人工智能与数据挖掘课程设计

报告题目

**组员信息：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 姓名 | 学号 | 成员贡献 | 实践成绩 |
| 1 | 刘思蕊 | 2021211918 |  |  |
| 2 | 李雨霏 | 2021211919 |  |  |
| 3 | 丁宁 | 2021211924 |  |  |
| 4 |  |  |  |  |

# **引言**

（1）命名实体识别任务分析

命名实体识别（Named Entity Recognition，简称NER），也称为专名识别，是自然语言处理（NLP）领域的一项关键技术。它的主要任务是在文本数据中识别出具有特定意义的实体，这些实体通常包括人名、地名、机构名、专有名词等。NER的目的在于将这些具有特定意义的实体提取出来，以便于后续的知识挖掘、信息抽取等任务。

命名实体识别的过程主要包括两个部分：

实体边界识别：确定文本中实体的起始和结束位置。例如，在句子“奥巴马是美国总统”中，需要识别出“奥巴马”和“美国”这两个实体的边界。

实体类别确定：判断识别出的实体属于哪个类别，如人名、地名、机构名等。在上述例子中，奥巴马和美国分别属于人名和地名类别。

目前，NER技术主要分为基于规则和字典的方法、基于统计模型的方法以及基于深度学习的方法。

基于规则和字典的方法：通过预先制定一套规则或构建一个实体词典，对文本进行扫描和匹配，从而实现实体的识别。这种方法的优点是易于实现，但缺点是扩展性差，对于新出现的实体或复杂场景难以应对。

基于统计模型的方法：通过对大量已标注的文本数据进行学习，训练出一种统计模型，用于对新的文本进行实体识别。常见的统计模型包括条件随机场（CRF）、支持向量机（SVM）等。这种方法的优点是具有一定的泛化能力，但缺点是需要大量的人工标注数据和特征工程。

基于深度学习的方法：利用深度学习技术，如卷积神经网络（CNN）、循环神经网络（RNN）、长短时记忆网络（LSTM）等，对文本进行端到端的建模，实现实体的识别。这种方法的优点是自动提取特征，减少人工干预，但缺点是需要大量的训练数据和计算资源。

（2）命名实体识别方法

a. BERT模型。BERT模型是基于Transformers Encoder，使用MLM获取双向融合信息，在海量连续语料上进行预训练，得到的预训练模型，基于下游任务只需简单的fine-tuning就能获得特别好的结果。

b.LSTM（循环神经网络）算法。是一种特殊的递归神经网络 。这种网络与一般的前馈神经网络不同，LSTM可以利用时间序列对输入进行分析。对于 LSTM，对于序列中的每个元素， 有相应的隐藏状态 ht​，原则上可以包含序列中较早的任意点的信息。

c.CRF（条件随机场）算法。CRF（条件随机场）算法是给定一组输入序列条件下另一组输出序列的条件概率分布模型。CRF中有两类特征函数，分别是状态特征和转移特征。前者只针对当前位置的字符可以被转换成哪些实体label，后者关注的是当前位置和其相邻位置的字符可以有哪些实体label的组合。

