

新闻检索平台

1 功能概述

本软件为一个新闻检索平台，实现了检索及显示10000条来自人民网的新闻的功能。

1.1 主页面

输入127.0.0.1:8000/search后，可得到以下初始界面。

新闻检索平台



引领中俄关系继续保持高水平发展--国际--人民网

Sept. 11, 2018, 5:38 a.m.

9月10日，在俄罗斯滨海边疆区阿尔乔姆市，宇博苏莫托利生产有限公司的工人正在对一辆汽车进行检查。 本报记者殷新宇摄 习近平主席将于9月11日至12日赴俄罗斯符拉迪沃斯托克出席第四届东方经济论坛。东方经济论坛由俄罗斯总统普京2015年亲自倡议举办，是俄罗斯推进远东合作的重要举措。中国国家元首首次出席东方经济论坛。

财经观察：中俄远东合作势头良好--国际--人民网

Sept. 10, 2018, 7:32 p.m.

新华社符拉迪沃斯托克9月10日电财经观察：中俄远东合作势头良好 新华社记者吴刚 得益于“一带一路”倡议的不断推进和与中国接壤的区位优势，俄罗斯远东地区与中国的合作呈现良好势头；连接两国的跨境基础设施建设迅速推进；农业和跨境物流成为现阶段双方合作的重点领域。 跨境基础设施打下合作基础 俄罗斯滨海边疆区与中国黑龙江省和

中俄携手打造东北亚共赢走廊--国际--人民网

Sept. 10, 2018, 8:41 a.m.

图为扎鲁比诺港。本报记者王峻岭摄 即将召开的第四届东方经济论坛涉及能源合作、农产品贸易、海洋经济、区域开发的相关议题将占据不小比重。其中，交通走廊和物流能力建设更是被摆在了显要位置。本报特派记者来到俄罗斯远东滨海边疆区东南部的扎鲁比诺港口，看到装卸机械正忙碌着将一箱箱铝锭装运上船。远处，一艘艘渔船载着刚打捞上来的海产品缓缓驶

俄国家杜马三读通过在俄远东设立特别行政区--国际--人民网

July 28, 2018, 2:31 a.m.

新华社莫斯科6月12日电（记者吴刚）俄罗斯国家杜马（议会下院）26日三读通过了一份关于在俄远东地区设立特别行政区的系列法律草案，以便在俄远东滨海边疆区设立经济区。 系列法律草案包括《关于在俄远东地区设立特别行政区的法律草案》和根据设立特别行政区的需要而对俄现行的国际公司法、民法典、税法典、外汇管制法、商船航行法及一些

第五届远东媒体峰会在俄罗斯举行--国际--人民网

June 13, 2018, 4:34 a.m.

本报莫斯科6月12日电（记者张晓东）第五届远东媒体峰会日前在俄罗斯远东滨海边疆区首府符拉迪沃斯托克举行。此次会议由俄罗斯滨海边疆区政府、俄罗斯记者协会和远东国立大学联合举办，主要议题包括传统媒体与新媒体的融合、现代传媒技术与发展、中俄两国媒体合作等。 俄罗斯记者联盟主席索洛维耶夫认为，在俄中合作不断走深的背景下，俄罗斯媒体应

0123456789>

找到结果10000个 用时0.000秒

可在本页面上端的搜索框中输入关键词进行搜索，也可通过下方的翻页键浏览新闻。

1.2 搜索结果界面



上图为搜索结果页面, 搜索关键词为"习近平 普京 特朗普", 搜索引擎会返回所有与关键词部分匹配的页面, 且匹配程度越高, 排名越靠前。对于出现在标题、摘要中的关键词, 会高亮显示。

同时, 搜索引擎还支持按时间检索, 若搜索关键词中包含用&&包起来的两个日期, 则会返回发布时间在这两个日期之间的文章。下图是输入&2017-01-01-2017-01-01&后的返回页面。

新闻检索平台



&2017-01-01-2017-01-01&

联合国安理会全票通过叙利亚问题决议--国际--人民网

Jan. 1, 2017, 11:33 a.m.

