/1103/09/5N6FUCIJ000120GR.html>

争建跨海大桥为哪般？看杭州湾三座跨海桥建桥热_网易新闻中心

正如火如荼建设中的宁波杭州湾跨海大桥。 不过更多的地方政府看中的是其长期效应。大桥未通，宁波环杭州湾产业带规划已实施，绍兴也提出产业兴市的规划。杭州发展的重心则从“西湖时代”转向“钱江时代”，即向杭州湾靠拢。“我们去长三角各城市考察，发现距离上海的车程控制在两小时

图 2 查询地点

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=中科院

中科院 搜索

关键词推荐:

- 中国科学院
- 研究所
- 刘先林
- 地学部
- 中国科学院心理研究所
- 副研究员
- 院士
- 中国林科院
- 陈俊勇
- 同济大学

网易新闻中心–路甬祥：中国科技发展再也不能停留于模仿与跟踪

新华网北京3月17日电（记者俞铮）中国科学院院长路甬祥17日在此间说，中科院要形成促进原始科学创新的根本机制，大力发展战略高技术，树立“创新跨越、引领未来、竞争合作、增值循环”的新的科技发展观。路甬祥在中科院2004年度工作会议上说：“中国科技发展再也不能停留于一般的模仿与跟踪，而必须具有实现跨越发展的胆识和魄力，增强做原始性科学创新、做世界一流技术创新与集成的信心和勇气。”他说，原始科学创新的动力源于人们对自然现象的好奇心和探索精神，促进原始科学创新的根本机制在于优先领域，鼓励 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12494/2004w03_1079509754799.html

网易新闻中心–独家专访 卢柯：最年轻的院士领跑纳米技术

2003年中国青年年度人物评选今年3月30日在北京揭晓并颁奖，38岁的中国科学院院士卢柯当选年度科学家。作为中科院最年轻的院士，他的实力与魅力都让人惊讶。 卢柯，1965年5月生，甘肃华池人，20岁时毕业于南京理工大学机械系，同年到中科院金属研究所攻读研究生；25岁在金属研究所获得博士学位；1993年28岁时被中科院金属研究所聘为研究员；刚刚30岁就成为博士生导师；32岁担任“快速凝固非平衡合金国家重点实验室”主任。2003年被增补为中国科学院院士；2004年3月30日在中国青年年度人物评选 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w04/12510/2004w04_1080892556785.html

网易新闻中心–全国首个超常班创建“神童班”并不制造天才

自开班后，媒体对超常班的关注已经让这些孩子对镜头习以为常了。 幼儿超常班3月1日开班，19个5岁幼儿通过中科院心理所的测试被选进这个超常班，经过层层测试、遴选出来的19个超常儿童已经在这里接受了一个多月的超常教育。 □□这些孩子究竟超常在哪里？超常教育真能造出“神童”吗？ ... [查看全文>>](#)

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=清华大学

清华大学 搜索

关键词推荐:

- 北京大学
- 浙江大学
- 中国农业大学
- 武汉大学
- 南开大学
- 北京师范大学
- 南京大学
- 同济大学
- 华东师范大学
- 中山大学

拒绝调解终胜诉 清华大学捍名誉告倒“李鬼”_网易新闻中心

因为滥用“清华大学”校名，北京中天华亿科技有限公司被清华大学告上法庭。记者今天9时获悉，海淀法院判决北京中天华亿科技有限公司立即停止使用含有“清华大学”或“清华”字样的宣传内容，赔礼道歉并赔偿公证费5000元。 法院的判决全部支持了清华大学的诉讼请求，因为开庭时被告就承认了一切并愿承担所有责任，但因为原告清华大学不同意调解，所以法院最终用判决书的方式使清华大学胜诉。 法院判决：被告立即停止使用含有“清华大学”或“清华”字样的宣传内容，停止使用并销毁含有侵权内容的宣传材料；北京中天华亿科技 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/41222/0/187VQTEG0001122E.html>

