角色识别系统

# 什么是角色标注系统

# Emory NLP的角色标注系统

# Emory NLP的自然语言处理系统

## 简介

Natural Language Processiong for JVM languages（NLP4J）项目提供的主要功能有：

* 可在各种研究过程中使用的工具
* 开发高效健壮的NLP组件
* 操作NLP特有数据结构的API（例如：dependecy graph）

该项目目前由 [Emory NLP](http://nlp.mathcs.emory.edu/) 研究组创建并管理。

* 最新释放版本：1.1.3
* 论坛（国内无法访问）
* demo（国内无法访问）

## 快速开始

### 怎样安装

提供两种安装方法，方法1为MVN安装包（在Windows下总提示文件系统IO有错误）。这里请使用mac os

方法2为直接下载汇编包，该包中包含了所有需要的class文件以及针对英文的模型。下载汇编包[nlp4j-appassembler-1.1.3.tgz](http://nlp.mathcs.emory.edu/nlp4j/nlp4j-appassembler-1.1.3.tgz)后解压并运行下边的命令，如果返回下表所示内容，表示安装成功。

|  |
| --- |
| > bin/version  =========================================  NLP4J Version 1.1.3  Contact: choi@mathcs.emory.edu  Webpage: https://emorynlp.github.io/nlp4j  ========================================= |

### 怎样训练

#### nlptrain的命令行参数

1. 命令参数

|  |  |
| --- | --- |
| 参数 | 意义 |
| -c | 配置文件 |
| -m | 模型文件 |
| -p | 已有模型文件 |
| -t|d | 训练集|验证集数据所在目录或单独一个文件 |
| -te|de | 训练数据文件的扩展名，默认为任意 |
| -cv | 交叉验证集比例 |
| -mode | 训练算法名称，可用算法见下表 |

2. NLP算法表

|  |  |  |
| --- | --- | --- |
| 算法名称 | 说明 | 备注 |
| custom | 定制NLP | 不知用途 |
| pos | 词性标注 |  |
| ner | 命名实体识别 |  |
| dep | 依存句法分析 |  |
| srl | 语义角色标注 |  |
| doc | 文档分类 |  |
| it | it分类 | 不知用途 |
| sentiment\_ensemble | 情感分析：ensemble方法 |  |

3. 配置文件

对模型的的配置说明，对训练数据格式的说明。具体参考

* <https://emorynlp.github.io/nlp4j/quickstart/train.html#configuration>
* <https://github.com/emorynlp/nlp4j/tree/master/api/src/main/resources/edu/emory/mathcs/nlp/configuration/>

#### 示例

这里给出一个训练依存句法分析模型的示例。

执行命令如下

|  |
| --- |
| > java -Xmx1g -XX:+UseConcMarkSweepGC java edu.emory.mathcs.nlp.bin.NLPTrain -mode dep -c config-train-sample.xml -t sample-trn.tsv -d sample-dev.tsv -m sample-dep.xz |

选项说明：

[-XX:+UseConcMarkSweepGC](http://www.oracle.com/technetwork/java/tuning-139912.html) 用于减少内存消耗。

[log4j.properties](https://github.com/emorynlp/nlp4j/blob/master/api/src/main/resources/edu/emory/mathcs/nlp/configuration/log4j.properties) 日志工具

模型文件的生成：

一旦训练结束，sample-dep.xz会被创建，该功能可以在configuration文件中的指定，要求dependency parsing功能。

训练过程的终端信息说明：

L：类别数量

SF：离散特征数量

NZW：非零特征数量

N/S：每秒中处理的节点数

### 怎样decode

## NLP组件

## 补充说明