

新闻信息整合与检索系统

Designer: 计 73 李家昊

主要功能

新闻展示功能

新闻展示界面统计新闻总数，并列出所有新闻的标题和日期，每个新闻标题可链接到新闻详情页面。

NewsHomeAll News

NEWS

All News

We Have 23625 News in Total

16部门“三定”方案集中公布 明确“内设机构”	2018-09-12 08:38:07
在贪念中迷失自我 李青海严重违纪违法案剖析	2018-09-12 08:33:38
八字没一撇，典型树起来：“速成典型”“盆景典型”在冒头	2018-09-12 08:24:36
打出组合拳 让“老赖”不能赖	2018-09-12 09:03:30
国有煤矿排污2年收33份通知书 整改为何多次搁置？	2018-09-12 07:34:58
中国核电标准跻身强国有时间表 完整体系弥补短板	2018-09-12 07:04:10
渠道商身子扑下去，农产品价格翻上来	2018-09-11 10:22:51
长租公寓甲醛超标问题缘何多发？	2018-09-11 10:14:22
“雪龙2”号下水 可双向破厚冰实现非夏季极地考察	2018-09-11 07:35:33
北京：中小学教师申报高级职称须“支教”1年	2018-09-11 07:56:38
越来越多人义务捡拾垃圾 清洁环境是件很酷的事	2018-09-11 08:24:24

由于新闻总数过多，因此制作了分页功能，每页展示 50 条新闻

应急管理部：立即开展文物建筑、博物馆等消防安全检查

2018-09-04 07:50:35

工伤认定奔走数年 劳动关系证明成“拦路虎”

2018-09-04 07:44:40

通讯：中国“铁路红利”惠及东非大地

2018-09-03 13:13:19

记者手记：如何打通跨省就医刷社保卡结账的最大堵点？

2018-09-03 10:00:59

全国水网数据库正式建成 收录333万余条水系实体数据

2018-09-03 07:48:23

Prev12345678910Next

查询功能

多关键词查询

支持多关键词查询，查询时对输入字符串进行分词，对每一个词在倒排列表中进行检索。

NewsHomeAll News

NEWS

习近平发表讲话

Find 2816 Results in 0.16174774113109153 seconds

From Date

年 / 月 / 日

To Date

年 / 月 / 日

习近平 同志推动厦门经济特区建设发展的探索与实践

2018-06-22 16:28:07

新华社北京6月22日电 新华社特约记者 东海之滨，鹭岛厦门。这里千年浪涌，潮涨风起。 当中国改革开放的航船扬帆出港，历史的坐标就将其定位为中国最早设立的四个经济特区之一。这个曾经偏僻的海防小城，在40年改革开放中破浪前行，昭示出中国城市蝶变的密码。 如今，海风海浪依旧，厦门却已旧貌换新颜。”习近平 总书记对这座城市充满感情。就在一年前，金砖国家领导人厦门会晤时，他回首厦门经济特区...

习近平 主持召开深入推动长江经济带发展座谈会并发表重要 讲话

2018-04-26 20:44:29

新华社武汉4月26日电 中共中央总书记、国家主席、中央军委主席 习近平 26日下午在武汉主持召开深入推动长江经济带发展座谈会并发表重要 讲话。他强调，推动长江经济带发展是党中央作出的重大决策，是关系国家发展全局的重大战略。新形势下推动长江经济带发展，关键是要正确把握整体推进和重点突破、生态环境保护和经济发展、总体谋划和久久为功、破除旧动能和培育新动能、自我发展和协同发展的关系，坚持新发展理念，坚持稳中求...

“脱贫攻坚战一定能够打好打赢”——记 习近平 总书记看望四川凉山地区群众并主持召开打好精准脱贫攻坚战座谈会

2018-03-14 11:52:20

根据日期查询

可限定搜索的日期范围，进一步筛选新闻

NewsHomeAll News

NEWS

习近平

Find 67 Results in 0.06152448899439378 seconds

From Date

2018/09/01

To Date

2018/09/16

习近平 同俄罗斯总统普京举行会谈

2018-09-11 19:35:10

新华社符拉迪沃斯托克9月11日电（记者胡晓光 骆珺 郑晓奕）国家主席 习近平 11日在符拉迪沃斯托克同俄罗斯总统普京举行会谈。两国元首一致认为，今年以来，中俄关系呈现更加积极的发展势头，进入更高水平、更快发展的新时期。一致同意，无论国际形势如何变化，中俄都将坚定发展好两国关系，坚定维护好世界和平稳定。 习近平 指出，今年以来，我同总统先生分别在北京和约翰内斯堡举行了富有成效的会晤。中俄保持密切的高层...

