**新闻信息整合与检索系统**

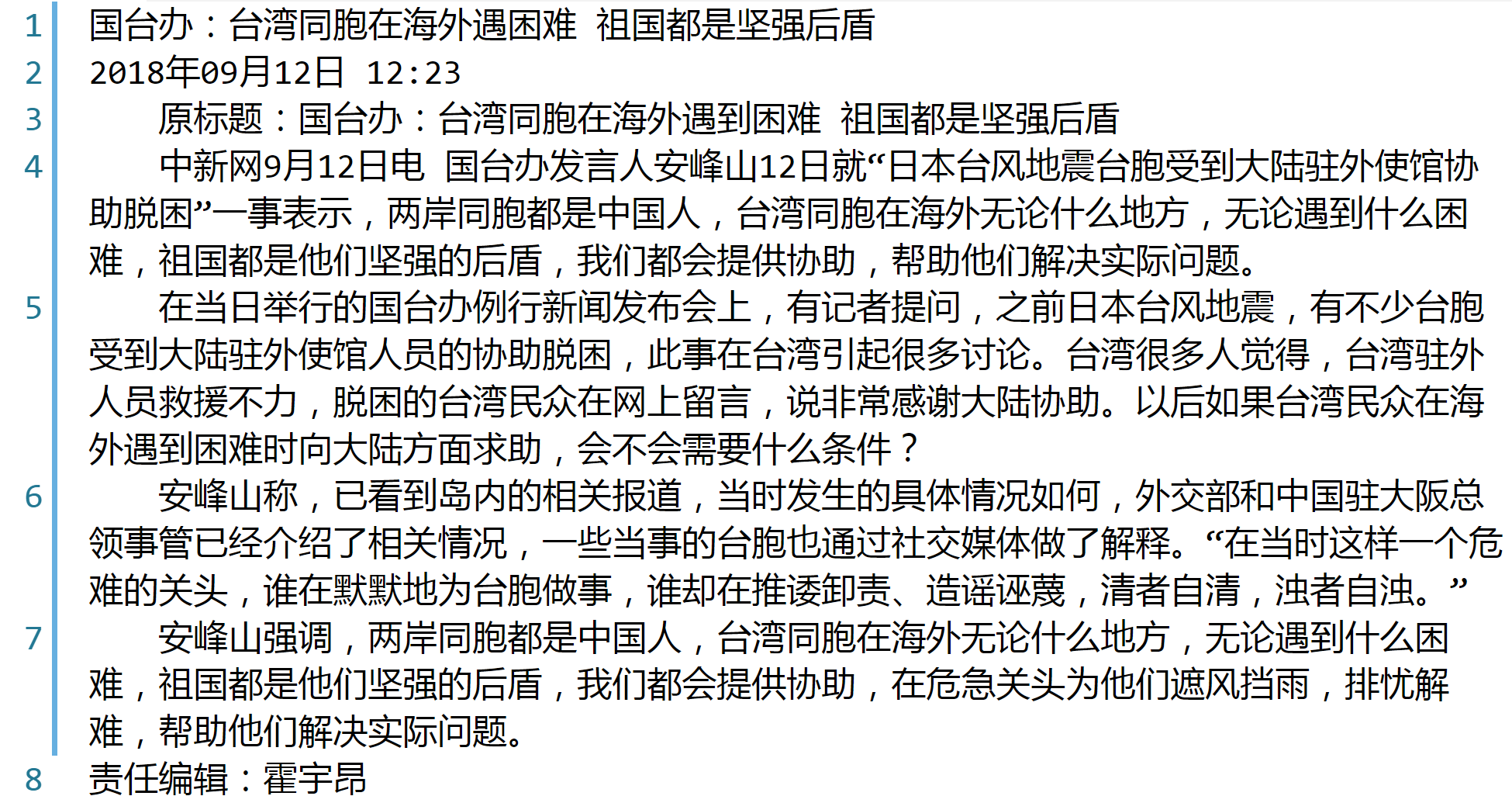
陈敬哲 2017011237

功能介绍：

1. 爬虫

爬虫的目标网站是<http://roll.news.sina.com.cn/news/gnxw/gdxw1/index_0.shtml>，即新浪的国内新闻列表，从这个列表中获取相应的新闻网站，然后用正则表达式对获取的信息进行处理，获得文章的标题、时间和正文，存储在文本文件中。

