

YSearch信息检索系统

殷翊文 2017011485

1 分词与词性标注

选用thulac c++版本对数据进行分词和词性标注。使用人民日报数据和一份搜狗数据共39,024,052条，每条为一个句子。

- 其中搜狗数据需要预处理，运行`./data/sogou.py`将原本的分词空格去除，并替换`<N>`为`0`。
- 下载thulac c++版本，运行`./thulac -t2s -input [inputfile.txt] -output [outputfile.txt]`，对数据进行分词和词性标注。生成的文件为`sogou_output.txt`, `rmrb_output.txt`，格式为每行一句，每句以空格分词，每个词后接'_'和相对应的词性。例如其中一行为：

`我_r 不悦_a 地_u 道_v 。_w`

2 索引构建

在本地运行elastic search，端口9200，运行`./data/add2es.py`对分词结果进行处理，连接elastic search，并批量加入数据，由es生成索引。

es中index名称为"sentences"，每条数据对应一句话，由三个字段组成，分别对应词汇、词性、词汇加词性。例如其中一条数据如下：

```
{  
    "content": "石头 与 石头 相击",  
    "part": "n c n v",  
    "mix": "石头_n 与_c 石头_n 相击_v"  
}
```

由于数据已经分好词，es中三个字段的`analyzer`即分词器都设置为`whitespace`。

3 搜索实现

`./flask`文件夹下为基于flask的demo项目。在该文件夹下运行`python ysearch.py`，在浏览器中打开`http://localhost:5000/`即可。

3.1 Demo说明

Demo包含首页和搜索结果页两个页面，预览如下。

The screenshot shows the YSearch homepage. At the top center is a large green logo "YSearch". Below it is a search bar with a placeholder and a green "搜索" button. Underneath the search bar is a section titled "输入格式" (Input Format) with a detailed explanation of search syntax, including examples like "我(r) 笑 (v) +a". Below this is a "词性提示" (Part-of-Speech Hint) section with a list of Chinese part-of-speech tags and their meanings. At the bottom of the page is a footer note: "@Supported by THULAC, ELASTICSEARCH & EVEN YIN".

YSearch

我 笑(v) +a	搜索																												
搜索耗时 117 ms, 搜索到 245 条结果																													
<table><thead><tr><th></th><th>相关性</th></tr></thead><tbody><tr><td>我笑, 我笑, 我笑, 笑, 笑!</td><td>12.897782</td></tr><tr><td>我冲它笑了笑, 它也冲我笑了笑。</td><td>12.004636</td></tr><tr><td>终于相遇了, 我笑, 我笑, 我开心地笑。</td><td>11.887672</td></tr><tr><td>我轻视的对他笑了笑, 他也对我笑了笑。</td><td>11.783374</td></tr><tr><td>我哈哈笑, 我哈哈哈!</td><td>11.782545</td></tr><tr><td>我笑我笑我继续笑, 没满意地看着雷洛愣神。</td><td>11.770879</td></tr><tr><td>我笑, 大家笑, 张含也非常含蓄地笑, 笑得我心里发毛。</td><td>11.675968</td></tr><tr><td>我说我笑什么, 我笑了吗?</td><td>11.623032</td></tr><tr><td>我看她冲我笑了笑。</td><td>11.611258</td></tr><tr><td>我看她对我笑了笑。</td><td>11.611258</td></tr><tr><td>我笑了笑说, 我知道。</td><td>11.611258</td></tr><tr><td>我笑了笑: “等我?”</td><td>11.611258</td></tr><tr><td>我笑了, 我真的笑了。</td><td>11.611258</td></tr></tbody></table>			相关性	我笑, 我笑, 我笑, 笑, 笑!	12.897782	我冲它笑了笑, 它也冲我笑了笑。	12.004636	终于相遇了, 我笑, 我笑, 我开心地笑。	11.887672	我轻视的对他笑了笑, 他也对我笑了笑。	11.783374	我哈哈笑, 我哈哈哈!	11.782545	我笑我笑我继续笑, 没满意地看着雷洛愣神。	11.770879	我笑, 大家笑, 张含也非常含蓄地笑, 笑得我心里发毛。	11.675968	我说我笑什么, 我笑了吗?	11.623032	我看她冲我笑了笑。	11.611258	我看她对我笑了笑。	11.611258	我笑了笑说, 我知道。	11.611258	我笑了笑: “等我?”	11.611258	我笑了, 我真的笑了。	11.611258
	相关性																												
我笑, 我笑, 我笑, 笑, 笑!	12.897782																												
我冲它笑了笑, 它也冲我笑了笑。	12.004636																												
终于相遇了, 我笑, 我笑, 我开心地笑。	11.887672																												
我轻视的对他笑了笑, 他也对我笑了笑。	11.783374																												
我哈哈笑, 我哈哈哈!	11.782545																												
我笑我笑我继续笑, 没满意地看着雷洛愣神。	11.770879																												
我笑, 大家笑, 张含也非常含蓄地笑, 笑得我心里发毛。	11.675968																												
我说我笑什么, 我笑了吗?	11.623032																												
我看她冲我笑了笑。	11.611258																												
我看她对我笑了笑。	11.611258																												
我笑了笑说, 我知道。	11.611258																												
我笑了笑: “等我?”	11.611258																												
我笑了, 我真的笑了。	11.611258																												

