YSearch信息检索系统 v2.0

殷翊文 2017011485

1 索引构建与运行方法

- 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。
- 选用thulac c++版本对数据进行分词和词性标注，标注后的格式为每句一行，以空格分词，每个词后接'_'和相对应的词性，如其中一行为：`石头_n 与_c 石头_n 相击_v`
- 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。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"
}
```

- `./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。

2 功能简介

支持的搜索格式

一句话概括功能：用`*`限定词性（词性可在首页查表），用`|`限定情感（n负向，p正向）。

分为两种搜索：

- **不限**，即没有位置要求，仅基于关键词和词性要求的搜索。只要出现输入词中的一个及以上即可。

  不限模式下可以规定词的词性，允许的格式包括：

  - `词`
  - `词*词性`
- **必须相邻**，即结果必须符合输入的位置顺序。输入的词（或要求）必须依次相邻出现在句子中。

  使用**必须相邻**模式，还可以限定词的情感，允许的格式包括：

  - `词`
  - `词*词性`
  - `*词性`
  - `*`（不限定词性，仅用于占位）

    - `*词性|情感`
    - `*|情感`（不限定词性，仅限定情感）

搜索举例：

- 不限
    - `我 笑 了`
    - `我 笑*v 了`
- 必须相邻
    - `开心 地 笑`
    - `开心*a 地 笑`
    - `* 地 笑`
    - `*a 地 笑`
    - `*a|p 地 笑`
    - `*|p 地 笑`

3 v2.0新增功能与实现

1. 增加了对词汇情感的分类，能够让用户搜索到一个词能够搭配哪些特定情感的词。最终效果很好（见4 效果演示），且搜索速度快，平均搜索总速度在200ms左右。

   具体实现如下：

    - 使用Chinese Open Wordnet汉语开放词网提供的中文WordNet，获取词汇对应的所有编号；

```
def load_net():
    lines = codecs.open("./app/static/cow-not-full.txt", "rb", "utf-8")
    net = dict()
    for line in lines:
        if line.startswith('#') or not line.strip():
            continue
        splited = line.strip().split("\t")
        if len(splited) == 3:
            (synset, lemma, status) = splited
        elif len(splited) == 2:
            (synset, lemma) = splited
            status = 'Y'
        if '+' in lemma:
            lemma = lemma.split('+')[0]
        if status in ['Y', 'O']:
            if not lemma.strip() in net.keys():
                net[lemma.strip()] = [synset.strip()]
            else:
                net[lemma.strip()].append(synset.strip())
    return net
```

- 使用nltk提供的wordnet和sentiwordnet库，获取所有编号对应的负向和正向情感值，取平均值；

```
def getSenti(word):
    # wordNet: dict{ word: [index, index, ...] }
    l = []
    if word in net.keys():
        l = net[word]

    if len(l) > 0:
        n = 0.0
        p = 0.0
        for index in l:
            info = wn.synset_from_pos_and_offset(str(index[-1:]),
int(index[:8]))
            info = swn.senti_synset(info.name())
            p += info.pos_score()
            n += info.neg_score()
        return n / len(l), p / len(l) # average sentiment
    else:
        return 0, 0
```

- 搜索过滤时，负向情感值>正向情感值的词判定为负向情感词，反之亦然，并根据用户要求进行过滤。

2. 修改了查询格式
    - 将手动输入"+a"改为选择标签限定“必须相邻”；且如果输入格式不符合”不限“的标准，可以自动判别为“必须相邻”。
    - 将词性限定的格式从`()`改为`*`，情感限定的格式设置成`|`。这是考虑到，用户在使用中文输入时，不用再频繁地切换中英文输入法来输入英文括号。`shift+8`和`shift+、`键可以直接输入`*`和`|`，而且如果使用搜狗拼音输入法，只需要再按一次shift就可以键入，不用在中英文输入之间来回切换。
3. 新增了占位查询，即`*`后不跟任何词性的查询，意在满足对>=1的位置距离的限定，如`北京 * * 烤鸭`可以查询到`北京的全聚德烤鸭`。而且，这种输入也更符合用户的认知习惯。

4 效果演示

占位搜索

YSearch

北京 * * 烤鸭 必须相邻 搜索

搜索耗时 53 ms，搜索到 8 条结果

	相关性
来北京吃北京烤鸭。	23.19911
其实，谁知道北京的全聚德烤鸭是怎样做的?	18.563635
提到吃鸭子，人们总是会想到北京的全聚德烤鸭、云南的宜良烤鸭，而今天在德宏，只要提到烤鸭，人们即刻想到沈鸭子，那色泽红艳，肉质细嫩，皮脆，味道醇厚，肥而不腻的特色烤鸭，真的让人垂涎三尺。	17.415642
火锅里他最爱吃的东西是鸭血，北京到处是烤鸭，却没有血。	16.146074
如北京的果木烤鸭，如武汉的煤火烤红薯，物体燃烧的烟气渗进食物，其味再怎么调制也不会纯正。	13.830326
1980年初，宋学良等人花公款让各科正、副科长到北京烤鸭店吃烤鸭，我不去，宋学良大发雷霆：“就是你张庆泰革命!	12.8096485
白薇说：“我很喜欢吃板鸭，我天生就喜欢吃鸭子，什么北京全聚德的烤鸭、便宜坊的挂炉焖鸭、还有什么咸水鸭，是鸭子我都喜欢吃。	11.929268
二、全国着名的烹调专家多招收一些学员，如北京仿膳、烤鸭，广东名菜，四川名菜等，使外宾在许多旅游胜地都可吃到有风味的佳肴。	11.767517

首页 1 末页

@Supported by THULAC, ELASTICSEARCH & EVEN YIN

情感搜索

- 不使用情感搜索，只搜索”笑“能搭配哪些形容词和副词：

YSearch

*a *u 笑 必须相邻 搜索

搜索耗时 185 ms，搜索到 300 条结果

	相关性
萧落雨笑了笑了，苦涩的笑。	9.903838
笑可以分很多种：最优美的笑是自然的笑，最诚挚的笑是发自内心的笑，最幸福的笑是甜蜜的笑，最高兴的笑是眉开眼笑，最可爱的笑是孩子露出两颗小虎牙抿着嘴笑。	9.897562
没有高级的笑，也没有低级的笑，笑就是笑。	9.857701
许淇灵始终笑着，温柔的笑，幸福的笑，轻轻的笑或是哈哈大笑。	9.8423605
笑到这就不由的笑了笑。	9.78843
姜昆总结笑之最说：“最愉快的笑是有说有笑；最高兴的笑是眉开眼笑；最自豪的笑是哈哈大笑；最没意思的笑是不笑装笑；最难为情的笑是捂面而笑；最幽默的笑是别人笑自己不笑；最呆痴的笑是莫名其妙跟着别人笑；最使人不高兴的笑是嘲笑；最使人捉摸不透的笑是假笑；最阴险的笑是皮笑肉不笑；最可怕的笑是奸笑；最难听的笑是狂笑；最残酷的笑是冷笑；最美丽的笑是心灵和眼睛一齐笑!	9.786034
我笑，大笑，夸张的笑，没有形象的笑。	9.749012
我的笑大致分为三种：喜悦的笑、甜蜜的笑和苦涩的笑。	9.719338
这一笑,是激动的笑,会心的笑。	9.651005
无极嘿嘿的笑了笑!	9.649936
易萧尴尬的笑了笑。	9.649936
易青微微的笑了笑。	9.649936
昌凡欣慰的笑了笑。	9.649936
永硕淡漠地笑了笑。	9.649936
简单憨憨的笑了笑。	9.649936