网易新闻中心–中国互联网发展大事记

1. 1986年，北京市计算机应用技术研究所实施的国际联网项目—中国学术网(Chinese Academic Network，简称CANET)启动，其合作伙伴是德国卡尔斯鲁厄大学(University of Karlsruhe)。 2. 1987年9月，CANET在北京计算机应用技术研究所内正式建成中国第一个国际互联网电子邮件节点，并于9月14日发出了中国第一封电子邮件：“Across the Great Wall we can reach every corner in the wor ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w04/12528/2004w04_1082442059215.html

清华大学51名首届国防生致信江泽民的前前后后_网易新闻中心

●也许我们以后的生活中没有大都市的繁华，更不可能腰缠万贯，但享乐安逸不属于我们，无私奉献才是我们的责任。 ●献身国防，是实现我们报国理想和个人成才的最佳选择。 ●蓬勃发展的国防事业为我们提供了一个大舞台。我们要在这个舞台上施展自己的才华，燃烧自己的青春和激情。 ——摘自清华大学首届国防生给江泽民主席的信*相关报道：江主席勉励清华大学毕业生国防建功立业 清华首届国防生给江主

图三：查询机构

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=帅气

帅气

给我搜

关键词推荐:

- 干练
- 温婉
- 蓝美
- 飘逸
- 端庄
- 活脱
- 优雅
- 气质
- 冷艳
- 打扮

网易新闻中心-豪门甘地新生代再战印度政坛 家族分裂历史延续

豪门甘地新生代再战印度政坛 家族分裂历史从上一代延续到新一代 印度总统2月初解散国会下院，新议会选举提前至4月中旬，政坛豪门尼赫鲁·甘地家族再次成为关注的焦点。这一次走在台前的是家族新生代——拉胡尔、普利扬卡和瓦伦。瓦伦于2月16日加入人民党，使原本一家亲的三兄妹成为政坛对手，家族分裂的历史也从上一代延续到新一代身上。 拉胡尔的沉着稳重和普利扬卡的高贵迷人为在野的国大党重新执政带来希望，而瓦伦的出身及诗性才华则是人民党手中的新王牌，但无论谁的成功，都将续写甘地家族的辉煌。 拉胡尔年龄 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12490/2004w03_1079207218747.html

网易新闻中心-有志者事竟成 清华大学食堂师傅托福考630分

中新网5月14日电托福满分670分，清华大学高材生考过600分也不易，可一个每天三顿为清华学子切菜卖饭的农民工，头回上场就爆冷门—— 北京日报报道，清华园里人才济济，但学生食堂的师傅张立勇，也为许多人知道，就不能不说是一件新鲜事了。他的艰辛，他的刻苦，他的顽强，让清华学子动容。水木清华BBS上，头一回为一个农民工的坚韧好学掀起波澜。 张立勇今年29岁，做农民工10年，如今即将拿到北大国际贸易专业大本文凭。 从包装箱上的英文说明学起 张立勇和千千万万农家子弟一样，做过考大学“跳龙门”的 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w05/12552/2004w05_1084502868482.html

美国女大学生爱上重庆农民 缘自一声英文招呼_网易新闻中心

“我于北京时间2005年2月7日，去了中国重庆开县的一个小镇，那是我男朋友的家乡，是一个非常美丽的地方，我们约定于2月18日在这里举行婚礼。过完中国春节就去办结婚证。” 据《重庆晨报》报道，2月11日，今年23岁的美国俄亥俄州女大学生罗娜，在开县赵家镇一网吧，兴奋地将这

新闻聚合搜索引擎 0.0.0.0:8080/?searchword=漂亮

漂亮

给我搜

关键词推荐:

- 很漂亮
- 好看
- 可爱
- 赏心悦目
- 潇洒
- 聪明
- 绝妙
- 华丽
- 美妙
- 秀气

“神秘女郎”引领成都迪吧新职业 喝酒玩要是工作_网易新闻中心

夜幕笼罩下的蓉城，各式迪吧酒吧里热闹非凡。近来，每到晚上9时，几名漂亮时髦的美眉就会准时出现在城南某大型迪吧，她们将在这里呆到凌晨1时才离开。她们作为一群特殊的“客人”，迪吧不仅为她们免费提供酒水和小吃，还将付给她们不菲的新手费。前日，记者近距离接触了这些“神秘女郎”。 22岁的陈真真是遂宁人，在成都读完大学后，她留在了这个不是很熟悉而又充满魅力的城市。眼看几个月时间都没能找到工作，她于上月接受了一家中介公司为她提供的这个特殊的职业机会。 一天的生活几乎都是从下午5时开始的。起床后，煲煲电 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/41205/2/16RDQ9KS0001122B.html>

