classifier text classifier

中文分词：CRF( conditional random field)——目前业界较好，较为成熟工具CRF++，可以直接安装

<https://blog.csdn.net/liu_zhlai/article/details/52335527>

结巴中文分词：<https://github.com/search?q=%E5%88%86%E8%AF%8D&type=Repositories>

HanLP：<https://github.com/hankcs/HanLP>

中科院计算所NLPIR http://ictclas.nlpir.org/nlpir/

ansj分词器 https://github.com/NLPchina/ansj\_seg

哈工大的LTP https://github.com/HIT-SCIR/ltp

清华大学THULAC https://github.com/thunlp/THULAC

斯坦福分词器 https://nlp.stanford.edu/software/segmenter.shtml

Hanlp分词器 https://github.com/hankcs/HanLP

结巴分词 https://github.com/yanyiwu/cppjieba

KCWS分词器(字嵌入+Bi-LSTM+CRF) https://github.com/koth/kcws

ZPar https://github.com/frcchang/zpar/releases

IKAnalyzer <https://github.com/wks/ik-analyzer>

哈工大的分词器：主页上给过调用接口，每秒请求的次数有限制。

清华大学THULAC：目前已经有Java、Python和C++版本，并且代码开源。

斯坦福分词器：作为众多斯坦福自然语言处理中的一个包，目前最新版本3.7.0， Java实现的CRF算法。可以直接使用训练好的模型，也提供训练模型接口。

Hanlp分词：求解的是最短路径。优点：开源、有人维护、可以解答。原始模型用的训练语料是人民日报的语料，当然如果你有足够的语料也可以自己训练。

结巴分词工具：基于前缀词典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。

**字嵌入+Bi-LSTM+CRF分词器：本质上是序列标注，这个分词器用人民日报的80万语料，据说按照字符正确率评估标准能达到97.5%的准确率，**各位感兴趣可以去看看。

ZPar分词器：新加坡科技设计大学开发的中文分词器，包括分词、词性标注和Parser，支持多语言，据说效果是公开的分词器中最好的，C++语言编写。

测试数据集

1、SIGHAN Bakeoff 2005 MSR,560KB

http://sighan.cs.uchicago.edu/bakeoff2005/

2、SIGHAN Bakeoff 2005 PKU, 510KB

http://sighan.cs.uchicago.edu/bakeoff2005/

3、人民日报 2014, 65MB

https://pan.baidu.com/s/1hq3KKXe

分词算法：

-基于词典分词算法

-基于统计的机器学习算法 HMM(隐马)、CRF(随机场)

=机器学习算法和词典相结合，一方面能够提高分词准确率，另一方面能够改善领域适应性