- 用户想搜索”不好的笑“，即情感为负向的笑。可以看到结果有苦涩、低级、虚假、悲哀等等，效果很好。

YSearch

*a|n *u 笑 必须相邻 搜索

搜索耗时 461 ms，搜索到 41 条结果

	相关性
萧落雨笑了笑了，苦涩的笑。	9.903838
没有高级的笑，也没有低级的笑，笑就是笑。	9.857701
姜昆总结笑之最说：“最愉快的笑是有说有笑；最高兴的笑是眉开眼笑；最自豪的笑是哈哈大笑；最没意思的笑是不笑装笑；最难为情的笑是捂面而笑；最幽默的笑是别人笑自己不笑；最呆痴的笑是莫名其妙跟着别人笑；最使人不高兴的笑是嘲笑；最使人捉摸不透的笑是假笑；最阴险的笑是皮笑肉不笑；最可怕的笑是奸笑；最难听的笑是狂笑；最残酷的笑是冷笑；最美丽的笑是心灵和眼睛一齐笑！	9.786034
我的笑大致分为三种：喜悦的笑、甜蜜的笑和苦涩的笑。	9.719338
曲瀚然不意外的笑了笑。	9.507423
我虚假的笑了笑。	9.481873
我无言的笑了笑。	9.481873
木木淫荡的笑了笑。	9.481873
札札悲哀的笑了笑。	9.481873
泪月残忍的笑了笑。	9.481873
束草轻贱的笑了笑。	9.481873
杨锐淫荡的笑了笑。	9.481873
水心紧张地笑了笑。	9.481873
梵若苦涩的笑了笑。	9.481873
皇后虚弱地笑了笑。	9.481873

- 用户想搜索”好的笑“，即情感为正向的笑。可以看到结果有优美、自然、幸福、温柔等等。

YSearch

*a|p *u 笑 必须相邻 搜索

搜索耗时 169 ms，搜索到 114 条结果

	相关性
笑可以分很多种：最优美的笑是自然的笑，最诚挚的笑是发自内心的笑，最幸福的笑是甜蜜的笑，最高兴的笑是眉开眼笑，最可爱的笑是孩子露出两颗小虎牙抿着嘴笑。	9.897562
没有高级的笑，也没有低级的笑，笑就是笑。	9.857701
许淇灵始终笑着，温柔的笑，幸福的笑，轻轻的笑或是哈哈大笑。	9.8423605
姜昆总结笑之最说：“最愉快的笑是有说有笑；最高兴的笑是眉开眼笑；最自豪的笑是哈哈大笑；最没意思的笑是不笑装笑；最难为情的笑是捂面而笑；最幽默的笑是别人笑自己不笑；最呆痴的笑是莫名其妙跟着别人笑；最使人不高兴的笑是嘲笑；最使人捉摸不透的笑是假笑；最阴险的笑是皮笑肉不笑；最可怕的笑是奸笑；最难听的笑是狂笑；最残酷的笑是冷笑；最美丽的笑是心灵和眼睛一齐笑！	9.786034
我笑，大笑，夸张的笑，没有形象的笑。	9.749012
我的笑大致分为三种：喜悦的笑、甜蜜的笑和苦涩的笑。	9.719338
这一笑,是激动的笑,会心的笑。	9.651005
易萧尴尬的笑了笑。	9.649936
邪气的笑、狡黠的笑、机智的笑、诡谲的笑、得意的笑、天真的笑……都那么极可爱而灿烂，让人忍俊不禁的同时怦然心动。	9.581715
那是笑--神的笑，美的笑；	9.568232
无极长老自信的笑了笑。	9.507423
美珍姐宽容地笑了笑。	9.507423
碎风影尴尬地笑了笑。	9.507423
我笑，大家笑，张含也非常含蓄地笑，笑得我心里发毛。	9.494794
战狼兴奋地笑了笑。	9.481873

- 不使用情感搜索，只想搜索对北京的形容词

YSearch

*a 的 北京 | 必须相邻 | 搜索

搜索耗时 245 ms，搜索到 37 条结果

	相关性
北京呵，伟大的北京！	9.169067
亲爱的北京，久违的北京！	9.168739
明星身后的北京，是古朴的北京，现代的北京，磅礴的北京，精致的北京，庄严的北京，时尚的北京，是各式各样的北京，那就来看看各个明星在这首mv中所站的各处代表风景及代表文化。	9.16719
资讯，汇聚最新的北京婚庆价格动向、提供全面的北京婚庆新闻、北京婚庆导购、北京婚庆查询、北京婚庆车型参数、北京婚庆图片等北京婚庆信息。	9.139065
户口所在地北京~地道的北京人！	8.877289
资讯，汇聚最新的北京除皱医院价格动向、提供全面的北京除皱医院新闻、北京除皱医院导购、北京除皱医院查询、北京除皱医院车型参数、北京除皱医院图片等北京除皱医院信息。	8.866469
资讯，汇聚最新的北京最好的整形医院价格动向、提供全面的北京最好的整形医院新闻、北京最好的整形医院导购、北京最好的整形医院查询、北京最好的整形医院车型参数、北京最好的整形医院图片等北京最好的整形医院信息。	8.8387
泡泡电影库提供最新最全的北京童话,北京童话下载,北京童话剧情,北京童话海报,北京童话免费下载等,更多精彩尽在泡泡电影网站。	8.619115