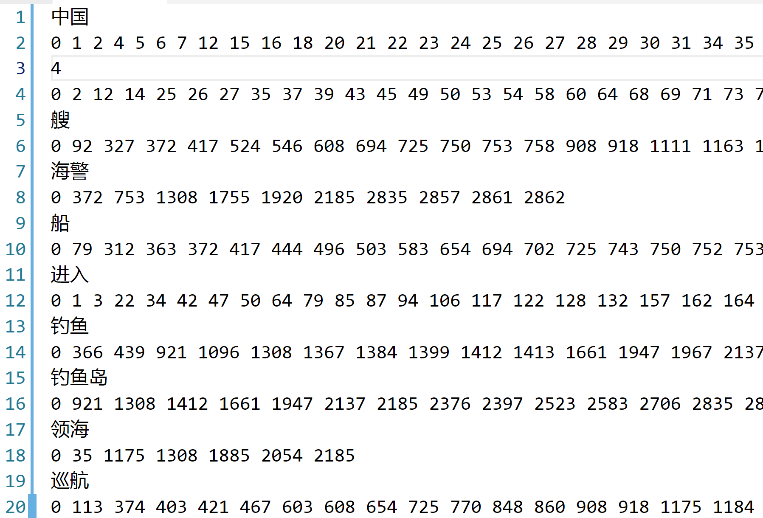
处理完后的文本示例：



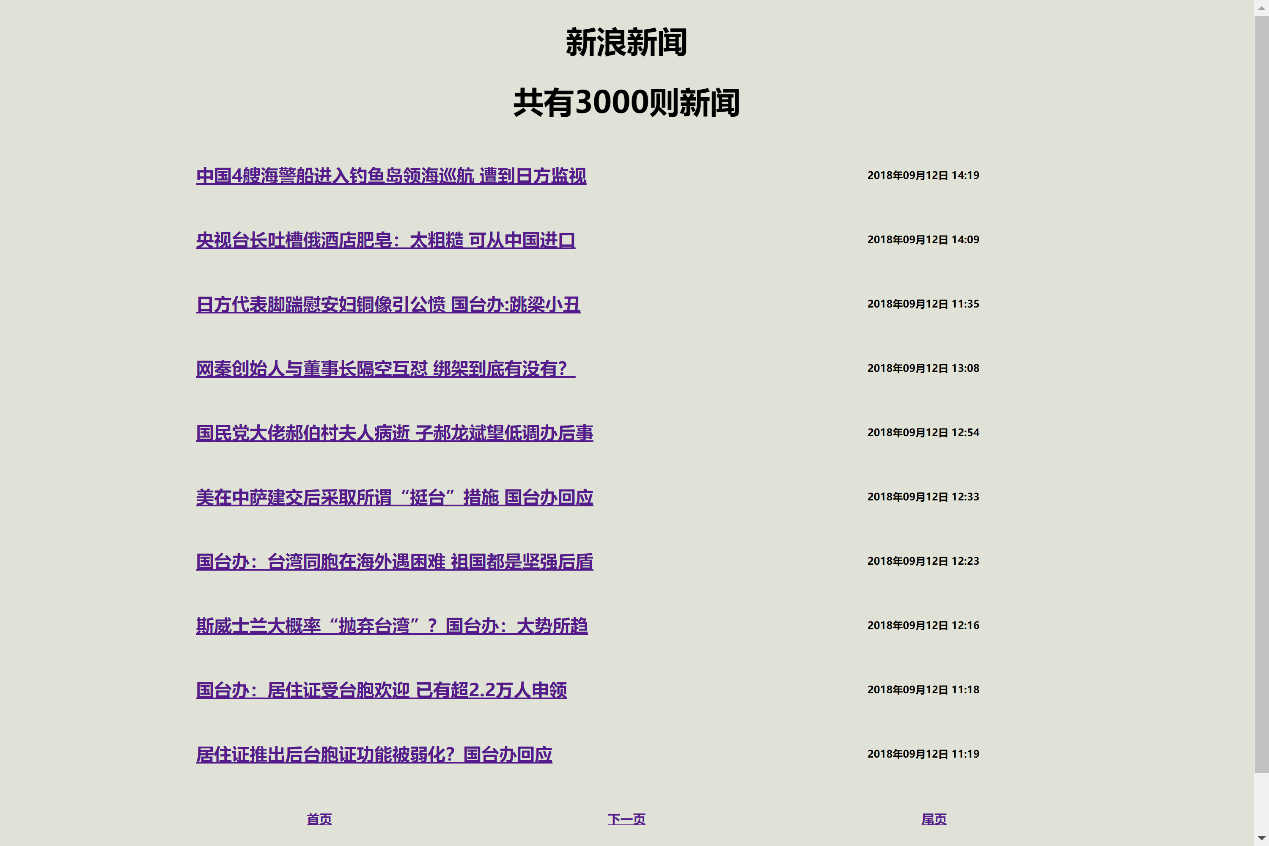
1. 新闻数据的预处理与索引

用jieba分词对每篇文章的标题和正文进行分词，然后对每个分词建立倒排列表索引表，存储在文本文件中。

存储索引表的文本文件：



1. 新闻展示

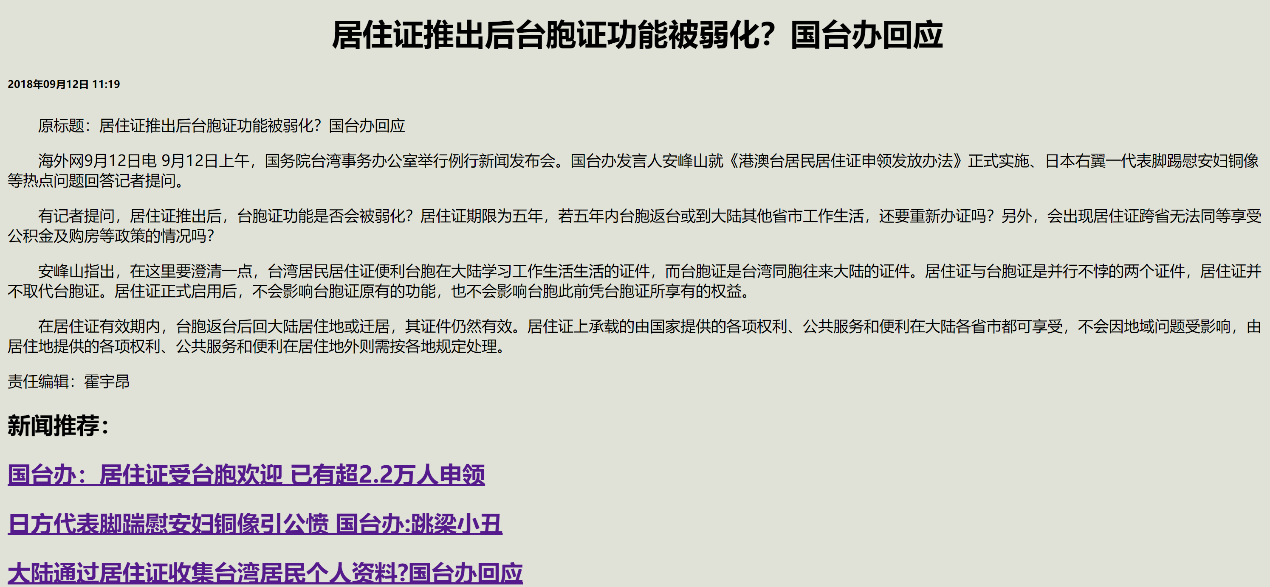


1. 新闻搜索





1. 新闻推荐：



总新闻量：3000

搜索耗时：一般在几毫秒和几十毫秒之间

查询算法：

如果是单关键词，在倒排列索引表中进行相关的查询，然后返回对应的id。

如果加上时间，就再加入一次筛选，选出对应时间的新闻返回。

如果是多关键词，统计每篇文章中所有关键词出现的次数，按照出现次数的多少进行排序，然后返回。

推荐算法：  
通过对各篇文章的分词，用tf-idf方法建立语料库，然后计算各篇文章的文本相似度，推荐文章是与这篇文章文本相似度排名前三高的文章。