我笑了笑，对我这个妹妹我还真没办法。	11.054337
正在发呆，一个和我差不多大的年轻人走都我身边坐下，冲我笑了笑，我也连忙点头，冲他笑了笑。	11.033688
我立马大笑了起来，众人见我笑起，也跟着笑了起来。	11.031856
我笑了，我不知道自己笑的理由，更不知道笑的结果。	11.031856
我叫她弟妹，她还冲我笑了笑。	10.973171
我喝酒，他喝酒，我笑，他也笑。	10.973171
我笑一笑，“还有什么我可以帮忙的？	10.973171
我笑了笑“怎么办，我好像也一样诶！	10.973171
我笑了笑说：“我才不想呢！	10.973171
我笑了笑，“你在，我没有目标。	10.973171
我笑了笑，“因为我可是彻底的享乐主义者。	10.973171
我笑了笑，“或许我说的有些冒昧。	10.973171

[首页](#) [上一页](#) [1](#) [2](#) [3](#) [4](#) [5](#) [下一页](#) [末页](#)

@Supported by THULAC, ELASTICSEARCH & EVEN YIN

- 点击YSearch图标可回到首页
- 每页显示50条数据，支持分页功能
- 最大搜索结果为300条
- query格式错误会回到首页，并给出提示

3.2 搜索格式

支持的搜索格式：

- 词汇1(词性1) 词汇2(词性2) ... 词汇n(词性n) [+a]
- +a 为可选项，表示n个词必须按顺序相邻。
- 如果没有相邻限制，那么n个词必须输入词汇，可以不限制词性；
- 如果有相邻限制，那么不输入词汇，仅限制词性也是被允许的。特别地，仅搜索一串相邻的词性序列也是支持的，如下面的举例中第四条。

搜索举例：

- 我 笑
- 我 笑(v)
- 我(r) (v) +a
- (r) (v) (u) +a

3.3 具体实现

查询使用python的 `elasticsearch` 库实现。

1. 无相邻条件的搜索：

根据有无词性要求，分别在 `content` 和 `mix` 两个字段进行查询，使用es的bool查询，条件为 `should`，即只要满足其中一个条件就好。

例如query为“我 笑(v) 了”，那么查询的body为：

```
body = {
    "from": from_,
    "size": size_,
    "query": {
        "bool": {
            "should": [{"match": {"content": "我 了"}}, {"match": {"mix": "笑_v"} } ]
        }
    }
}
```

from和size用来分页。搜索平均耗时在60ms左右。

2. 有相邻条件的搜索：

首先按照无相邻条件的情况来搜索，es返回500条数据后，对数据进行筛选。筛选方法如下：

- 对每个句子，首先定位query中的一个词，即找出所有可能的该词出现的位置。这个词选择的优先度为词汇+词性>词汇>词性，为了得到尽可能少的位置。
- 对每个可能的位置，检查query中其它每个词汇或词性要求，是否符合相对位置，如果不符，则删去该位置。检查的优先度同上，为了快速过滤掉不可能的位置。
- 如果最后不存在这样的位置，则过滤掉该条数据。