网易新闻中心-宁静敞开心扉实话实说 缺少明星的气质(图)

“我真的不适合做演员，我的性格太不随和了。我没有做明星的气质，当我出门时候，有人追着签名合影，我说不清楚是高兴还是不高兴，我就是激发不起来那种高昂的情绪。我不愿意接受采访，是因为我不晓得自己说什么，我也不愿意拍照片。我已经习惯自己这种个性了，跟着自己没心没肺的身体走了30多年，何必去改呢？我总结了一下，像我这种人，往往能够把戏演好，因为我不太注意人际关系，只专注一件事情。我非常高兴我有这样的性格，这是我最大的缺陷，也是我最大的优点。” 宁静气喘吁吁小跑着进来，嘴里问着几点了，脸上带着 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w04/12523/2004w04_1082015293664.html

网易新闻中心-枭龙试飞员与研制者,潜在市场价值1000亿(图)

枭龙01 副总工程师师朝林 枭龙试飞员梁万俊(左)与枭龙试飞员王文江(右) 新闻背景： 中国“枭龙”一飞冲天 “枭龙”飞机是中国和巴基斯坦共同合作开发的,1999年中国和巴基斯坦正式签订研制合同,这是中国军机研制的历史跨越。2003年8月25日,“枭龙”01架实现首次飞行。

图四：查询形容词

新闻聚合搜索引擎

0.0.0.0:8080/?searchword=我爱你

给我搜

关键词推荐:

- 好想你
- 阿杜
- 爱上你
- 命中注定
- 我心
- 生日快乐
- 圣诞节快乐
- 祝你成功
- 真美
- 祝你幸福

网易新闻中心-人民网特别策划:那些甜蜜、浪漫的——动人情话

1.《简爱》——你以为，我因为贫穷，低微，矮小，不美我就没有灵魂没有心吗？你想错了我的灵魂跟你一样充实，我的内心跟你一样地丰富。我们站在上帝的脚跟前，我们是平等的！ 2.《第一次亲密接触》——如果还有一天寿命，那我要做你女友。我还有一天的命吗？.....没有。所以，很可惜。我今生仍然不是你的女友。如果把整个浴缸的水倒出，也浇不熄我对你的爱情的火。整个浴缸的水全部倒得出吗？.....可以。所以，是的。我爱您..... 3.《大话西游》——曾经有一份真诚的爱情放在我面前，我没有珍惜，等我失去的时候我才后悔莫 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w02/12460/2004w02_1076605775888.html

网易新闻中心-小试身手大获成功 埃尔兰总理千金走红文坛(图)

现年22岁的塞西莉亚·埃亨是爱尔兰总理伯蒂·埃亨的千金。去年，她凭首部小说杀进文坛，取得巨大成功。迄今为止，这部小说已经卖出上百万册。除在爱尔兰热销，世界各国出版商也在争相出版这本书，好莱坞著名制片人甚至还出巨资买下版权，准备将这个催人落泪的爱情故事搬上银幕。塞西莉亚一夜之间声名大噪，并坐拥上百万美元收入。 小试身手 大获成功 塞西莉亚的首部小说题为《补充一句，我爱你》，去年在爱尔兰首都都柏林出版后，立即成为畅销书，首印5万册很快售罄。 小说还引起国际 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12484/2004w03_1078670286703.html

网易新闻中心-男子钟情三陪女遭拒 血书“我爱你”后将其斩死

南方网讯 22岁的北京男子赵某，网恋一个三陪女子遭拒后，将其杀死。为示钟情，赵某行凶前在“情人”所住楼外血书“我爱你”3个大字。昨天（19日），朝阳检察院以涉嫌故意杀人罪将其批准逮捕。 去年12月16日，朝阳一小区内发生命案。警方勘察现场时发现，女死者身上有12处锐

