

# 信息检索第三次作业报告

罗峻骁 计75

2017011364

## 摘要

本次作业搭建了一个简单的信息检索系统。系统支持：输入若干词语及其词性（可以只输入词语或词性），并可对输入的词语进行位置约束的限制，寻找相关的文档并显示结果。

本次作业的工作主要分为先后以下几项：

1. 下载人民日报数据作为系统检索文档集；
2. 使用 THULAC 工具，使用其提供 demo 的基础上对所有文档分词并进行词性标注，保存分词和词性标注结果；
3. 使用 Elasticsearch 作为系统搜索引擎实现，在 Ubuntu20.04 计算机上部署；
4. 使用 Elasticsearch 提供的相关接口创建索引，并将分词结果作为文档上传索引；
5. 开发检索系统的 web 应用，使能够通过浏览器使用检索系统。

## 分词与词性标注

使用 [THULAC](#) 工具（C++）版进行分词和词性标注。实现时为了便利，引用了[一个C++的json库](#)和[一个C++的进度条库](#)，并对 THULAC 的源代码进行了微小的修改使其可以输出用于 Elasticsearch Bulk API 的文件。文档经过分词和词性标注后，首先将词语与其对应词性以下划线 `_` 连接，随后再将这些结果用空格拼接为一个完成的句子。

## 部署 Elasticsearch

作业在 Ubuntu 20.04 操作系统上完成，参照[官方给出的基于 deb 包的方法](#)安装了 Elasticsearch 最新版（7.10.0），安装好后即通过如下命令创建并启动 Elasticsearch 服务：

```
sudo /bin/systemctl daemon-reload
sudo /bin/systemctl enable elasticsearch.service
sudo systemctl start elasticsearch.service
```

于是便有一个 Elasticsearch 结点在本机 9200 端口上工作。

## 创建索引

创建的索引只有 `content` 这一属性，类型为 `text`。为了支持词语或词性不同的匹配方法，使用了 Elasticsearch 的 `pattern capture token filter`，其作用为：输入一个 token，可根据相关规则输出多个 token，并且新的 token 和输入 token 的位置信息相同。于是，对于已经经过空格分隔得到若干 token（形如 `词语_词性`），可以通过编写相关规则除了使其输出自身外，还输出 `词语`，`词性` 两个新 token，由此便能实现通过词语或词性进行检索的目的。

由于 Elasticsearch 的 Bulk API 能上传的文件大小上限为 100M，而分词后的文件达 3G，无法直接使用 CLI，于是使用 Elasticsearch 的 python API 将文档集分批上传，分词后在 `cut` 目录下执行 `index.py` 脚本即可。

## 搜索

搜索的输入为若干单词及其词性（单词内容及词性都为可选，但至少有一），以及这些词语之间的一些位置约束。

对于输入的单词及相关词性，仍然是先将单词及词性用下划线连接，再将这些结果以空格分隔形成一个完整的字符串，最后使用 Elasticsearch 的 `match` 查询方法查询。值得注意的是，搜索过程中 token 不能和建索引时作相同的处理，否则本来较严的搜索条件会便宽。例如，如果查询 `吃_v` 时如果使用相同的处理方式，则会得到 `吃_v`，`吃`，`v` 三种 token，可能会错误地匹配上其它词语。

对于位置约束，以左相邻为例，只需将两个词语用空格分隔，再使用 `match_phrase` 方法即可。由于 `match_phrase` 方法会要求以相同顺序匹配到查询串中的所有词语，因此可以以此实现相邻的位置约束；若为不左邻，则将查询从 `must` 中放入 `must_not` 中；对于右邻，即为将两个 token 位置交换的左邻；对于相邻，即要么左邻要么右邻，只需要将左邻和右邻的查询用一个 `should` 包含即可。

## web 应用

基于 Node.js 的 express 框架搭建系统后端，后端处理前端传来的搜索请求，包含词语以及相关约束，生成查询并调用 Elasticsearch API，再将查询结果发送给前端。

前端页面主要包含输入词语、约束以及结果展示三个板块。用户可以添加或删除词语和约束，选择词语的词性，约束的类型，点击搜索按钮后前端将搜索请求发至后端，收到结果后分页展示。

应用页面示例如下：

Orange 搜索 主页

添加词语

添加约束

进行搜索

1	生产	动词	删除
2	部队	选择词性	删除
3	会议	选择词性	删除

1

2

相邻

删除

排名	内容	评分
1	第6版()专栏：欧洲伟大的社会主义明灯发扬自力更生的革命精神进一步开展部队生产活动阿人民军召开部队生产积极分子会议巴卢库同志在会上号召全军加强政治思想工作	30.37633
2	针对新形势下部队生产经营中暴露出来的“漏洞”，这个集团军专门召开会议，重申了部队生产经营的项目、范围，要求严格遵守地方现行的政策、法令，并制订出部队管理教育、预防犯罪等8条规定。	29.35349
3	据阿尔巴尼亚《战士报》二日报道，阿尔巴尼亚人民军一日在地拉那召开了部队生产积极分子会议。	28.399311
4	她坚持用毛主席的革命路线带领知识青年，用古田会议决议指导连队建设，使连队成为北京部队生产建设部队的先进集体。	27.756256
5	北京部队生产建设部队某部党委会	27.07439
6	广州部队生产建设部队某部知识青年何地章	26.870518
7	兰州部队生产	26.779823
8	据东北、华北、华东、中南、西北（仅包括陕西、甘肃、宁夏三省和新疆部队生产）和北京郊区等地所订的粮食增产计划的统计，计划增产的粮食共达一百四十五亿斤，超过全国农业生产会议原订增产计划数（原计划未包括广东、京郊和新疆部队生产在内）四十四点五亿斤，计增加百分之四十四强。	26.611689
9	到达生产部队垦区	26.53769
10	驻新疆的农业生产部队，在紧张的秋收中订出了增产节约计划，要求各生产部队做好今年的秋收工作，同时为明年的生产作好准备。	26.439987

1

2

3

4

5

31

Go