最后返回符合要求的数据。搜索平均耗时在500ms左右。

四、问题与解决

1. Elastic search的写入问题。一开始为了测试，我在同一个程序里先写入再查询一批少量数据，结果发现查询结果为空。实际上数据库的写入也需要一点时间，两个进程是并行的，不能这么测试。写入后直接用postman查询即可。
2. 批量写入。我发现数据的写入非常慢，很少量的数据就需要耗费一秒的时间，所以必须用python elasticsearch库的helpers.bulk()，打包数据后批量写入。
3. 为了提高有词性限制的查找速度，我在数据库中新增了一个"mix"字段，同时记录词汇和词性信息，这样查找有词性限制的词汇时，只需要直接在该字段查找"词性_词汇"即可。
4. 相邻条件限制如果直接用es搜索，实现起来比较麻烦。由于数据要求量不大，在这里直接使用搜索后过滤的办法，对相邻词汇进行过滤。但这个算法的耗时较长。

五、demo演示

录屏见附件 `screenrecord.mov`。部分搜索结果截图如下：

YSearch

清华

搜索

搜索耗时 184 ms, 搜索到 300 条结果

	相关性
清华同方()最新清华同方笔记本报价，清华同方笔记本大全为您提供清华同方笔记本最新报价，包含清华同方笔记本报价、清华同方笔记本价格、清华同方笔记本品牌、清华同方笔记本大全、清华同方笔记本图片、清华同方笔记本查询、清华同方笔记本导购、清华同方笔记本评测、清华同方笔记本论坛、清华同方最新笔记本尽在泡泡笔记本网。	15.179211
清华紫光()最新清华紫光移动硬盘报价，清华紫光移动硬盘大全为您提供清华紫光移动硬盘最新报价，包含清华紫光移动硬盘报价、清华紫光移动硬盘价格、清华紫光移动硬盘品牌、清华紫光移动硬盘大全、清华紫光移动硬盘图片、清华紫光移动硬盘查询、清华紫光移动硬盘导购、清华紫光移动硬盘评测、清华紫光移动硬盘论坛、清华紫光最新移动硬盘尽在泡泡移动硬盘网。	14.78259
清华同方数码录音笔,清华同方数码录音笔大全,清华同方数码录音笔报价,清华同方数码录音笔图片,清华同方数码录音笔参数,清华同方数码录音笔经销商;	14.579893
清华紫光移动U盘清华紫光移动U盘专区提供清华紫光移动U盘报价、清华紫光移动U盘价格、清华紫光移动U盘新闻、清华紫光移动U盘行情、清华紫光移动U盘评测和清华紫光移动U盘经销商等综合信息。	14.289686
综观清华历史，新竹清华所呈现的气象，实与之前的清华南辕北辙。	14.1689
紧邻清华西门和清华附中~！	14.114695
可清华是全国的清华，不是北京的清华，它要面向全国。	14.032232
电脑之家清华同方清华同方液晶电视大全，包括各种型号清华同方清华同方液晶电视全面、详细的介绍信息。	13.991362
清华	13.954538
我清华的，什么时候来清华啊？	13.912225

