

NLP算法入门系列：不同领域文章的热词提取

原创 IT可达鸭 IT可达鸭 5月9日



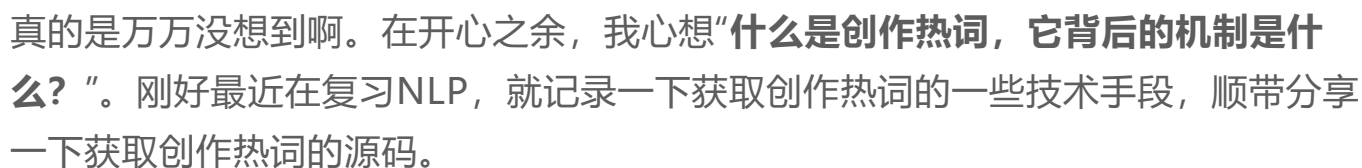
点击上方蓝字关注我们!!!
FOLLOW US


文/IT可达鸭

图/IT可达鸭、网络

• 前言

昨天突然收到头条官方的私信，让我很意外。说我获得近一周【创作热词】的体验特权，本周科技领域热词：“阿里、百度、疫情、日本、马化腾、淘宝、微信、华为手机、社交软件、智能家居。”



头条号 

举报 ...

10:43

头条号

亲爱的IT可达鸭，恭喜获得【创作热词】体验特权！发布热词文章，将更大概率获得高推荐和高阅读，进而创造爆文！期待你成为下一个爆文创作者！

本周科技领域用户喜爱热词：阿里、百度、疫情、日本、马化腾、淘宝、微信、华为手机、社交软件、智能家居

 @IT可达鸭

• 热词详解

热词指的是，在同一个领域中，最近一周的文章出现频率较高而且非无用的词语，在一定程度上代表了这同一领域在这一周的所有文章的焦点所在，可以视为一种关键词。

在自然语言处理NLP中，热词提取就是高频词提取，就是NLP中TF (Term Frequency) 策略。主要有以下两个干扰项：1. 标点符号（一般标点符号无任何价值，需要去除）；2. 停用词（例如“的”、“了”、“是”等常用无任何意义，也需要剔除）。

在进行热词提取之前，首先要对文章进行分词。规则匹配与统计两种切词算法，其中基于规则：《NLP算法入门系列：中文切词算法，基于规则匹配》，基于统计：《NLP算法入门系列：隐含马尔可夫链(HMM)模型的简单介绍》。

实际项目使用一个比较著名的python NLP开源库，**jieba**包。这里只使用到了里面的分词操作，词频统计是自己单独采用python写法写的，大家可以参考。

• 环境配置

python版本：3.6.0

编辑器：pycharm

语料准备：这里提供部分医疗领域的新闻稿，近两千份文档，采用GBK编码

项目所需要的环境安装包：


jieba: 主要用于切词，切词算法内置，原理是规则切词+统计切词

glob: 主要用于获取语料文件路径

• 具体实现

第一步，进行数据的读取，该函数用于加载指定路径下的数据。


```
7      """ 数据读取 """
8      def get_content(path):
9          content = ''
10         with open(path, 'r', encoding='gbk', errors='ignore') as rf:
11             for line in rf:
12                 content += line.strip()
13         return content
```



第二步：定义高频词统计函数，输入是一个词的数组。

其中，`sorted()` 对字典`tf_dic`进行排序，`key`参数指定一个`lambda`表达式，指定以字典的值大小进行排序，`reverse=True`表示倒序（字频从大到小进行排序），`[:topK]`表示从列表中取前`topK`个元素。

```
14
15      """ 定义高频词统计函数 """
16      def get_tf(words, topK=10):
17          tf_dic = dict()
18          for w in words:
19              tf_dic[w] = tf_dic.get(w, 0) + 1
20          return sorted(tf_dic.items(), key=lambda x: x[1], reverse=True)
21
```



