信息检索大作业 Report

计 62 胡致远

2016011260

一、项目框架

整个 Project 由索引构建、后端、前端三个部分组成。

1.1 索引构建

索引主要是以 Elasticsearch 为轮子构建的。对于已经完成分词和词性标注的文本，在数据库中保留两条数据，第一条数据包括词构成的数组和词性构成的数组，第二条数据由分词并进行词性标注后的词/词性对构成。例如：

原始数据：我/n 吃/v 苹果/n

Index1: text: [我，吃，苹果], attr: [n, v, n]

Index2: text: [我/n，吃/v，苹果/n]

Index1 主要是为了方便在不区分词性的情况下对词进行查询，Index2 主要是为了在有词性限制的情况下进行查询。

数据库采用的是 Sougou0002 中的前 100 万行数据，其实对于检索而言完全可以承担更大的数据集，但是构建索引太花时间了，因此只用了 100 万条数据。

1.2 后端

用户可以通过对后端进行 query 请求获得想要的结果。每一个 query 包含关键词(keywords)(可以多个，按空格隔开即可)，词性限制(attr)，距离限制(length)3 部分。

对于用户的每个 query，后端首先对 keywords 按照空格切开，但是对于每个 keyword 不会进行分词，然后在数据库中查询**包含所有 keyword** 的 doc，将这些 doc 中所有的除 keyword 以外的词收集在一起，作为备选集合(注：同一词若有不同词性则算所多个备选词)。然后将对备选集合进行多次过滤，最后计算每个词的得分并进行排序。

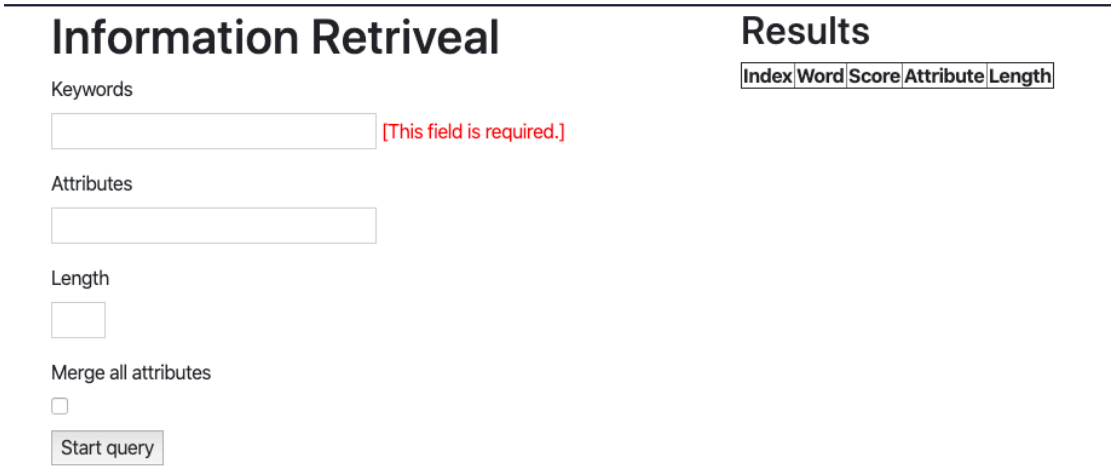
第一次过滤(基本过滤以及用户词性限制):首先会将词性为标点、数量词、叹词、助词等的词直接过滤掉，而只保留名词、动词、形容词等信息量较高的词。然后根据用户给出的词性限制选择符合用户需求的词(如果用户没有输入 attr 则保留所有词)。另外，如果用户在输入中要求将同一词的不同词性合并为同一个词(此时 attr 必定为空)，则将会将不同词性的同一个词合并为一个，并将其 attr 标记为 all。

第二次过滤(基于距离):为了方便，这里计算的距离是检索词与关键词在数组中 index 的差，即“词距离”而非“字距离”。如果用户输入了多个关键词，则只考虑索引词和第一个关键词的距离。这一次过滤仅保留符合距离限制的词。

