工作记录5之R中的文本挖掘包

2016-09-06 15:54:06

Rwordseg包——文本分词(建议数据量<1G)
Rwordseg分词原理以及功能详情 Rwordseg 是一个R环境下的中文分词工具，使用 rJava 调用 Java 分词工具 Ansj。 Ansj 也是一个开源的 Java 中文分词工具，基于中科院的 ictclas 中文分词算法， 采用隐马尔科夫模型（Hidden Markov Model, HMM）。作者孙健重写了一个Java版本， 并且全部开源，使得 Ansi 可用于人名识别、地名识别、组织机构名识别、多级词性标注、 关键词提取、指纹提取等领域，支持行业词典、 用户自定义词典。 1、分词原理 n-Gram+CRF+HMM的中文分词的java实现. 分词速度达到每秒钟大约200万字左右（mac air下测试），准确率能达到96%以上 目前实现了.中文分词. 中文姓名识别 . 用户自定义词典,关键字提取，自动摘要，关键字标记等功能 可以应用到自然语言处理等方面,适用于对分词效果要求高的各种项目.（官方说明文档来源：http://pan.baidu.com/s/1sj5Edjf） 该算法实现分词有以下几个步骤： 1、 全切分，原子切分； 2、 N最短路径的粗切分，根据隐马尔科夫模型和viterbi算法，达到最优路径的规划； 3、人名识别； 4、 系统词典补充；5、 用户自定义词典的补充；6、 词性标注（可选） 2、Ansj分词的准确率 这是我采用人民日报1998年1月语料库的一个测试结果，首先要说明的是这份人工标注的语料库本身就有错误。 P（准确率）：0.984887218571267 R（召回率）：0.9626488103178712 F（综合指标F值）：0.9736410471396494 3、歧义词、未登录词的表现 歧异方面的处理方式自我感觉还可以，基于“最佳实践规则+统计”的方式，虽然还有一部分歧异无法识别，但是已经完全能满足工程应用了。 至于未登录词的识别，目前重点做了中文人名的识别，效果还算满意，识别方式用的“字体+前后监督”的方式，也算是目前我所知道的效果最好的一种识别方式了。 4、算法效率 在我的测试中，Ansj的效率已经远超ictclas的其他开源实现版本。 核心词典利用双数组规划，每秒钟能达到千万级别的粗分。在我的MacBookAir上面，分词速度大约在300w/字/秒，在酷睿i5+4G内存组装机器上，更是达到了400w+/字/秒的速度。 参考文献： Rwordseg说明：http://jianl.org/cn/R/Rwordseg.html ansj中文分词github：https://github.com/NLPchina/ansj_seg ansj中文分词作者专访:http://blog.csdn.net/blogdevteam/article/details/8148451