新闻聚合搜索引擎

0.0.0.0:8080/?searchword=哈哈

给我搜

关键词推荐:

- 呵呵
- 啊
- 哎
- 唉
- 呀
- 哎呀
- 哇
- 哩哩
- 嗯
- 哟

网易新闻中心-娃哈哈特邀香港女作家梁凤仪“三八节”书赠杭城知识女性

在被誉为“最具女人味的城市”，在“三八”节这样一个纯粹女性的节日里，娃哈哈特别邀请梁凤仪，以特殊的方式向杭州的女性祝贺节日。 3月8日下午，梁凤仪出现在娃哈哈美食城一楼圆厅，向杭州的女性朋友送上自己现场签名的系列丛书。 作为杭州的一张“金名片”，娃哈哈以2003年全年营业收入突破100亿元、在全球饮料企业排名中与两乐等跨国企业共同跻身五强等业绩再次引发业界的强烈关注。作为杭州市民而言，娃哈哈连续3年承办“娃哈哈西湖狂欢节”的大手笔深得杭城百姓的好评。同时，娃哈哈以不断开发生产健康饮品的进 ... [查看全文>>](#)

原始网页链接:http://news.163.com/2004w03/12486/2004w03_1078814398755.html

2.14撞车_网易新闻中心

春节、情人节撞车怎么过？小情侣网上发帖公开吵架被誉为2010年最大杯具的2.14，很快就要来临了。今年的2月14号是考验一个男人要亲情还是要爱情还是要自己的日子：大年初一、情人节、NBA全明星周末，在这一天三合一，比老妈和老婆同时掉河里救谁的选择还要难。请各位男士注意，这不是一个段子，关于情人节、春节撞车的杯具真的已经开演了。近日在天涯社区，一对情侣为了2.14在哪儿过，上演了一出发帖吵架秀。男友发帖女友跟2月5日，天涯网友“咖喱煎蛋”发帖倒苦水，讲述“2.14引发的战争”。他说：“有谁比我还郁闷 ... [查看全文>>](#)

原始网页链接:<http://news.163.com/10/0210/02/5V4K66L7000120GR.html>

爱孩子 爱未来_网易新闻中心

城市让生活更美好，孩子让未来更美好！每个宝宝都是可爱的天使，是爸爸妈妈的希望，是整个大家庭的希望，是国家未来的希望。先进的教育理念、科学的关爱观是建立一个孩子良好成长环境的重要保证。为迎接2010年上海世博会的到来，更好地关心下一代的成长，向您征集爱的寄语，爱孩

图五：查询短句

综上，可以看出查询返回的文档是按照相关性排序，同时我们注意到关键字推荐达到了很高的准确率，有明显的效果。

3.6 相关搜索推荐测试

分别点击查看原文，观察推荐的新闻：

The screenshot shows a web browser window with the URL 0.0.0.0:8080/news?id=64442. The main content area displays an article titled "网易新闻中心--美国田纳西州井水遭受污染 居民不满激增". The article discusses how residents in Dickson County, Tennessee, are unhappy due to polluted wells. On the right side of the page, there is a sidebar titled "相关新闻" (Related News) which lists three other news items:

- 世纪逢春 千年一梦 资都景德镇再度雄起
- 网易新闻中心--弱女遭前夫绑架和暴打被
- 网易新闻中心--广东江门134票香港货物

The screenshot shows a web browser window with the URL 0.0.0.0:8080/news?id=34425. The main content area displays an article titled "美国梦或是美国空想？_网易新闻中心". The article discusses the 1992 US presidential election. On the right side of the page, there is a sidebar titled "相关新闻" (Related News) which lists three other news items:

