Atitit java的nlp自然语言处理类库

目录

[1.1. OpenNLP是一个基于Java机器学习工具包， 1](#_Toc10966)

[1.2.  Java自然语言处理 LingPipe 1](#_Toc15158)

[1.3. 2.中文自然语言处理工具包 FudanNLP 1](#_Toc5292)

[1.3.1.  5、Stanford CoreNLP 斯坦福大学NLP 2](#_Toc10488)

[1.4. 1.IKAnalyzer 2](#_Toc2229)

[1.5. 3.FudanNLP 2](#_Toc22414)

[1.6. jcseg 3](#_Toc5381)

## FNLP主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。

本工具包及其包含数据集使用LGPL3.0许可证。

## **功能(Functions)**

信息检索： 文本分类 新闻聚类

中文处理： 中文分词 词性标注 实体名识别 关键词抽取 依存句法分析 时间短语识别

结构化学习： 在线学习 层次分类 聚类

## OpenNLP是一个基于Java机器学习工具包，

用于处理自然语言文本。支持大多数常用的 NLP 任务，例如：标识化、句子切分、部分词性标注、名称抽取、组块、解析等。

## Java自然语言处理 LingPipe

LingPipe是一个自然语言处理的Java开源工具包。LingPipe目前已有很丰富的功能，包括主题分类（Top Classification）、命名实体识别（Named Entity Recognition）、词性标注（Part-of Speech Tagging）、句题检测（Sentence Detection）、查询拼写检查（Query Spell Checking）、兴趣短语检测（Interseting Phrase Detection）、聚类（Clustering）、字符语言建模（Character Language Modeling）、医学文献下载/解析/索引（MEDLINE Download, Parsing and Indexing）、数据库文本挖掘（Database Text Mining）、中文分词（Chinese Word Segmentation）、情感分析（Sentiment Analysis）、语言辨别（Language Identification）等API。

下载链接：<http://alias-i.com/lingpipe/web/download.html>

## 2.中文自然语言处理工具包 FudanNLP

FudanNLP主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。

演示地址: <http://jkx.fudan.edu.cn/nlp/query>

FudanNLP目前实现的内容如下：

1. 中文处理工具
   1. 中文分词
   2. 词性标注
   3. 实体名识别
   4. 句法分析
   5. 时间表达式识别
2. 信息检索
   1. 文 本分类
   2. 新闻聚类
   3. [Lucene中文分词](http://code.google.com/p/fudannlp/wiki/lucene)
3. 机 器学习
   1. Average Perceptron
   2. Passive-aggressive Algorithm
   3. K-means
   4. Exact Inference

### **5、[Stanford CoreNLP](http://nlp.stanford.edu/software/corenlp.shtml) 斯坦福大学NLP**

很牛叉的一个库

下载地址 http://search.maven.org/#browse%7C11864822

学习自然语言这一段时间以来接触和听说了好多开源的自然语言处理工具，在这里做一下汇总方便自己以后学习，其中有自己使用过的也有了解不是很多的，对于不甚了解的工具以后学习熟悉了会做更新的

## 1.IKAnalyzer

IK Analyzer是一个开源的，基于Java语言开发的轻量级的中文分词工具包。从2006.12推出1.0版本开始，IK Analyzer已经推出了多个版本，当前最新版本为2012 u6，最初基于Luence，从3.0开始成为面向Java的公用分词组件，独立于Luence，下载地址为：[http://git.oschina.net/wltea/IK-Analyzer-2012FF](http://git.oschina.net/wltea/IK-Analyzer-2012FF" \t "https://www.cnblogs.com/aeexiaoqiang/p/_blank)。IK支持细粒度和智能分词两种切分模式，支持英文字母、数字、中文词汇等分词处理，兼容韩文、日文字符。可以支持用户自定义的词典，IKAnalyzer.cfg.xml文件来实现，可以配置自定义的扩展词典和停用词典。词典需要采用UTF-8无BOM格式编码，并且每个词语占一行。配置文件如下所示

## 3.FudanNLP

FudanNLP主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。FudanNLP及其包含数据集使用LGPL3.0许可证。

主要功能包括：

信息检索：文本分类，新闻聚类。

中文处理：中文分词，词性标注，实体名识别，关键词抽取，依存句法分析，时间短语识别。

结构化学习：在线学习，层次分类，聚类，精确推理。

工具采用Java编写，提供了API的访问调用方式。最新版本为FudanNLP-1.6.1，下载地址为：[http://code.google.com/p/fudannlp/](http://code.google.com/p/fudannlp/" \t "https://www.cnblogs.com/aeexiaoqiang/p/_blank)。

在使用时将fudannlp.jar以及lib中的jar部署于项目中的lib里面。models文件夹中存放的模型文件，主要用于分词、词性标注和命名实体识别以及分词所需的词典；文件夹example中主要是使用的示例代码，可以帮助快速入门和使用；java-docs是API帮助文档；src中存放着源码；PDF文档中有着比较详细的介绍和自然语言处理基础知识的讲解。

初始运行程序时初始化时间有点长，并且加载模型时占用内存较大。在进行语法分析时感觉分析的结果不是很准确。

## **jcseg**

**Jcseg**是基于mmseg算法的一个轻量级中文分词器，同时集成了关键字提取，关键短语提取，关键句子提取和文章自动摘要等功能，并且提供了一个基于Jetty的web服务器，方便各大语言直接http调用，同时提供了最新版本的lucene, solr, elasticsearch的分词接口！**Jcseg**自带了一个 jcseg.properties文件用于快速配置而得到适合不同场合的分词应用，例如：最大匹配词长，是否开启中文人名识别，是否追加拼音，是否追加同义词等！

FNLP是一套Java写就的中文NLP工具包 - Java开发社区 \_ CTOLib码库.html