习近平 会见日本首相安倍晋三

2018-09-12 14:04:08

新华社符拉迪沃斯托克9月12日电（记者骆珺 郝薇薇）国家主席 习近平 12日在符拉迪沃斯托克会见日本首相安倍晋三。 习近平 首先对不久前日本关西地区和北海道分别遭受台风、地震灾害，造成重大人员伤亡和财产损失表示慰问。 习近平 指出，当前，国际形势正在发生深刻复杂变化，不稳定不确定因素增多。中日作为世界主要经济体和地区重要国家，应该共同担起责任，为维护世界和地区和平稳定和发展繁荣发挥建设性...

习近平 发表视频祝贺纪念“一带一路”倡议在哈萨克斯坦提出5周年商务论坛开幕

2018-09-07 16:17:28

摘要生成及高亮显示

后端收到生成搜索结果后，在搜索结果的正文定位到用户输入的关键词，并生成包含关键词的摘要。具体逻辑是，若正文前 200 字包含关键词，则取前 200 字作为摘要，若前 200 字不包含关键词，则定位到关键词第一次出现的位置，取其前 100 字和后 100 字之间的内容作为摘要。

前端使用 mark.js 的代码，将摘要中包含关键字的部分高亮显示。

输入：中国

卡洛斯·阿基诺：中国发展如同列车飞驰

2018-09-08 11:47:10

从上世纪80年代末到今年的7月，我有幸到访中国20次，足迹遍及十几个省市。最初我只是一名游客，后来我作为东亚和中国经济社会学者前往中国参加学术会议。除了与中国同行交流，我更愿意用自己的脚丈量中国土地，用自己的眼睛观察中国社会的变迁。1989年我第一次到上海浦东时，浦东中心地带似乎没有超过5层的楼房，边缘地区完全是落后的农村。28年后我重返浦东时，这里出现了此前无法想象的新元素：直插云霄的东方...

输入：勇气

三代火车司机见证“中国速度”

2018-09-01 07:50:26

...着像吵架。”姜爱舜说，机车经常发生漏水漏油，操纵台红灯一亮司机心里就发慌，随时有可能停车，停车就是机车故障。“开火车讲究安全、正点和平稳。父亲一辈子开车平平安安，了不起！”姜爱舜敬佩父亲的干劲和勇气。2006年7月份，沪宁铁路迎来电气化时代，客货列车逐步更换为由我国生产的新一代电力机车牵引。姜爱舜也告别了ND5机车，开上了SS4电力机车。电力机车马力大，运行速度快，干净、噪音小，没有柴...

输入：习近平发表讲话

习近平主持召开深入推动长江经济带发展座谈会并发表重要讲话

2018-04-26 20:44:29

新华社武汉4月26日电 中共中央总书记、国家主席、中央军委主席习近平26日下午在武汉主持召开深入推动长江经济带发展座谈会并发表重要讲话。他强调，推动长江经济带发展是党中央作出的重大决策，是关系国家发展全局的重大战略。新形势下推动长江经济带发展，关键是要正确把握整体推进和重点突破、生态环境保护和发展、总体谋划和久久为功、破除旧动能和培育新动能、自我发展和协同发展的关系，坚持新发展理念，坚持稳中求进...