第三步：加载数据，随机选择一个文本进行分词和词频统计，数据存放在`data/news/`下的所有`txt`文件。

```

23 def demo_topk():
24     files = glob.glob('.*\\data\\news\\*.txt')
25     corpus = [get_content(file) for file in files]
26     sample = choice(corpus)
27     split_words = list(jieba.cut(sample))
28
29     print(' 样本之一: ' + sample + "\n\n")
30     print(' 样本分词效果: ' + ' / '.join(split_words) + "\n\n")
31     print(' 样本top(10): ' + str(get_tf(split_words, 10)) + "\n\n")
32

```



运行结果如下。

```

样本分词效果: 国家/ 食品/ 药品监督管理局/ 南方/ 医药/ 经济/ 研究所/ 主办/ 的/ 首届/ "/ 中国/ 制药/ 工业/ 百强/ 年会/ 暨/ 第三/ 终端/ 高峰论坛/ "/
样本top(10): [('的', 8), ('医药', 6), ('.', 6), (' ', 6), ('、', 5), ('2005', 5), ('年', 5), ('企业', 5), ('

```



通过上面的结果，我们可以发现，诸如“的”、句号、逗号等词占据很高的位置，而这类词对把控文章的焦点与热点并无关系。我们需要的是，真正意义上的热词。


常用的方法，是自定义一个停用词词典，当遇到这些词时，自动过滤掉。

第四步，接入停用词，同时改进热词提取算法，每次分词后，对词语进行过滤。把过滤后的词做词频统计。


```

34  """ 获取停用词 """
35  def get_stopwords(path):
36      with open(path, encoding='utf-8') as rf:
37          return [line.strip() for line in rf]
38
39  def demo_topk_with_stopwords():
40
41      files = glob.glob('..\\data\\news\\*.txt')
42      corpus = [get_content(file) for file in files]
43      sample = choice(corpus)
44      split_words = list(jieba.cut(sample))
45
46      stopwords = get_stopwords('..\\data\\stop_words.utf8')
47      split_words_stopwords = [word for word in split_words if word not in stopwords]
48
49      print('样本之一: ' + sample + '\n\n')
50      print('样本分词效果: ' + ' / '.join(split_words) + '\n\n')
51      print('样本top(10): ' + str(get_tf(split_words_stopwords, 10)) + '\n\n')
52

```



但是，这个时候，还存在一点点地方可以优化。就是网络新词，通过手工录入 user_dict.utf8 文件，作为新词，这样 jieba 分词器就会识别它为一个词。

例如：“大数据工程师”，如果没有录入新词库，就会出现“大数据 / 工程师”两个词，录入词典后，分词器自动将其识别为一个词，减少误判的概率。关于新词库，对不同的业务，单独维护一个词库即可。

最后，完善的热词提取算法。

```

53  def demo_topk_with_userdict():
54      jieba.load_userdict('..\\data\\user_dict.utf8')
55      files = glob.glob('..\\data\\news\\*.txt')
56      corpus = [get_content(file) for file in files]
57      sample = choice(corpus)
58      split_words = list(jieba.cut(sample))
59
60      stopwords = get_stopwords('..\\data\\stop_words.utf8')
61      split_words_stopwords = [word for word in split_words if word not in stopwords]
62
63      print('样本之一: ' + sample + '\n\n')
64      print('样本分词效果: ' + ' / '.join(split_words) + '\n\n')
65      print('样本top(10): ' + str(get_tf(split_words_stopwords, 10)) + '\n\n')
66

```



本文介绍了文章热词提取的相关技术，讲解了在实际项目中如何使用jieba分词器。作为NLP的入门学习算法，提取高频词是分词算法最常应用的一个业务场景。算是一个非常基础的功能模块，当然，真正的热词提取还有很多地方可以做，除了高频之外，还有逆词频，TF-IDF等算法来提取文章热词。

如果有疑问想获取源码，可以关注后，在后台私信我，回复：**python提取热词**。我把源码发你。持续关注"**IT可达鸭**"，每天除了分享有趣Python源码，还会介绍NLP算法。最后，感谢大家的阅读，祝大家工作生活愉快！

长按二维码
获取更多精彩

IT可达鸭