YSearch

德国 德语 官方 语言 瑞士

搜索

搜索耗时 63 ms, 搜索到 300 条结果

	相关性
瑞士的官方语言是德语、法语和意大利语。	39.537537
瑞士是德国、法国和意大利人的后裔，它的官方语言是德语、法语还有意大利语。	38.204464
瑞士虽小，但有德语、法语、意大利语等3种官方语言。	35.642265
除德国外，欧洲国家奥地利、瑞士、卢森堡等也将德语当作国语或官方语言，法国、意大利、比利时、美国等国家的部分地区也使用德语，全世界使用德语的人口约有一亿。	33.14856
瑞士的语言主要是德语和法语，所在城市不同所用的语言不同。	29.2271
由于德语语言关难过，建议学生争取通过语言关之后再来德国。	28.12857
比如0年0月0日奥地利首相下令在波希米亚将捷克语和德语作为同等的内部官方语言结果受到整个帝国德国民族主义者的抨击。	27.968784
瑞士国家虽小，却有四种不同的官方语言。	25.69318
说瑞士人把单词的重音落在不同的地方，所以他听起来很费劲，但是瑞士德语比德国德语要柔和很多，大概是被法语和意大利语软化了。	25.391022
入幼儿园后开始学德、法两种官方语言，其中德语更为迫切，因为它是教堂宣教的语言。	25.317059
瑞士德语校对工具尚未安装。	24.503555
讲解语言是德语。	24.503428
瑞士官方语言五位数	24.000000

YSearch

干(v)

搜索

搜索耗时 57 ms, 搜索到 300 条结果

	相关性
我干，我干，我干，干，干。	10.489383
说干就干，干就干好。	10.389469
不能是干不干一个样，干多干少一个样，干好干坏一个样。	10.327392
要纠正干与不干一个样，干好干坏、干多干少一个样的思想。	10.271525
说干就干，要干就得干好。	10.221682
靠科技创业，吕道龙越干越想干，越干越敢干，越干越会干。	10.161587
改革，就要消除干与不干、干好干坏、干多干少一个样的现象。	10.161587
平时，社员干啥我干啥，早干晚干我都跟着干。	10.152863
到处出现了越干越想干，越干越敢干，越干越会干的动人情景。	10.107495
这样，改变了过去那种干与不干、干多干少、干好干坏一个样的现象。	10.107495
给自己干，其余的给别人干、单位干、国家干、人民干。	10.088133
防止干好干坏一个样，干与不干一个样。	10.059227
出工一窝蜂，下地磨洋工，反正干多干少一个样、干好干坏一个样、干与不干一个样。	10.053978

YSearch

干(a) 毛巾

搜索

搜索耗时 79 ms, 搜索到 300 条结果

	相关性
洗完后用干毛巾擦干。	24.687035
油漆打水磨后再用干毛巾擦干。	23.803055
我们只能用湿毛巾和干毛巾来进行洗漱。	23.55254
用干毛巾擦干刮水器和窗台上的水。	22.983175
记得用干毛巾把宝贝头上的汗擦干。	22.983175
2、采用干毛巾擦汗。	22.59912
拧干毛巾，我站起身。	22.59912
拿了条毛巾擦干身体。	22.59912
最后以干毛巾轻轻擦拭即可。	22.59912
杰克给绿娇娇递过干毛巾。	22.59912
用干净的毛巾擦干双手。	22.59912
用毛巾擦干餐具或水果。	22.59912
洗干净了，用毛巾擦干。	22.070862

YSearch

我 笑(v) 了 +a|

搜索

搜索耗时 275 ms, 搜索到 263 条结果

	相关性
我冲它笑了笑， 它也冲我笑了笑。	14.122153
我笑了， 我真的笑了。	13.913677
我轻视的对他笑了笑， 他也对我笑了笑。	13.845684
我笑了笑， 她也笑了。	13.726701
我笑了笑， "我回去了。	13.714271
我笑了笑， '我回去了。	13.714271
我笑了笑： "我已经完成了！	13.520499
我笑了笑说道！	13.478026
许丹冲我笑了笑。	13.478026
我看她冲我笑了笑。	13.461585
我看她对我笑了笑。	13.461585
我笑了笑说， 我知道。	13.461585
我笑了笑： "等我？	13.461585