相关新闻推荐

使用了 tf-idf+余弦距离计算文本相似度的算法，

打开正文链接后，正文右方会显示推荐的新闻

例如打开标题为“中国东盟携手推进文化创意产业发展促进民心相通”的新闻时，系统会推荐与“东盟”“文化”有关的新闻

中国东盟携手推进文化创意产业发展促进民心相通

新华网 · 2018-09-11 19:20:04

新华社南宁9月11日电（记者唐荣桂）以“传承创新 发展共赢——中国—东盟文化创意产业的交流与合作”为主题的第13届中国—东盟文化论坛11日在广西南宁开幕。与会代表表示，中国与东盟在文化创意产业交流与合作上取得了丰硕成果，双方将进一步探索合作机制与路径，携手推进文化创意产业发展，促进民心相通。

中国文化和旅游部党组成员于群介绍，中国高度重视与东盟各国在文化创意产业领域开展交流合作。在中国—东盟博览会框架下，中国—东盟动漫游戏展已经成功举办两届。2017年4月，由中国国际动漫游戏博览会组委会牵头组成的中国动漫游戏展团首次亮相泰国动漫展。9月底，中国还将派代表团赴柬埔寨参加澜湄流域国家文化类中小企业研讨会。

老挝新闻文化旅游部新闻合作司副司长坎普·皮尔萨卡认为，中国和东盟在文化保护方面有很多可以相互借鉴之处，在文化交流方面取得的成果可以增进相互间的友谊。

广西壮族自治区文化厅厅长张虹介绍，在“一带一路”框架下，广西加大与东盟各国在数字影视动漫产业领域的交流合作，在东盟国家举办广西电视展播周，与东盟国家合作拍摄《海上新丝路》等影视节目，广西动画片《铜鼓传奇》《喀斯特神奇之旅》在泰国中央电视台播放。近5年来，广西共组派40多个代表团近千人次赴东盟十国开展文化交流合作。东盟各国的优秀演艺节目也走进广西，通过中国—东盟（南宁）戏剧周等平台展示，让中国观众领略东盟各国的人文风情。

相关推荐

[中国和东盟科技工作者探讨加强人工智能领域合作](#)

新华网 · 2018-09-11 11:22:19

[王毅会见东盟常驻代表委员会一行](#)

新华网 · 2018-09-10 18:26:15

[第五届“跨越太平洋——中国艺术节”在旧金山开幕](#)

新华网 · 2018-09-08 16:54:10

[歆慕而往 学有所用——“一带一路”的中东故事](#)

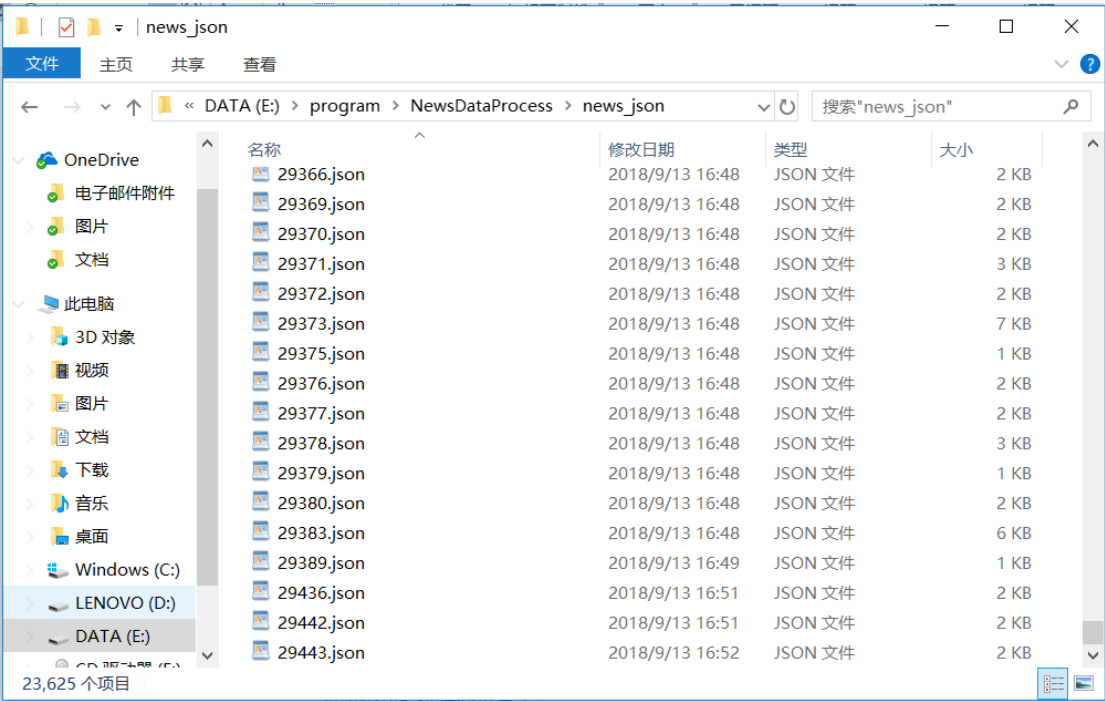
新华网 · 2018-09-11 13:14:42

[中国—东盟环境信息共享平台正式启动](#)

