人工智能与数据挖掘课程设计

命名实体识别

**组员信息：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 姓名 | 学号 | 成员贡献 | 实践成绩 |
| 1 | 郭嘉琪 | 2021522001 | 代码运行环境布置、报告编写 |  |
| 2 | 王雅曈 | 2021522005 | 代码编写、报告编写 |  |
| 3 | 赵欣然 | 2021522013 | 代码编写、报告编写 |  |
| 4 |  |  |  |  |

# **引言**

paddlepaddle是百度的一个深度学习框架,该框架的生态环境中，有一个paddleNLP开源子项目，该项目提供了当前大部分NLP预训练模型训练好的模型参数，我们可以使用paddleNLP提供的预训练模型来做NER命名实体识别任务。

# **实验数据**

## 数据获取或预处理

采用conda方式配置环境，可参考下面链接：

<https://www.paddlepaddle.org.cn/install/quick?docurl=/documentation/docs/zh/install/docker/linux-docker.html>

主要依赖包paddlepaddle-gpu==2.5.2、paddlenlp==2.6.1

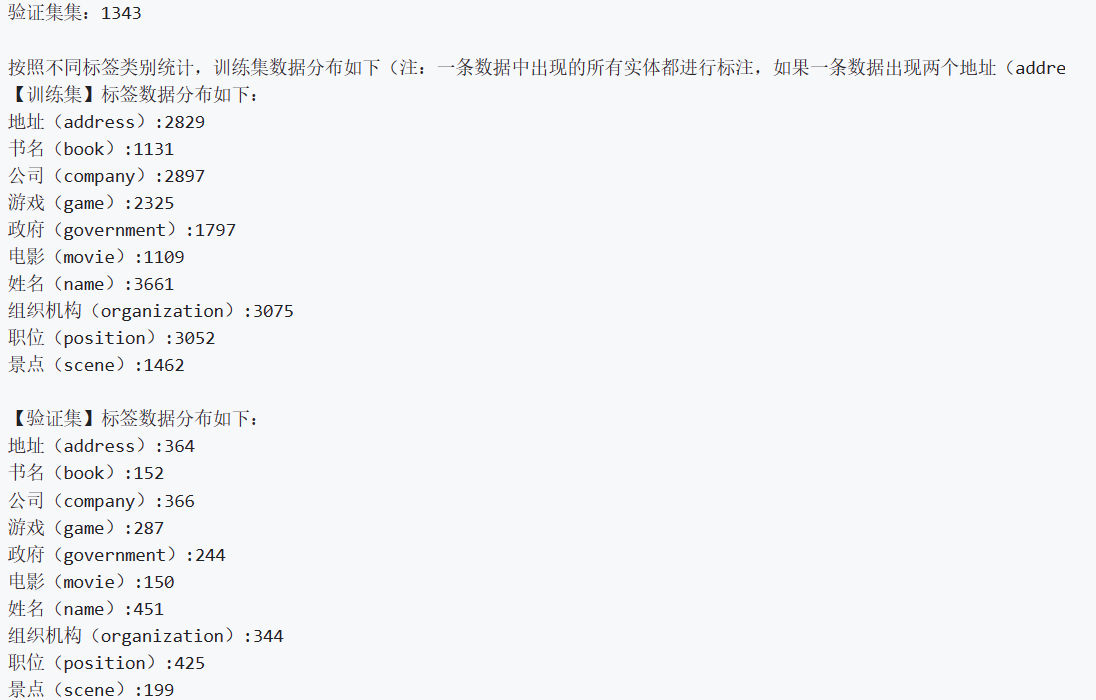
本次数据是在清华大学开源的文本分类数据集THUCTC基础上，选出部分数据进行细粒度命名实体标注，原数据来源于Sina News RSS.

训练数据。其中，文件模型训练数据集采用clue命名实体识别数据集。

首先安装 Anacodna，再安装相应第三方库bash，进入anaconda的python环境

pip install -i https://pypi.tuna.tsinghua.edu.cn/simple paddlepaddle

pip install -i https://pypi.tuna.tsinghua.edu.cn/simple paddlenlp



## 数据统计

方法基于paddlepaddle深度学习架构，采用bert-wwm-chinese预训练模，结合交叉熵损失函数，实现命名实体标签预测。

首先克隆项目bash

git clone https://gitee.com/lingcb/paddle-ner.git

cd paddle-ner

python src/predict.py      #预测

python src/eval.py         #评估模型

python main.py             #训练模型

# **方法/模型**

在运行上面项目之前，需要设置一些parameter.py文件里的一些参数

首先训练bash

cd paddle-ner

python main.py

其次预测和评估bash

cd paddle-ner

python src/predict.py      #预测

python src/eval.py         #评估模型

训练与评估需要改的参数，相似，其中checkpoint\_base是上面百度链接文件夹的路径