Elasticsearch-ik\_smart中文分词效果测试报告

1. 测试环境
2. 软件环境

[Elasticsearch-7.8.0](https://www.elastic.co/cn/downloads/elasticsearch)

[Ik -7.8.0](https://github.com/medcl/elasticsearch-analysis-ik/releases)

1. 脚本编写语言

[Python-3.8](https://www.python.org/downloads/release/python-383/)

1. 测试数据集

Second International Chinese Word Segmentation Bakeoff公开的数据源：

[icwb2-data.zip](http://sighan.cs.uchicago.edu/bakeoff2005/)

1. 测试内容

通过黄金标准（Golden standard）、评价指标公式，评测elasticsearch插件ik\_smart分词器的中文分词效果。

1. 测试方法

测试过程分为三个部分叙述，一个是待测数据的处理过程，一个是黄金标准数据处理的过程，最后是数据的获取与计算。

1. 待测数据处理过程：
   1. 将 icwb2-data\testing\pku\_test.txt 中的数据进行切分，每10行作为一条入库数据。并将数据转为符合入库的json格式。
   2. 使用es的api将a) 中的json索引至elasticsearch中。
   3. 通过es的api将库中数据的分词结果取出。
   4. 计算分词结果中，每个分词的偏移量（offset）。
2. 黄金标准数据处理过程：
   1. 将 icwb2-data\gold\pku\_test\_gold.txt 中的数据每10行合成一条数据，便于比对测试数据。
   2. 重新处理 pku\_test\_gold.txt 文件，将两个空格缩进为一个空格。
   3. 计算黄金标准中分词的偏移量。
3. 最后计算的处理过程：
   1. 对比黄金标准与待测数据的偏移量，计算出每条数据的召回率、精准率、F得分、错误率，以及整体平均召回率、精准率、F得分、错误率。

下面对测试所使用的分词效果评测算法做简单的介绍：

变量说明：

N 黄金标准的分词个数

e 分词器错误标准的单词数

c 分词器正确标注的单词数

R 召回率

P 精准率

F f得分

ER 错误率

公式运用：

R = c / N

P = c / (c + e)

F = (2\*P\*R) / (P + R)

ER = e / N

说明：

其中错误单词数的计算有两种算法，一种是以黄金标准为基准计算错误单词数，一种是以测试单词数为基准计算错误单词数，本文采用后者，即如果待测数据将一个词标注成n个，那么错误分词数就是n个；如果结果将多个正确的分词连在一起，那么错误标准也算1个。

1. 测试目标

通过中文分词效果的评测方法以及标准的数据集，对elasticsearch中的ik\_smart分词器插件的分词效进行直观的参考。

1. 测试结果与分析
2. 测试基本情况

数据总条数：194条

总字数：172733个

1. 测试主要结果（保留三位小数）

平均召回率：0.554

平均精准率：0.757

平均f得分：0.638

平均错误率：0.176

1. 测试其他结果

具体在report\report.txt 中查看。

1. 过程数据

具体在 process 文件夹中，有测试方法中每个步骤的输出数据。

1. 整体分析

通过测试数据得知，ik分词器的中文分词效果一般，平均错误率较高，但相比于elasticsearch原生的分词方法，ik分词器也能起到一定的作用。

召回率较低而精准度较高的原因：从样本中抽取例子分析，大部分黄金标准形如

“0,2;2,4;4,6;”，一般分词器会分为“0,2;2,4;0,4;0,6;”等，分词结果涵盖正确结果，但也夹杂着错误结果。所以会导致召回率较低而精准度较高。

1. 参考资料

1．https://blog.csdn.net/Real\_Myth/article/details/50036007?utm\_medium=distribute.pc\_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-3.nonecase&depth\_1-utm\_source=distribute.pc\_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-3.nonecase

2. A Compression-based Algorithm for Chinese Word Segmentation

W. J. Teahan, The Robert Gordon University

Rodger McNab, University of Waikato

Yingying Wen, University of Waikato

Ian H. Witten, University of Waikato