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

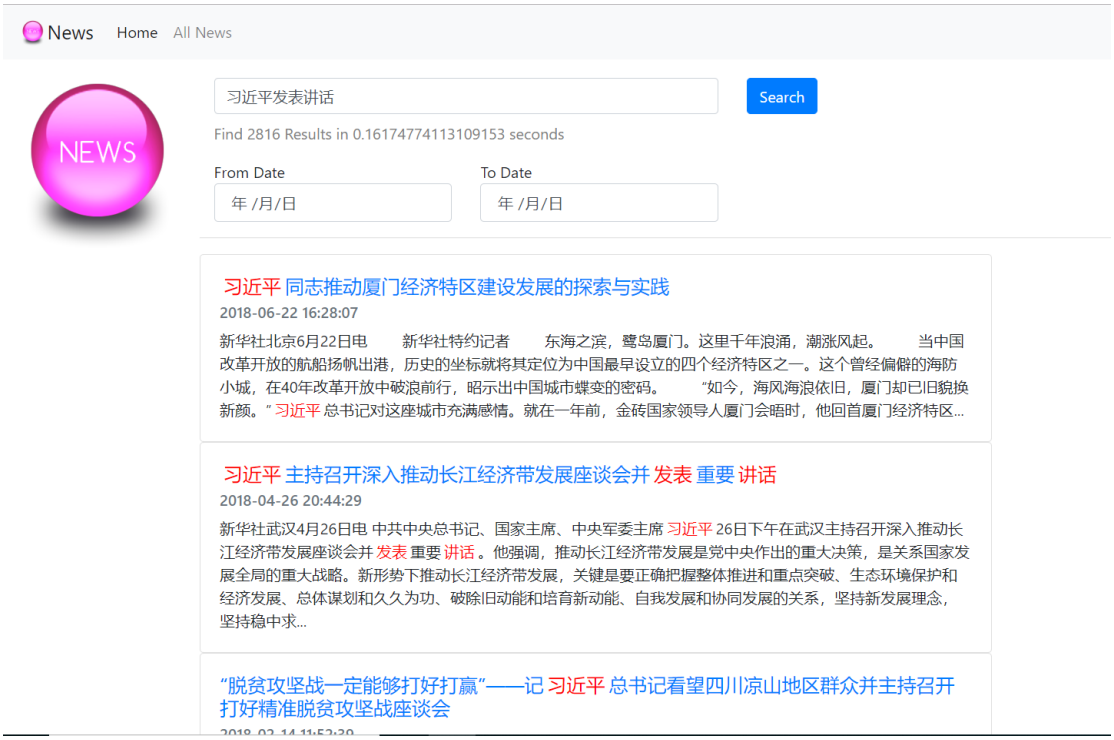
新闻展示界面统计新闻总数，并列出所有新闻的标题和日期，每个新闻标题可链接到新闻详情页面。