中新社联合国12月31日电联合国安理会31日一致通过第2336号决议，欢迎并支持日前由俄罗斯和土耳其促成的叙利亚停火协议。中国常驻联合国副代表吴海涛称，中方希望叙利亚政府和有关反对派切实、全面执行停火协议，也呼吁其他反对派尽快加入停火安排。主要由俄罗斯和土耳其促成的叙利亚停火协议于当地时间12月30日零时开始生效。安理会3

默克尔发表新年讲话：“我们比恐怖主义更强大”--国际--人民网

Jan. 1, 2017, 1:38 p.m.

2017年新年前夕，德国总理默克尔发表新年讲话，呼吁在德国生活的人们，乐观迎接新年，团结抵抗恐怖主义。德国总理默克尔在对全国发表的新年谈话中称，过去的一年是德国人民经受“重大考验的一年”，其中恐怖主义“无疑是最严峻的考验”。在提及维尔茨堡、安斯巴赫和柏林圣诞市场的袭击时，她说这些“宣称要在我们的国家寻求保护的人”却从事恐怖袭击是“

吴哥古迹见证中国东盟友谊--国际--人民网

Jan. 1, 2017, 5:12 a.m.

扫描二维码 看更多内容 悠扬的古乐在千年巴戎寺响起，优美的舞姿在古老吴哥胜境跃动。2016年12月28日晚，2016中国—东盟联合文化展演在柬埔寨暹粒吴哥世界文化遗产所在地举行，来自中国和东盟十国的文化团体为现场观众带来了一场艺术盛宴，也为中国与东盟建立对话关系25周年纪念活动画上一个圆满的句号。当

巴基斯坦外交国务部长法塔米：巴基斯坦举国支持中巴经济走廊建设--国际--人民网

Jan. 1, 2017, 1:41 p.m.

12月30日，巴基斯坦外交国务部长法塔米接受新华网记者专访。新华网记者何险峰摄 新华网北京1月1日电（巩阳）“巴基斯坦的确有很多党派，但是有一点，没有一个政党反对中巴经济走廊建设。”正在北京访问的巴基斯坦外交国务部长法塔米30日告诉新华网记者。法塔米日前来京参加中巴经济走廊远景规划联合委员会第六次

古特雷斯新年履新 联合国迎来“新掌门人”--国际--人民网

Jan. 1, 2017, 1:43 p.m.

2016年4月12日，葡萄牙前总理、联合国前难民事务高级专员安东尼奥·古特雷斯在纽约联合国总部。（新华社记者李木子摄）2017年1月1日，新年伊始，联合国新任秘书长古特雷斯安东尼奥·古特雷斯正式履新，成为联合国历史上第九任秘书长。任期自2017年1月1日起至2021年12月31日。他究竟是怎样一个人？他是如何成为联合国这个最

0 1 2 >

找到结果11个 用时0.006秒

1.3 详情界面

点击上述两个页面中的新闻标题后，可进入新闻的详情界面。

新闻检索平台



巧克力飘香里斯本斗牛场--国际--人民网

新华社里斯本2月12日电(记者章亚东、张柯)一年一度的里斯本巧克力节12日落下帷幕,西班牙、比利时、法国、秘鲁和巴西等国80家巧克力厂家带来的各式顶级巧克力食品,给市中心著名建筑里斯本斗牛场带来了浓郁的巧克力醇香。

在春意盎然的里斯本,一杯香气浓郁的热巧克力对参观者来说必不可少。即使再挑剔的食客,也能在这里找到自己的最爱,水果巧克力、酒心巧克力、干果巧克力、无糖巧克力.....种类繁多,不一而足。

里斯本巧克力节今年已是第四年举办。主办方工作人员路易松对记者说,主办方希望将展会办成一个巧克力盛宴,既能让食客们品尝到各类顶级巧克力食品,也能更全面地介绍巧克力文化。

