

2018-2019 学年第一学期
《大数据分析与应用》 期末课程项目

天线宝宝搜索引擎说明文档

组 长：陈婉月

组 员：杨敏 范德宝

学 院：计算机学院

任课教师：张华平

1 系统功能说明

(1) 爬取新闻：数据量需要尽量多而广，并且对于爬取的数据要进行一定程度的处理，提取关键的、我们所需要的数据，对于一些半结构化数据和计算机无法识别的数据或者文本进行去噪、清洗、转码等，比如去除我们不需要的广告等信息，然后按照需求整理成结构化数据，最后将格式化的数据进行存储。

(2) 建立索引：建立索引是利用爬虫存储好的结构化数据进行字典和记录表的建立。要关注的问题是尽可能减小索引列表的存储空间，还有就是尽可能提高搜索效率。

(3) 推荐功能：需要我们对新闻进行较为准确的分析，推荐与选定新闻尽可能相似或者有一定关联的新闻。

(4) 检索功能：检索是将用户输入的信息进行处理，然后和我们建立的索引库进行匹配，按照相关性输出新闻。这要求我们准确地理解用户想要搜索的内容，按照相关度、时间、或者热度尽量提供给用户个性化的排序选择。

2.使用方法

1. 安装 python 3.6 环境
2. 安装 lxml html 解析器，命令为 `pip install lxml`
3. 安装 jieba 分词组件，命令为 `pip install jieba`
4. 安装 Flask Web 框架，命令为 `pip install Flask`
5. 进入 web 文件夹，运行 main.py 文件
6. 打开浏览器，访问 `http://127.0.0.1:5000/` 输入关键词开始测试
7. 如果想抓取最新新闻数据并构建索引，一键运行 `setup.py`，爬取新闻。

3 系统使用详解

3.1 搜索引擎首页

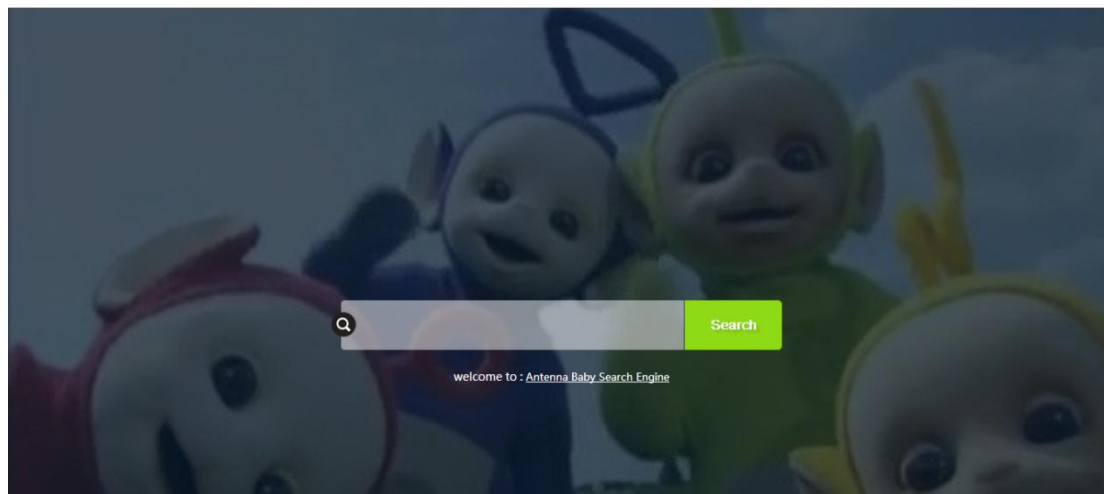


图 5.1 搜索引擎首页展示

图 5.1 是搜索引擎界面，用户在输入栏中输入需要查询的关键词，点击右侧的 Search 就能实现对关键词的搜索。

3.2 新闻搜索页



图 3.2 新闻搜索页展示

图 3.2 是新闻搜索页面，当用户在首页的搜索栏输入关键词搜索后，系统将

输入框中的关键词进行 jieba 分词，随后通过 BM25 算法和索引表将其与数据库中出現关键词的文档进行比较，计算出关联度和热度，最后根据关联度和热度排序来输出新闻信息。从图中实际的检索结果来看，该系统实现了较好的检索功能，能根据用户输入准备的查找出需要的新闻链接。

3.3 新闻展示页



图 3.3 新闻展示页

图 3.3 是我们打开上一步搜索到的新闻，可以看到新闻的标题，时间，链接，新闻的具体内容，还有左下角本项目实现的推荐阅读，我们从推荐阅读的新闻标题和检索的关键词可以看出系统推荐的文章和关键词之间具有较好的相似性。