新华网 · 2018-09-11 19:38:58

总新闻量

系统的新闻全部从新华网上爬取，共爬取 23625 条新闻。每个.json 文件包括一条新闻的 id，标题、正文、日期信息。



查询时间及性能

例如输入中国，可找到 12218 条结果，用时 0.057 秒



Search

Find 12218 Results in 0.057255311537375064 seconds

From Date

To Date

年/月/日

年/月/日

输入习近平, 可找到 1407 条结果, 用时 0.055 秒



Search

Find 1407 Results in 0.05480238247445186 seconds

From Date

To Date

年/月/日

年/月/日

查询算法

制作了倒排索引列表，对每篇新闻进行分词后，将每个关键词对应的新闻 id 和出现频率存放在数据库内，共 8,000,000 余条记录。若关键词在新闻正文中出现，则词频 $FREQ+=1$ ，若关键词在新闻标题中出现，则词频 $FREQ+=20$ 。下面是表的末页：

Grid view		Form view	
			8198
	KEYWORD	IDX	FREQ
R197087	证史	29362	1
R197088	杭厂	29362	1
R197089	允后	29362	1
R197090	季夏	29373	1
R197091	侯伟	29373	1
R197092	票量	29373	1
R197093	闽琼	29373	1
R197094	席为	29373	1
R197095	各航司	29373	2
R197096	航变	29373	1
R197097	售取票	29373	1
R197098	机设	29373	1
R197099	机设备	29373	1
R197100	计算机设备	29373	1
R197101	杜柯欣	29378	1
R197102	牛书楠	29378	1
R197103	刘曼	29378	1
R197104	广慧	29378	1
R197105	78000	29378	1
R197106	6025	29378	1
R197107	家非	29378	1
R197108	4802	29378	1
R197109	要客	29378	1
R197110	孙尧	29378	1
R197111	孙学玉	29378	1
R197112	冯润	29378	1
R197113	师晨冰	29389	1
R197114	加拿大航空公司	29436	1
R197115	财湖	29443	2

当用户提交查询请求时，先对用户输入的字符串进行分词，得到分词列表，然后在倒排列表

中检索每一个分词对应的 id 和词频列表，然后将不同分词对应的相同 id 的词频相加，得到最终的 id 和词频列表，对其进行倒序排序，对应词频最高的 id 被排在搜索结果的第一位，最低的则排在最后一位。

推荐算法

采用 tf-idf+余弦距离的方式计算两篇新闻的文本相似度。tf 为一篇新闻中某个词出现的频率，idf 为所有新闻(语料库)中的逆文档频率。

比较两篇新闻时，具体做法是，先分别将两篇新闻出现次数最高的 40 个关键词提取出来，求其交集，不妨设交集中关键词个数为 m ，然后分别计算这 m 个词在两篇文章中的 tf-idf 值，生成两个 m 维 tf-idf 向量 \vec{x}, \vec{y} ，第 k 个分量即为第 k 个词的 tf-idf 值。

为保证余弦距离的可比性，需要保证两个 tf-idf 向量维数相同，因此定义第 k 个分量的排序指标为

$$key(k) = \vec{x}(k) + \vec{y}(k)$$

并对这些分量根据 key 进行倒序排序，然后分别取两个向量的前 20 个分量

$$\vec{x} = \vec{x}(0:20) \quad \vec{y} = \vec{y}(0:20)$$

作为新的 tf-idf 向量，计算这两个向量的余弦距离，即为两篇新闻的文本相似度。

$$\cos \theta = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|}$$

接下来将 23625 篇新闻拆分为 24 个子集，每个子集约 1000 篇新闻，在每个子集中对新闻两两进行文本相似度计算，最后对每篇文章提取出与之相似度最高的五篇文章，作为相关推荐内容。