# **实验数据**

## 数据获取或预处理

（1）数据获取

数据来源于人民日报NER数据集

简介：本NER数据集由人民日报语料库1998版和2014版生成，包含了人名(PER)、地名(LOC)和机构名(ORG)3类常见的实体类型。

语种：Chinese

实体类别数量：3

链接：

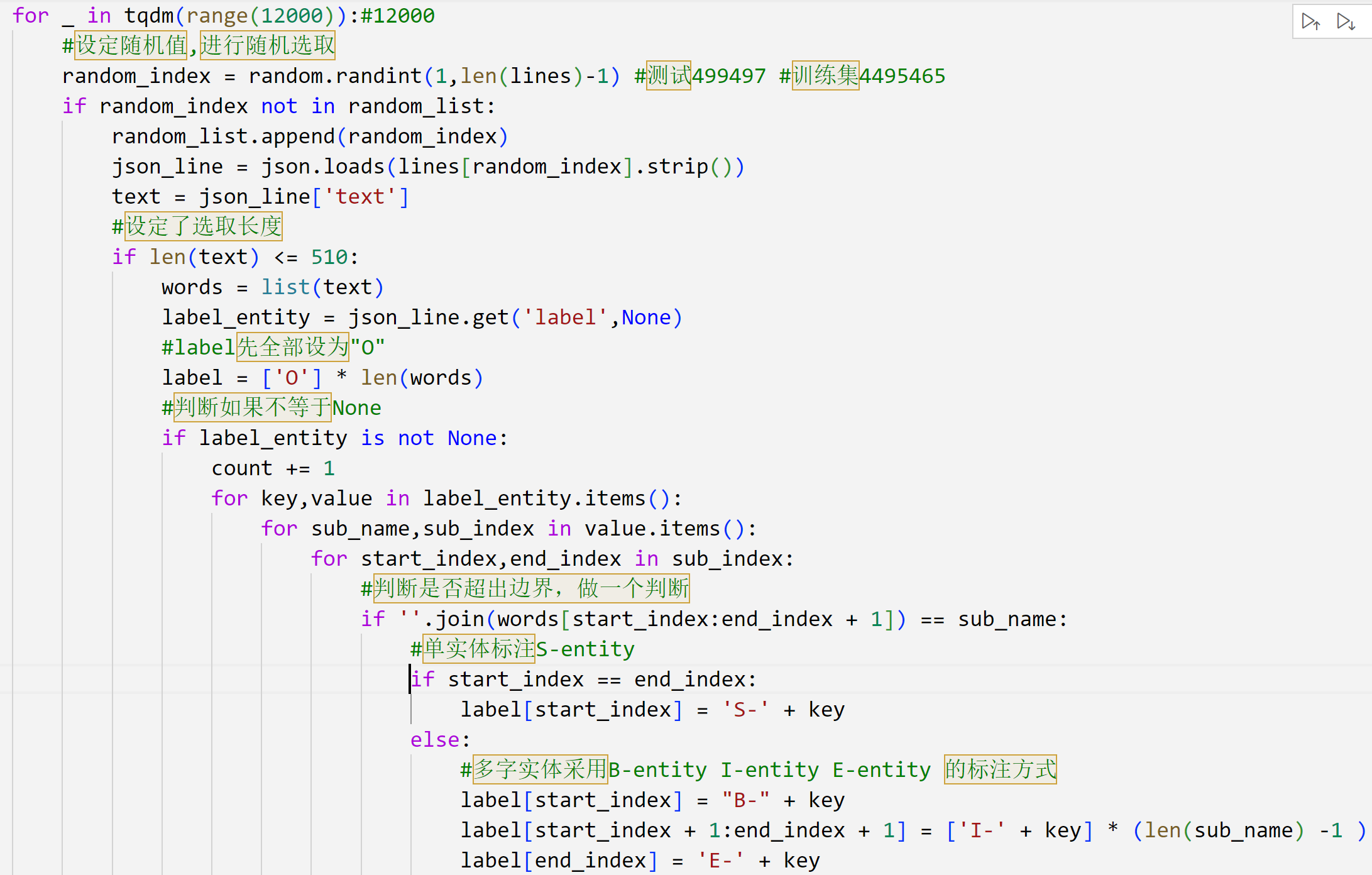
<https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/renMinRiBao>

<https://raw.githubusercontent.com/InsaneLife/ChineseNLPCorpus/master/NER/renMinRiBao/renmin.txt>

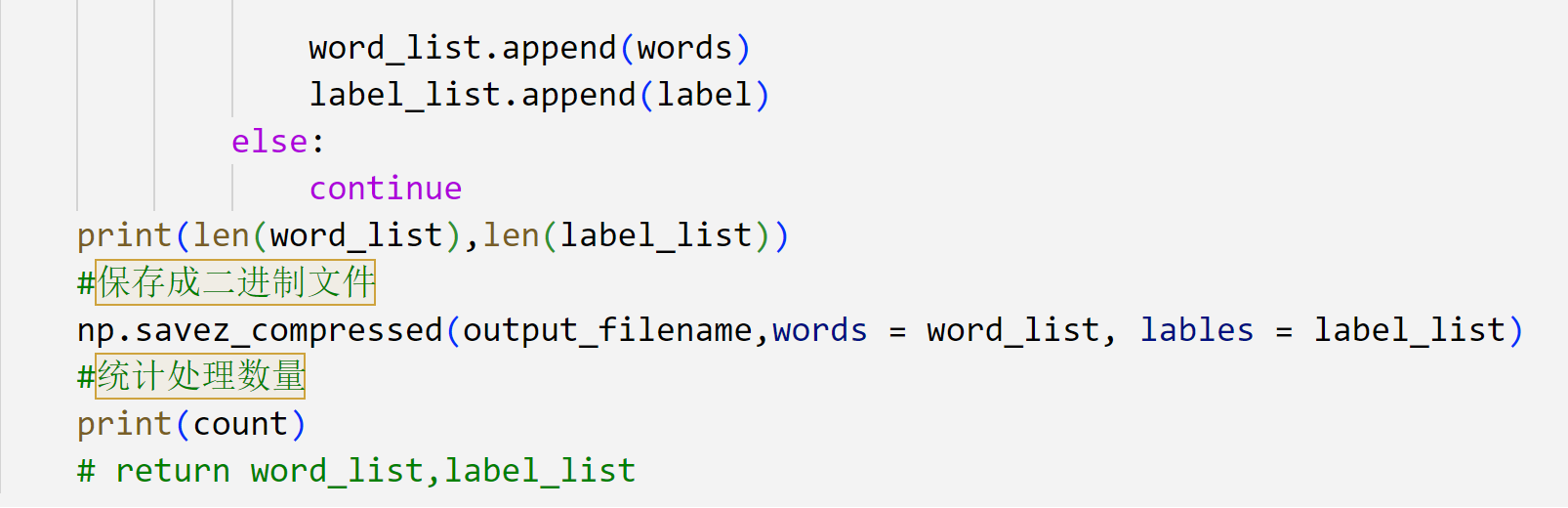


（2）数据预处理

a. 单实体标注S-entity，多字实体采用B-entity I-entity E-entity 的标注方式。具体代码如下：



b. 保存为二进制文件并统计数据处理数量。



## 数据统计

方法介绍

# **方法/模型**

功能描述与实现方法介绍。

# **结果评测/展示**

方法评测或功能展示。

# **总结**

工作总结、局限性分析。