打分与排序:这里采用了类似 tf-idf 的方法对索引词进行打分并按照分数排序。首先计算索引词和关键词同时出现的文档数，记为 tf，然后统计索引词在所有文档中出现的频率，记为 df。最终该索引词的得分就是 $s=tf*\log(D/df)$，其中 D 为文档总数。若 df 为 0 则直接置 s 为-1

1.3 前端

前端基于 flask 搭建，其界面如图所示：

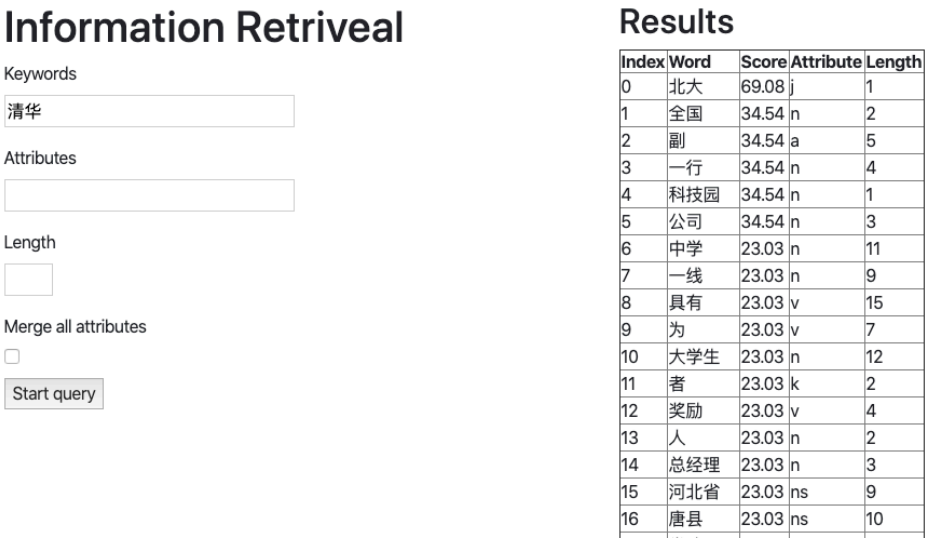


用户通过 Keywords, Attributes, Length 指定关键词，词性限制和距离限制，如果不填则默认没有限制。用户可以勾选 Merge all attributes 对索引结果不区分词性。

用户点击 Start query 之后，Results 中会列出索引结果。最多显示 50 条索引结果。

二、功能展示

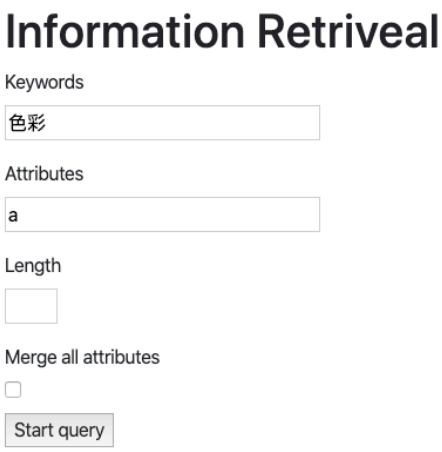
2.1 基本检索



Index	Word	Score	Attribute	Length
0	北大	69.08	j	1
1	全国	34.54	n	2
2	副	34.54	a	5
3	一行	34.54	n	4
4	科技园	34.54	n	1
5	公司	34.54	n	3
6	中学	23.03	n	11
7	一线	23.03	n	9
8	具有	23.03	v	15
9	为	23.03	v	7
10	大学生	23.03	n	12
11	者	23.03	k	2
12	奖励	23.03	v	4
13	人	23.03	n	2
14	总经理	23.03	n	3
15	河北省	23.03	ns	9
16	唐县	23.03	ns	10