同往年一样,主办方今年仍举办了一系列讲座和现场展示,包括来自米其林餐厅的大厨现场展示巧克力甜品制作,鼓励食客互动,创造带有个性标签的巧克力。

60岁的伊蕾娜是一家本土巧克力品牌的代表,这是她连续第二年参加里斯本巧克力节。今年她带来了品牌特色产品——松露巧克力和已有20年历史的独家巧克力糕点。她对记者说,参加巧克力节使企业在售卖产品的同时也为品牌做了宣传,今年的销售额与去年基本持平,很满意。

- 葡萄牙总统德索萨表示欢迎难民来葡--国际--人民网 Feb. 9, 2017, 2:26 p.m.
- 葡萄牙总理科斯塔--国际--人民网 Oct. 8, 2016, 4:38 a.m.
- 葡萄牙人唱中文歌比赛在里斯本举行--国际--人民网 April 17, 2018, 4:51 a.m.

详情页面展示了新闻的全文,并推荐3篇数据库中相似性最高的文章。

2 代码实现

2.1 相关配置

爬虫使用python2,其版本号为2.7.13,使用的第三方库有 `requests 2.19.1`, `pandas 0.23.4`

其余部分使用python3,其版本号为3.7.0,所使用的第三方库有: `Django 2.1.1`, `jieba 0.39`, `numpy 1.14.5`, `pandas 0.23.3`

2.2 爬虫

从网页<http://world.people.com.cn/n1/2018/0911/c1002-30284709.html>开始,对网页的相关新闻版块进行bfs,提取每个网页的标题、发表时间、正文,从0开始按爬取时间对所有爬取到的网页进行编号,把标题、发表时间存至 `info.csv` 文件中,将正文存到以编号命名的txt文件中。爬虫代码位于 `scripts/crawler.py` 中

2.3 数据处理

从爬取到的文件中提取文章摘要,存到 `info.csv` 中。

使用 `jieba` 的search模式对每篇文章的标题及正文进行分词,记录文中出现的词及出现次数(词在标题中每出现一次按照5次计入),使用出现次数在2次及以上的词建立倒排索引,在索引表中同时存储出现次数,倒排索引存储于 `mix.csv` 中

上述代码主要位于 `scripts/segment.py` 中

使用 `tf-idf+jaccard similarity` 进行相关文章推荐,首先计算每个关键词的 `idf` 值,导入到 `mix.csv` 中,再枚举所有文章对,计算每一对文章的相似程度,存到 `similarity` 文件夹中以文章编号命名的csv文件中。此部分代码位于 `scripts/similarity.py` 中

之后，将 `info.csv`、`mix.csv`、`similarity/*.csv` 中内容导入Django创建的sqlite3数据库中，此部分代码位于主要位于 `database_insert.py` 中

2.4 搜索算法

使用打分的方法进行搜索。

若开启了时间检索，则为位于这个时间区间内的文章加20000分，并将 `threshold_time` 值设为20000，否则将 `threshold_time` 值设为0。

若使用了关键词检索，则将 `threshold_text` 值设为1000，否则设为0。对于检索使用的每个关键词，若在某篇文章中出现，则为该篇文章加 $1000 + \log(t)$ 次，其中 t 表示关键词在文章及标题中的出现次数（标题中每出现一次按5次计算）。每个关键词有1000分的基础分保证了出现关键词较多的文章得分会较高，使用 `log` 函数是为了适当平衡出现次数多的关键词和出现次数少的关键词的贡献。

之后，将得分大于等于 `threshold_time+threshold_text` 的文章返回，返回前将文章按得分从高到低排序。

2.5 服务器搭建

服务器使用Django进行搭建。

前端分为2个页面 `index.html` 与 `article.html`，`index.html` 用于展示主页面及搜索结果，并使用JavaScript实现关键词高亮。`article.html` 用于展示新闻详情及新闻内容。

这两个页面都是在网上下载的网页模板的基础上添加了搜索栏，更改了上部的网站名称而得到的，网站模板位于 `template` 文件夹中。

后端使用数据库进行交互，当用户点击文章详情时，读取本地静态文件进行显示，当用户进行搜索时，调用2.4中所描述的搜索算法进行搜索。