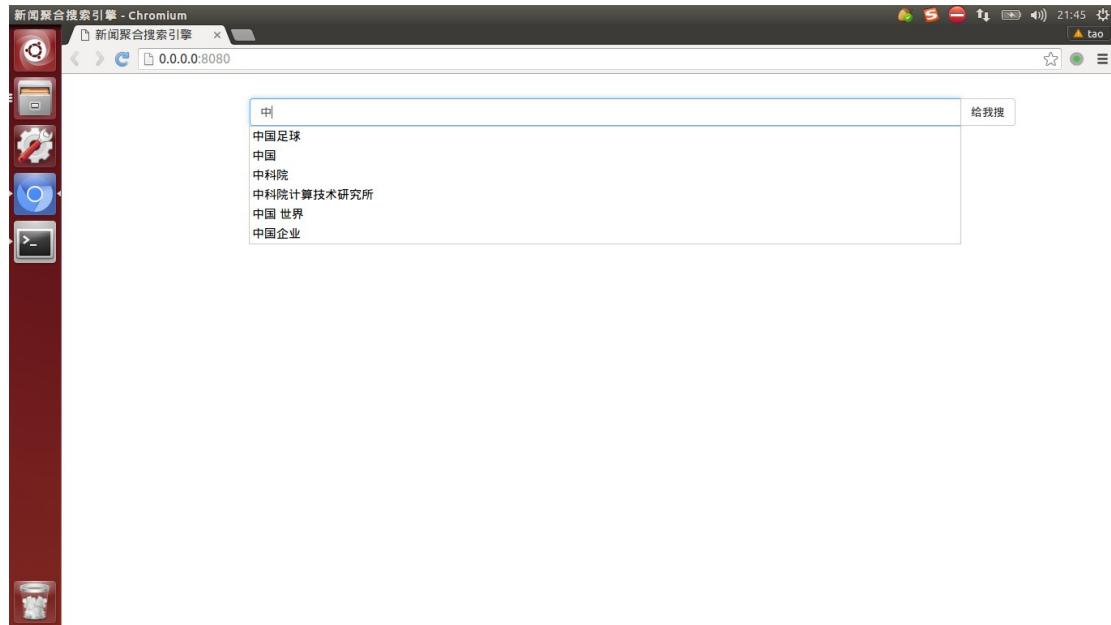
- 网易新闻中心--浙江温州掀起“效能革命”
- 网易新闻中心--“效能革命”席卷温州向
- 网易新闻中心--广东各界人士纪念辛亥革命

3.7 网页快照测试

在浏览器中输入查询字符串，点击某条记录的查看全文即为网页的快照。。

3.8 查询自动补全测试

在浏览器中输入查询字符串“中”，结果如下。



4 创新&总结

4.1 创新点

- (1) 通过建立神经网络模型，建立词向量模型，关键字推荐模块的准确性很高。
- (2) 构建指向倒排索引词项字节地址的字典文件，实现快速从磁盘中读取词项的倒排记录表，适合倒排记录表文件大。

4.2 总结

在各位组员的共同努力之下，我们从刚开始的无从下手，到逐渐明确所需要做的事情，再到逐步实现我们的系统，最终完成这次的大作业——新闻搜索

引擎的设计与实现，这是一个曲折而具有一定难度的过程。虽然我们实现的新闻搜索引擎还略微简陋，不过“麻雀虽小，五脏俱全”，它已经包含了一个完整搜索引擎的所有基本模块，通过逐步实现这些基本模块，我们基本掌握了一个完整搜索引擎的工作过程和构建流程，并且加深了对在课堂上所学到的知识的理解和运用，同时也学到了不少新的知识和技能，锻炼了我们的团队协作能力。

5 致谢

首先非常感谢老师给了我们这次锻炼的机会，让我们提高了编程的动手能力，加深了对信息检索以及搜索引擎相关知识的深刻认识和理解。同时感谢所有组员的辛勤付出，从开始的没有头绪到过程中的一点一滴的反复思考、认识和实践，最终完成此次项目作业。另外还感谢那些当我们遇到问题时求助过的所有同学们。

6 参考文献

1. scrapy 手册 http://scrapy-chs.readthedocs.org/zh_CN/1.0/intro/tutorial.html
2. 结巴分词 手册: <https://github.com/fxsjy/jieba>
3. webpy 手册 <http://webpy.org/>
4. python 手册 <https://docs.python.org/3/tutorial/index.html>
5. 《信息检索导论》，王斌
6. 《自然语言处理》，宗成庆。