可见能够实现基本的检索功能

2.2 词性限制

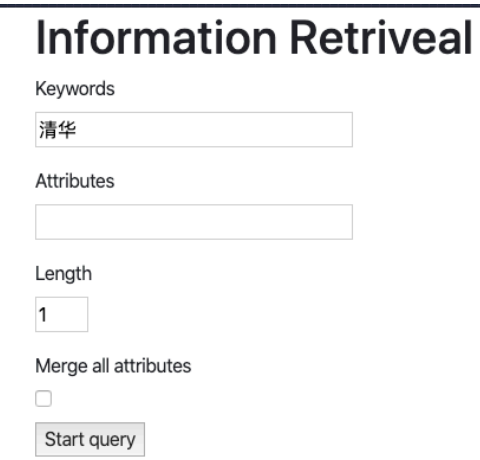


Results

Index	Word	Score	Attribute	Length
0	最新	34.54	a	10
1	精美	23.03	a	1
2	高	23.03	a	6
3	丰富	23.03	a	3
4	全	23.03	a	12
5	亮艳	13.82	a	1
6	诱人	11.51	a	40
7	漂亮	11.51	a	27
8	炫丽	11.51	a	11
9	美观	11.51	a	3
10	短	11.51	a	25
11	新颖	11.51	a	12
12	鲜艳	11.51	a	1
13	多	11.51	a	5

可见能够按照词性检索出想要的词

2.3 距离限制



Results

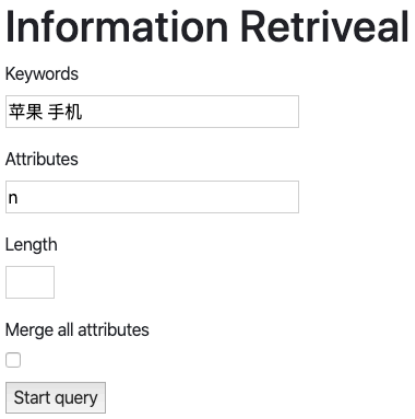
Index	Word	Score	Attribute	Length
0	北大	69.08	j	1
1	科技园	34.54	n	1
2	万博	12.72	nz	1
3	走红	11.51	v	1
4	考上	11.51	v	1
5	含	11.51	v	1
6	同方	11.51	nz	1
7	访问	11.51	v	1
8	浙江	11.51	ns	1
9	长三角	11.51	nz	1
10	机构	11.51	n	1

当距离限制为 1 时，能够找出和 “清华” 最常用的搭配词

2.4 多个关键词

Results

Index	Word	Score	Attribute	Length
0	香蕉	57.56	n	2
1	赌机	27.63	n	27
2	橙子	23.03	n	4
3	菠萝	23.03	n	4
4	媒体	23.03	n	1
5	电脑	23.03	n	17
6	图案	23.03	n	6
7	新奇	13.82	n	21
8	创报	13.82	n	39
9	阅报率	13.82	n	55
10	单板机	13.82	n	24
11	果泥	13.12	n	19
12	瓜类	13.12	n	18
13	自由时报	13.12	n	64
14	中国时报	12.72	n	50
15	评比会	12.43	n	10
16	苹果机	12.43	n	10



Results

Index	Word	Score	Attribute	Length
0	市场	57.56	n	2
1	消息	34.54	n	7
2	版	34.54	n	2
3	电影	34.54	n	3
4	媒体	23.03	n	3
5	大事	23.03	n	6
6	用户	23.03	n	4
7	行货	23.03	n	9
8	工具	23.03	n	11
9	厂商	23.03	n	10
10	量	23.03	n	16
11	drunknbass	13.82	n	8
12	苹果迷	13.12	n	5
13	舶来品	12.02	n	7

可见，在给定多个关键词的情况下，可以有效地消除歧义，挖掘出用户真正想表达的语义

三、改进内容

在后续的开发中，将持续对原有的项目进行改进，主要集中在以下两个方面：

1. 增大数据集规模。可以考虑将整个项目迁移到服务器上或者采用更高效的索引构建的方法。

2. 采用更多评分方式。目前只采用了 tf-idf 的策略，事实上还有 BM25，word2vec 等多种评分策略可用，后续将实现不同的策略。