YSearch

(r) (v) 了 +a|

搜索

搜索耗时 901 ms, 搜索到 300 条结果

	相关性
行了行了我知道了， 拜拜了您那！	3.6137395
还有， 谁谁谁下台了， 谁谁谁离婚了， 谁谁谁自杀了， 谁谁谁发财了， 谁谁谁叛逃了， 谁谁谁破产了， 谁谁谁把谁谁谁操练了， 谁谁谁宴请谁谁谁没请谁谁谁了。	3.611896
最耐人寻味：“0年， 我非典了、我隔离了、我萨达姆了、我神五了、我0了、我国荣了、我艳芳了、我朝核了、我终结者0了、我彩信了、我数码了、我毒鼠强了， 你呢？	3.605273
杀了他杀了他杀了他！	3.5891492
我杀人了我杀人了我杀人了。	3.5891492
我来了我来了我来了。	3.5891492
我累了， 我怕了， 我疯了， 我饿了。	3.5826154
我愁了因为你瘦了， 我瘦了因为你病了， 我笑了因为你壮了， 我有钱了因为把你卖了.....猪啊！	3.5796363
我冷了， 我热了， 我失去了， 我获得了。	3.57494
我们跟了你多久， 就忍了多久了。	3.5625138
我以为他长大了懂事了， 但是我错了， 他病了。	3.5592825
我问了问他们我进去了多久了。	3.5520234

YSearch

(a) (n) (v) +a

搜索

搜索耗时 939 ms, 搜索到 146 条结果

	相关性
直跑式楼梯显得严肃冷峻，悬弧式楼梯显得沉稳霸气，七字型楼梯精巧干练，旋转楼梯温馨浪漫，独龙骨楼梯开朗大方。	1.2869847
中小型企业合资多，大型跨国集团直接投资少；非生产性投资多，生产性投资比重小；加工装配型多，高新技术投入少；产品轻型化严重，新型产业少；用汇项目多，创汇项目少。	1.2865622
英红彩瓦采用挤压成型制造、外型美观生动、瓦表光滑平整、尺寸严格准确、密度均匀一致。	1.2765616
百纳宏顺奶业售大小黑白花奶牛是最大最全的宏顺奶业售大小黑白花奶牛网上贸易市场,提供宏顺奶业售大小黑白花奶牛产品信息,宏顺奶业售大小黑白花奶牛价格行情以及宏顺奶业售大小黑白花奶牛相关的供应信息,是商人们销售产品、拓展市场及网络推广的首选网站!	1.274933
每一个人都有对自己来说非常喜欢的动物，我也一样，比如小猪小猫小狗小鸡小鸭小兔小鹤小鸟小羊小鱼小马小牛小蛇小象小恐龙小老鼠小老虎小狮子小狐狸大猪大猫大狗大鸡大鸭大兔大鹅大鸟大羊大鱼大马大牛大蛇大象大恐龙大老鼠大老虎大狮子打狐狸等等等等等等，但是今天我要说的是我最好的好朋友——乔天。	1.274799
百纳全自动微型快速热水器是最大最全的全自动微型快速热水器网上贸易市场,提供全自动微型快速热水器产品信息,全自动微型快速热水器价格行情以及全自动微型快速热水器相关的供应信息,是商人们销售产品、拓展市场及网络推广的首选网站!	1.2726709
深层净白科技有效阻止肤色暗哑，美白效果能减少大小及深淡不一的色斑。	1.2674642
这款无线宽带路由器主要特性如下：高安全性：支持最新无线安全WPA认证方式，用户可完全控制无线网络安全。	1.2672952
小母鸡一般头小体圆脚干细，喙短颈短腿较矮；小公鸡一般头大体长脚干粗，喙长颈长腿也长。	1.2664641
蓝眼美女猫咪翻唱希望蓝眼美女猫咪翻唱希望。	1.2653621
确保无重特大治安事件、无重特大群体性事件、无重特大安全事故发牛，切实关心困难群众生活，让老百姓过一个欢乐、祥和、安定的新春佳	