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

查询功能

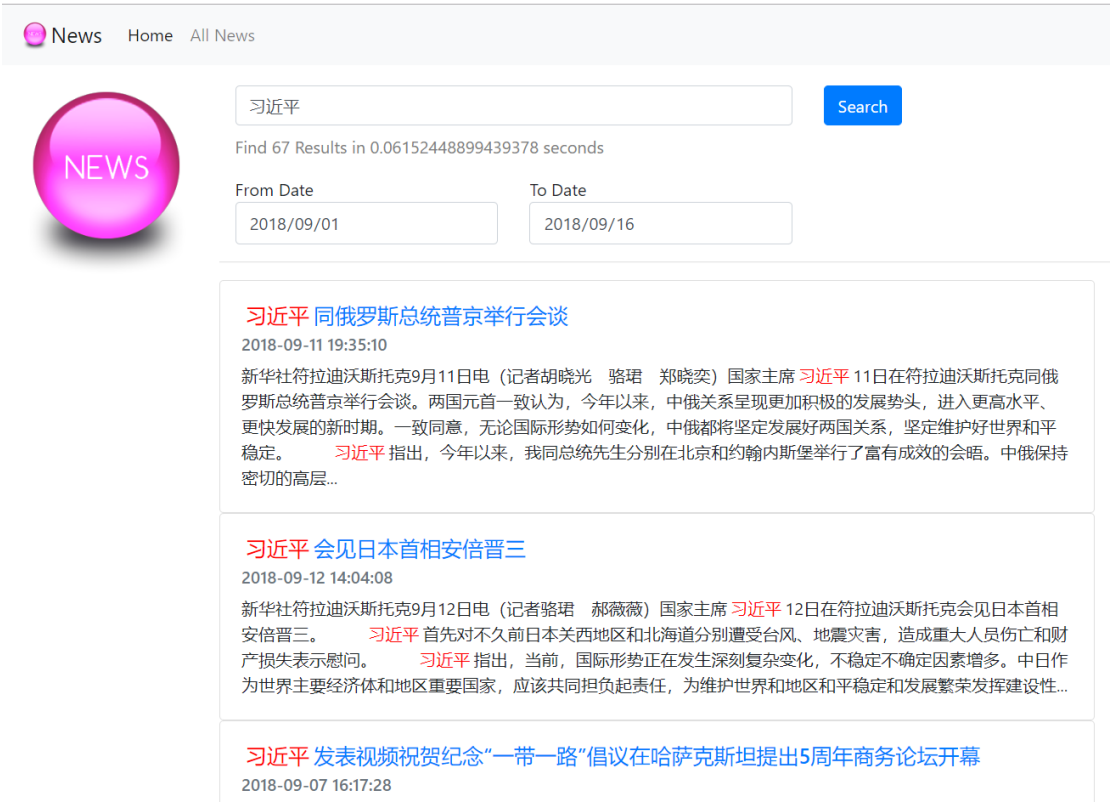
多关键词查询

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。



根据日期查询

可限定搜索的日期范围，进一步筛选新闻

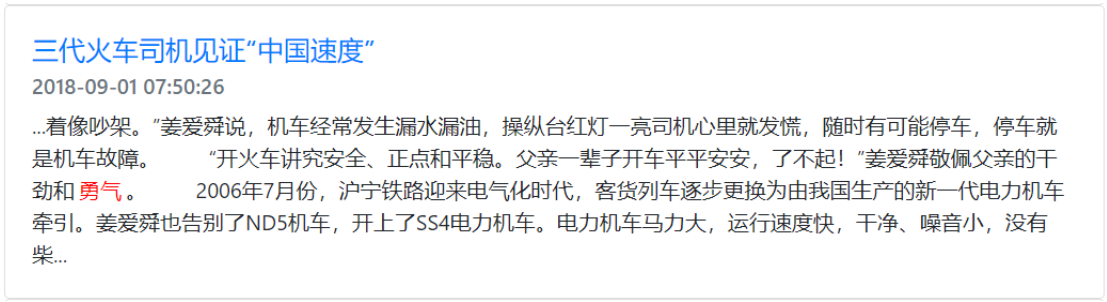


摘要生成及高亮显示

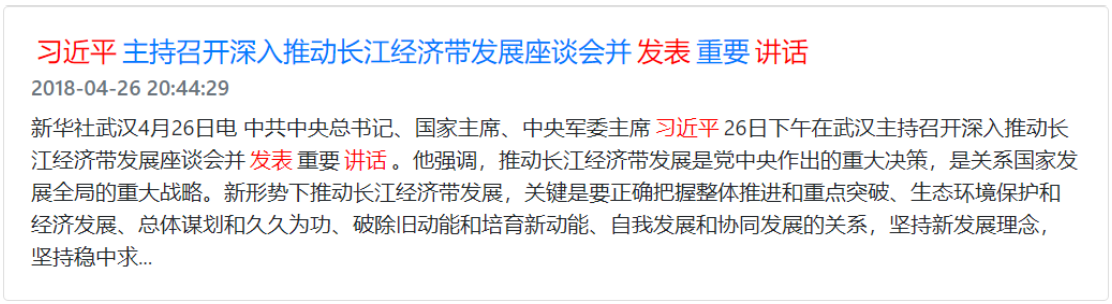
后端收到生成搜索结果后，在搜索结果的正文定位到用户输入的关键词，并生成包含关键词的摘要。具体逻辑是，若正文前 200 字包含关键词，则取前 200 字作为摘要，若前 200 字不包含关键词，则定位到关键词第一次出现的位置，取其前 100 字和后 100 字之间的内容作为摘要。
前端使用 mark.js 的代码，将摘要中包含关键字的部分高亮显示。

