**中文分词测评报告**

平台：

系统：Windows XP

CPU：Intel(R) Pentium(R) [G2030@3，00GHz](mailto:G2030@3.00GHz)

内存：4G

硬盘：100G，7200转

参与测评的分词工具有：

ICTCLAS windows版C内核

ICTCLAS windows版JAVA内核

Fnlp

IKAnalyzer2012\_u6

测试数据源：

源文本：icbw\_msr\_test.txt

标准词汇文本：icbw\_msr\_golden.txt

分词工具介绍：

Anjs\_seg：

简介：这是一个基于google语义模型和条件随机场模型的中文分词的java实现，目前实现了、中文分词、中文姓名识别、用户自定义词典可以应用到自然语言处理等方面,适用于对分词效果要求搞的各种项目。

作者 孙健

分词原理 HMM（隐马尔科夫模型）、CRF（条件随机场）

分词速度 200万字/秒（大约4000K/S）

分词准确率 96%以上

支持语言 中文简体/繁体

最新更新版本 2014年6月

特点：开源分词工具。有三种模块，base模块速度非常快，能达到300万字/秒，但不支持自定义词典；to模块速度和效果较为平衡，同时支持自定义词典和人名识别；nlp模块效率降低，但在分词效果上进一步增强，同时支持机构名识别和新词发现。

ICTCLAS:

简介：ICTCLAS1.0在国内973专家组组织的评测中活动获得了第一名，ICTCLAS2.0在第一届国际中文处理研究机构SigHan组织的评测中都获得了多项第一名，具体的参见系统评测部分。这些都是权威机构进行大规模现场开放测试的结果，真实可信。 目前，ICTCLAS已经向国内外的企业和学术机构颁发了30,000多份授权，其中包括3721、NEC、中华商务网、硅谷动力、云南日报等企业，新疆大学、清华大学、华南理工、麻省大学；同时，ICTCLAS广泛地被《科学时报》、《人民日报》海外版、《科技日报》等多家媒体报道。

作者 中科院计算机研究所

分词原理 HHMM（层叠隐马尔科夫模型）

分词速度 25万字/秒（大约500K/）

分词准确率 98.45%

支持语言 中文简体

最新更新版本 2011年

特点：有开源版本和商业版本，分词效果非常好，但是相对性能较低，支持自定义词典、人名识别、机构名识别、新词发现等，以及语义理解等高级应用。

Jieba：

简介：是一款基于Python写的分词组件，提供三种模式的分词：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义搜索引擎模式；搜索引擎分词模式，在精确模式的基础上，对长词再次切分，提高召回率。

作者

分词原理 基于Trie树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)；采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法

分词速度 1.5 MB / Second in Full Mode

400 KB / Second in Default Mode

分词准确率 98.45%

支持语言 中文（简体&繁体）；英文

支持及扩展性 分词；添加自定义词典；关键词提取；词性标注；

并行分词；返回词语在原文的所在位置；

ChineseAnalyzer for Whoosh搜索引擎

现有版本 Python 2.x；Python 3.x；Java版；C++版；Node.js版；

Erlang版

接口 组件只提供jieba.cut 方法用于分词

特点 在Python编程领域的高准确率、高效率的分词组件。

最新更新版本 2014年6月

CRF中文分词器：

简介：采用由字构词技术路线，同时优化了分词速度，使得极速分词引擎具备了分词速度快和分词精度高的双重优势。国内主要通过CRF++开源软件包来学习该分词方法，但是CRF++过于复杂的代码结构，导致了该算法的普及率。CRF中文分词开源版仅仅包含CRF++软件包中分词解码器部分，简化了CRF++复杂代码结构，清除了分词解码器不需要的代码，大大提高了分词解码器的可读性和可懂度。

作者 langiner工作室

分词原理 CRF（条件随机场，目前分词效果最好的一种算法）

分词速度 350万字/秒

分词准确率 95%

支持语言 中文简体

最新更新版本 2013年4月

特点：开源包中的分词知识库较小，仅用了200万字的分词精度较低，仅供学习Crf分词算法使用。