**HanLP**是一系列模型与算法组成的NLP工具包，由大快搜索主导并完全开源，目标是普及自然语言处理在生产环境中的应用。**HanLP**具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。

**HanLP**提供下列功能：

* 中文分词
  + HMM-Bigram（速度与精度最佳平衡；一百兆内存）
    - [最短路分词](https://github.com/hankcs/HanLP#1-%E7%AC%AC%E4%B8%80%E4%B8%AAdemo)、[N-最短路分词](https://github.com/hankcs/HanLP#5-n-%E6%9C%80%E7%9F%AD%E8%B7%AF%E5%BE%84%E5%88%86%E8%AF%8D)
  + 由字构词（侧重精度，可识别新词；适合NLP任务）
    - [感知机分词](https://github.com/hankcs/HanLP/wiki/%E7%BB%93%E6%9E%84%E5%8C%96%E6%84%9F%E7%9F%A5%E6%9C%BA%E6%A0%87%E6%B3%A8%E6%A1%86%E6%9E%B6)、[CRF分词](https://github.com/hankcs/HanLP#6-crf%E5%88%86%E8%AF%8D)
  + 词典分词（侧重速度，每秒数千万字符；省内存）
    - [极速词典分词](https://github.com/hankcs/HanLP#7-%E6%9E%81%E9%80%9F%E8%AF%8D%E5%85%B8%E5%88%86%E8%AF%8D)
  + 所有分词器都支持：
    - [索引全切分模式](https://github.com/hankcs/HanLP#4-%E7%B4%A2%E5%BC%95%E5%88%86%E8%AF%8D)
    - [用户自定义词典](https://github.com/hankcs/HanLP#8-%E7%94%A8%E6%88%B7%E8%87%AA%E5%AE%9A%E4%B9%89%E8%AF%8D%E5%85%B8)
    - [兼容繁体中文](https://github.com/hankcs/HanLP/blob/master/src/test/java/com/hankcs/demo/DemoPerceptronLexicalAnalyzer.java#L29)
    - [训练用户自己的领域模型](https://github.com/hankcs/HanLP/wiki)
* 词性标注
  + [HMM词性标注](https://github.com/hankcs/HanLP/blob/master/src/main/java/com/hankcs/hanlp/seg/Segment.java#L584)（速度快）
  + [感知机词性标注](https://github.com/hankcs/HanLP/wiki/%E7%BB%93%E6%9E%84%E5%8C%96%E6%84%9F%E7%9F%A5%E6%9C%BA%E6%A0%87%E6%B3%A8%E6%A1%86%E6%9E%B6)、[CRF词性标注](https://github.com/hankcs/HanLP/wiki/CRF%E8%AF%8D%E6%B3%95%E5%88%86%E6%9E%90)（精度高）
* 命名实体识别
  + 基于HMM角色标注的命名实体识别 （速度快）
    - [中国人名识别](https://github.com/hankcs/HanLP#9-%E4%B8%AD%E5%9B%BD%E4%BA%BA%E5%90%8D%E8%AF%86%E5%88%AB)、[音译人名识别](https://github.com/hankcs/HanLP#10-%E9%9F%B3%E8%AF%91%E4%BA%BA%E5%90%8D%E8%AF%86%E5%88%AB)、[日本人名识别](https://github.com/hankcs/HanLP#11-%E6%97%A5%E6%9C%AC%E4%BA%BA%E5%90%8D%E8%AF%86%E5%88%AB)、[地名识别](https://github.com/hankcs/HanLP#12-%E5%9C%B0%E5%90%8D%E8%AF%86%E5%88%AB)、[实体机构名识别](https://github.com/hankcs/HanLP#13-%E6%9C%BA%E6%9E%84%E5%90%8D%E8%AF%86%E5%88%AB)
  + 基于线性模型的命名实体识别（精度高）
    - [感知机命名实体识别](https://github.com/hankcs/HanLP/wiki/%E7%BB%93%E6%9E%84%E5%8C%96%E6%84%9F%E7%9F%A5%E6%9C%BA%E6%A0%87%E6%B3%A8%E6%A1%86%E6%9E%B6)、[CRF命名实体识别](https://github.com/hankcs/HanLP/wiki/CRF%E8%AF%8D%E6%B3%95%E5%88%86%E6%9E%90)
* 关键词提取
  + [TextRank关键词提取](https://github.com/hankcs/HanLP#14-%E5%85%B3%E9%94%AE%E8%AF%8D%E6%8F%90%E5%8F%96)
* 自动摘要
  + [TextRank自动摘要](https://github.com/hankcs/HanLP#15-%E8%87%AA%E5%8A%A8%E6%91%98%E8%A6%81)
* 短语提取
  + [基于互信息和左右信息熵的短语提取](https://github.com/hankcs/HanLP#16-%E7%9F%AD%E8%AF%AD%E6%8F%90%E5%8F%96)
* [拼音转换](https://github.com/hankcs/HanLP#17-%E6%8B%BC%E9%9F%B3%E8%BD%AC%E6%8D%A2)
  + 多音字、声母、韵母、声调
* [简繁转换](https://github.com/hankcs/HanLP#18-%E7%AE%80%E7%B9%81%E8%BD%AC%E6%8D%A2)
  + 简繁分歧词（简体、繁体、臺灣正體、香港繁體）
* [文本推荐](https://github.com/hankcs/HanLP#19-%E6%96%87%E6%9C%AC%E6%8E%A8%E8%8D%90)
  + 语义推荐、拼音推荐、字词推荐
* 依存句法分析
  + [基于神经网络的高性能依存句法分析器](https://github.com/hankcs/HanLP#21-%E4%BE%9D%E5%AD%98%E5%8F%A5%E6%B3%95%E5%88%86%E6%9E%90)
  + [MaxEnt依存句法分析](https://github.com/hankcs/HanLP/blob/master/src/main/java/com/hankcs/hanlp/dependency/MaxEntDependencyParser.java)
* [文本分类](https://github.com/hankcs/HanLP/wiki/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB%E4%B8%8E%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90)
  + [情感分析](https://github.com/hankcs/HanLP/wiki/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB%E4%B8%8E%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90#%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90)
* [word2vec](https://github.com/hankcs/HanLP/wiki/word2vec)
  + 词向量训练、加载、词语相似度计算、语义运算、查询、KMeans聚类
  + 文档语义相似度计算
* [语料库工具](https://github.com/hankcs/HanLP/tree/master/src/main/java/com/hankcs/hanlp/corpus)
  + 默认模型训练自小型语料库，鼓励用户自行训练。所有模块提供训练接口，语料可参考[OpenCorpus](https://github.com/hankcs/OpenCorpus)。

在提供丰富功能的同时，**HanLP**内部模块坚持低耦合、模型坚持惰性加载、服务坚持静态提供、词典坚持明文发布，使用非常方便，同时自带一些语料处理工具，帮助用户训练自己的模型。