届时北京青年报小红帽将率先组织4000名北京人游北京，让这些普通的北京人在遍赏了北京时代美景之后推选出心中希望入选MTV之中的美丽景致。	8.609531
欢乐的北京	8.605607
伟大的北京	8.605607
本北京自考公司凭借卓越的北京自考人才实力为您提供值得信赖的北京自考服务。	8.501001
百纳北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收是最大最全的北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收网上贸易市场,提供北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收产品信息,北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收价格行情以及北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收相关的供应信息,是商人们销售产品、拓展市场及网络推广的首选网站！	8.45348

- 用户想搜索对北京的正向的评价

YSearch

*a|p 的 北京 | 必须相邻 | 搜索

搜索耗时 205 ms，搜索到 21 条结果

	相关性
明星身后的北京，是古朴的北京，现代的北京，磅礴的北京，精致的北京，庄严的北京，时尚的北京，是各式各样的北京，那就来看看各个明星在这首mv中所站的各处代表风景及代表文化。	9.16719
资讯，汇聚最新的北京婚庆价格动向、提供全面的北京婚庆新闻、北京婚庆导购、北京婚庆查询、北京婚庆车型参数、北京婚庆图片等北京婚庆信息。	9.139065
资讯，汇聚最新的北京除皱医院价格动向、提供全面的北京除皱医院新闻、北京除皱医院导购、北京除皱医院查询、北京除皱医院车型参数、北京除皱医院图片等北京除皱医院信息。	8.866469
资讯，汇聚最新的北京最好的整形医院价格动向、提供全面的北京最好的整形医院新闻、北京最好的整形医院导购、北京最好的整形医院查询、北京最好的整形医院车型参数、北京最好的整形医院图片等北京最好的整形医院信息。	8.8387
泡泡电影库提供最新最全的北京童话,北京童话下载,北京童话剧情,北京童话海报,北京童话免费下载等,更多精彩尽在泡泡电影网站。	8.619115
欢乐的北京	8.605607
本北京自考公司凭借卓越的北京自考人才实力为您提供值得信赖的北京自考服务。	8.501001
百纳北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收是最大最全的北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收网上贸易市场,提供北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收产品信息,北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收价格行情以及北京双盈二手库存积压回收电脑回收 / 北京二手电脑回收相关的供应信息,是商人们销售产品、拓展市场及网络推广的首选网站！	8.45348
第11版(美术世界)专栏：多彩的北京——北京画院作品展为庆祝中华人民共和国建国五十周年、北京和平解放五十周年，北京画院于5月11日至16日在中国美术馆举办《多彩的北京———北京画院画展》。	8.439059
新京报》是对北京之外影响力最大的北京报纸。	8.416867
作为北京市城市规划顾问，刘太格多次对北京的发展提出建议，他认为北京应该实现新旧城市分开发展，在古城内外建设两个辉煌的北京———个是历史的北京、一个是现代的北京。	8.331191
未来几年，北京可以投巨资建设北京，可以组织起充足的志愿者大军，可以使北京拥有更高的科技、环保水平，但如果没有北京人优良的整体素质这个基础，再好的北京也只能徒有其表。	8.319818

- 用户想搜索对北京的负向的评价，结果较少，”普通“和”通用“都被归在了负向情感的词汇集合里。

YSearch

*a|n 的 北京 | 必须相邻 | 搜索

搜索耗时 223 ms，搜索到 2 条结果

	相关性
届时北京青年报小红帽将率先组织4000名北京人游北京，让这些普通的北京人在遍赏了北京时代美景之后推选出心中希望入选MTV之中的美丽景致。	8.609531
不过，“北京时间”并非是真正的北京的地方时间，因为北京的地理位置是一百一十六度二十二分，因而现在全国通用的北京时间和真正的北京的地方时间还相差十四分二十四秒。	8.393039

首页 1 末页