人工智能之nlp

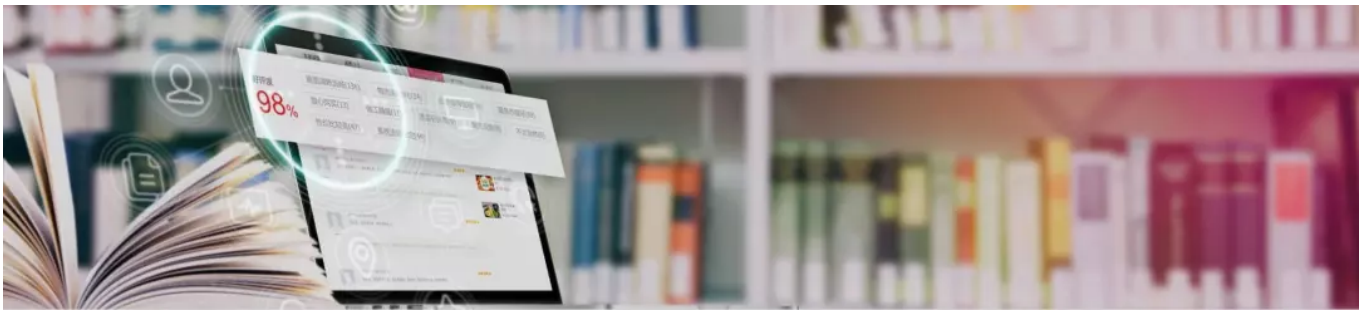
原创 yhJavaWeb 艳学网 2019-01-21

人工智能之nlp

最近，在写自媒体文章，词穷的我写不出一篇优秀的原创文章，对语言的能力掌控只有ctrl加c。 听别人说，人工智能可以自动写代码，那自动写文章也可以吧。写了2年博客的我还在坚持原创， 但是某些操作需要专业的能力才可以做到，如关键字提取，读完一篇文章，能快速准确提取出本文的重点关键字吗？ 高中以前应该还可以，我记得以前读书考试的时候有这么一题，请归纳本文的主要思想， 看完文章之后，一般和作者的想法产生了矛盾。

百度nlp：http://ai.baidu.com/tech/nlp?hmsr=huabiao&hmpl=%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86&hmcu=&hmkw=banner

有个想法，就是输入一个词，例如人工智能，然后就可以输出一篇伪原创热文。 或者另一个想法，输入一个关键字，可以生成一个网页，过年之前给自己一个小目标，做一个一键生成网站的网站。



关键词提取

```
List<String> keyWordList = NaturalLanguageApi.getKeyword(content);
                mav.addObject("keyWord1", keyWordList.get(0));
                mav.addObject("keyWord2", keyWordList.get(1));
                mav.addObject("keyWord3", keyWordList.get(2));
                mav.addObject("keyWord4", keyWordList.get(3));
                mav.addObject("keyWord5", keyWordList.get(4));
```

提取5个关键词

常见的关键词提取方法有：TF-IDF关键词提取方法、Topic-model关键词提取方法和RAKE关键词提取。

TF-IDF：

使用TF-IDF提取关键词的方法十分好理解，TF衡量了一个词在文档中出现的频率，一个文档中多次出现的词总是有一定的特殊意义，但是并不是所有多次出现的词就都是有意义的，如果一个词在所有的文档中都多次出现，那么这个词就没有什么价值了。

TF-IDF就很好地衡量了这些因素：TF= （词在文档中出现的次数）/ （文章总词数），IDF= log(语料库中文档综述/(包含该词的文档数+1))

TF-IDF= TF* IDF

TF-IDF值越大，则这个词成为一个关键词的概率就越大。

Topic-model：

使用主题模型提取关键词的关键思想是认为文章是由主题组成的，而文章中的词是以一定概率从主题中选取的，即文章与词之间存在一个主题集合。不同的主题下，词出现的概率分布是不同的。 根据LDA主题模型的学习可以获取文档的主题词集合。

RAKE关键词提取：

RAKE(Rapid Automatic Keyword Extraction)算法的原作者是Alyona Medelyan，RAKE的更新版本就是她完成的，muai indexer也是她的杰作，她的GitHub上有很多关键字提取的项目。

RAKE提取的关键词并不是单一的单词，有可能是一个短语。

每个短语的得分有组成短语的词累加得到，而词的得分与词的度与词频有关：score = degree / freq

其中，当与一个词共现的词越多，则该词的度就越大。

摘要提取

TextRank是自然语言处理领域一种比较常见的关键词提取算法，可用于提取关键词、短语和自动生成文本摘要。TextRank是由PageRank算法改进过来的，所以有大量借鉴PageRank的思想，其处理文本数据的过程主要包括以下几个步骤：

（1）首先，将原文本拆分为句子，在每个句子中过滤掉停用词（可以不选），并只保留指定词性的单词，由此可以得到句子和单词的集合。

（2）每个单词作为PageRank中的一个节点。设窗口大小为k，假设一个句子所组成的单词可以表示为w1,w2,w3,..., wn.

则w1,w2, ..., wk、w2,w3,...,wk+1、w3,w4,...,wk+2等都是一个窗口，在一个窗口内任意两个单词之间存在一条无向无权的边。

（3）基于上面的节点和边构成图，可以据此计算出每个节点的重要性。最重要的若干单词可以作为区分文本类别和主题的关键词。

ai仿写

NLP的四个经典的“AI完成”问题：问答，重播，摘要等。如果只解决其中一个问题，其他三个解决。问题和答案是让机器人非常开放，回答你提到的各种问题，就像真人一样。复述是让机器用另一种方式表达它，摘要是告诉你一篇长篇文章，并让你写出一篇100字的摘要很难做到这一点，翻译也非常困难，英文思维模式和中文思维模式都被转换。中间会出现很多复杂的问题。

内容重写的生态非常复杂。我们无法用简单的自然语言处理技术解决所有问题。过去，自然语言处理相对简单。甚至提出了一个单词列表来解决所有问题。随着电子商务的生态。扩张需要非常复杂的技术。因此，我们需要一个完整的高性能自然语言处理技术。高性能体现在算法的准确性和执行效率上。

git开源：https://gitee.com/490647751/yanhui-sdk

群名称：艳听AI工具交流群
群　号：926426148

加群926426148一起学习AI

如需获取本文源码，请加QQ490647751回复“开通VIP-人工智能之nlp”