输入：中国

输入：勇气



输入：习近平发表讲话



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

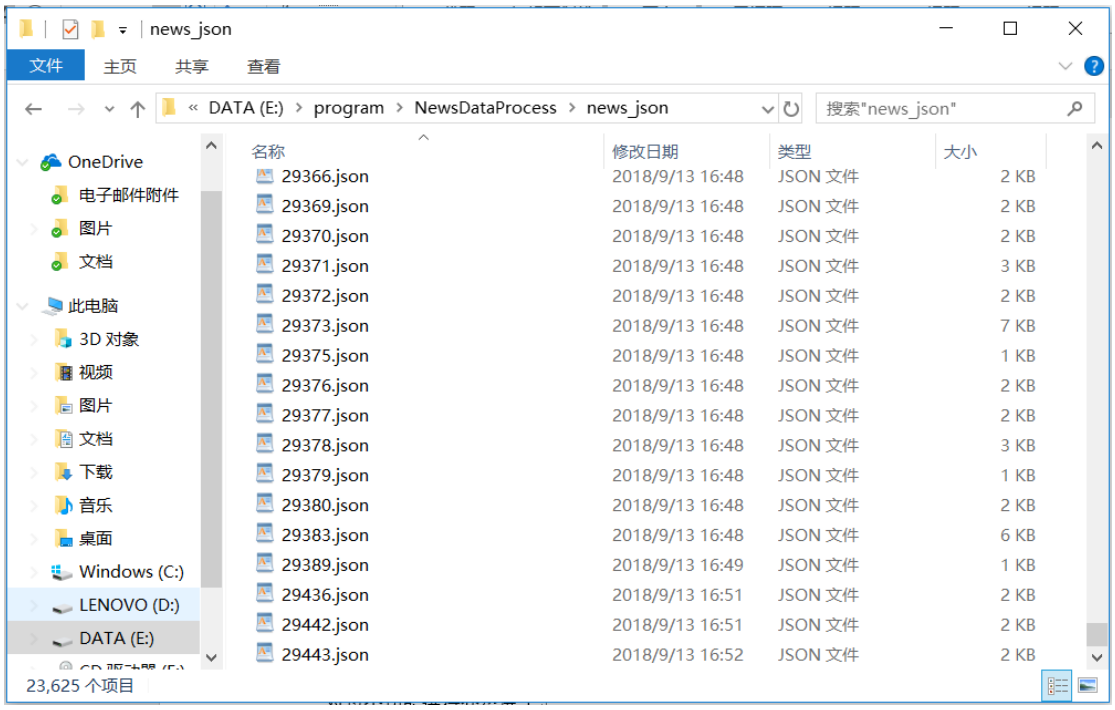
第五届“跨越太平洋——中国艺术节”在旧金山开幕
新华网 · 2018-09-08 16:54:10

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

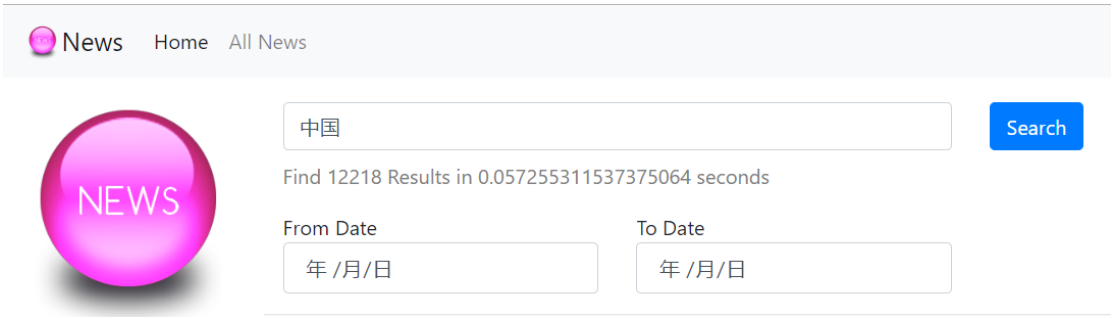
总新闻量

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。

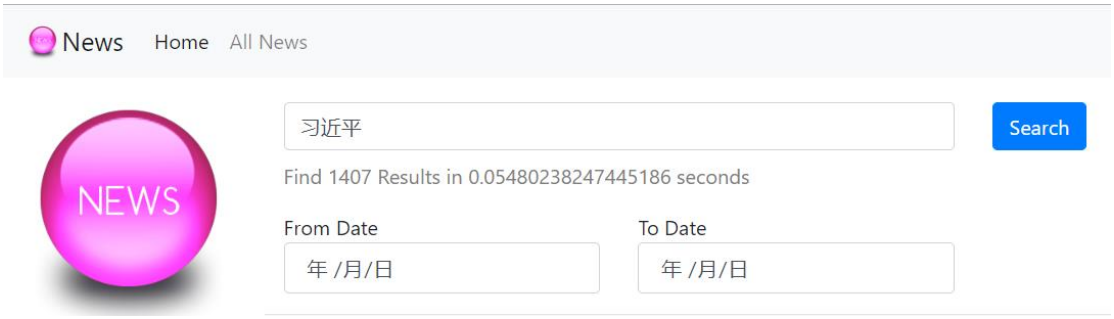


查询时间及性能

例如输入中国，可找到 12218 条结果，用时 0.057 秒



输入习近平，可找到 1407 条结果，用时 0.055 秒



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
8197087	证史	29362	1
8197088	杭厂	29362	1
8197089	允后	29362	1
8197090	季夏	29373	1
8197091	侯伟	29373	1
8197092	票量	29373	1
8197093	闽琼	29373	1
8197094	席为	29373	1
8197095	各航司	29373	2
8197096	航变	29373	1
8197097	售取票	29373	1
8197098	机设	29373	1
8197099	机设备	29373	1
8197100	计算机设备	29373	1
8197101	杜珂欣	29378	1
8197102	牛书楠	29378	1
8197103	刘曼	29378	1
8197104	广慧	29378	1
8197105	78000	29378	1
8197106	6025	29378	1
8197107	家非	29378	1
8197108	4802	29378	1
8197109	要客	29378	1
8197110	孙尧	29378	1
8197111	孙学玉	29378	1
8197112	冯润	29378	1
8197113	师晨冰	29389	1
8197114	加拿大航空公司	29436	1
8197115	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。