

# 2024知识工程第一次大作业

徐帅 xuanshen.818@gmail.com学号: 1320221095

## 1 引言

这是一段利用词性作为特征，进行指代消解的代码。

在最开始我是利用词频来作为特性来完成这个任务的，最后内存不够跑不出来就放弃了，发现词性真的十分便捷，详情我会写README在压缩包里，这里不过多赘述。

## 2 代码详细解释

### 2.1

首先是数据处理部分，这里我直接利用了我做词频时的数据预处理，可能会有不需要的部分不过没关系。

数据预处理共有几个部分，我在刚刚读完要求的时候就写了几个相关的预处理，不过最后应用的时候发现没有太大作用，包括删除除了中文、英文、数字和空格之外的所有特殊字符，除去空行等等。

关于比较重要的部分有，对名词的提取，提取出来之后会有很多的名词，数量很大，具体是因为有重复的，所以我进行了去重，及如果有重复的便合起来（因为最开始用词频做的，所以我当初计算了名词的出现次数之后才合起来的）。合起来之后发现有的词前面带有【，例如中国和【中国就变成了两个词了，所以我对【进行了删除，然后再合并。

之后在读取json文件时我发现了有部分json文件为空文件，在这里可以选择跳过或删除，我选择了删除，主要原因是跳过的话每次都要跳过，删除是一次性的。

### 2.2

然后是正式代码部分

首先是对资源的读取。包括读取原始语料库txt文件，还有读取预处理完毕的词库文件n\_sort\_delete\_number.txt，这是处理完毕之后的名词（命名是在做词频的时候起名的，是按词频排序了，并且删除了词语出现的次数的数字，故此命名），然后还有相关的训练测试等的json文件。

然后是获取句子并提取词性，利用python的nltk库，接下来是生成特征向量和标签向量，并计算最长句子长度（这里的目的是主要是调整向量的维度）。之后将特征和标签转换为numpy数组，这里其实是有点害怕内存爆了，毕竟之前词频的时候爆过一次。

接下来计算正负样本的权重（因为正负类的差距其实很大），将数据转换为PyTorch张量，创建DataLoader。之后定义神经网络模型（不是完全应用神经网络，主要还是模型的训练）这个模型是一个简单的前馈神经网络，它有一个输入层，一个隐藏层，和一个输出层，输入层的大小是词性标签的数量，隐藏层大小是128，输出层大小是1。

创建损失函数和优化器，采用二元交叉熵损失函数和Adam优化器。并且将模型移动到GPU上（不必须）。

---


然后基本就是最后的模型构建和训练和在测试集上评估模型（model.eval（））了。迭代训练模型，同时计算损失，在每个epoch 完成后，在测试集上评估模型的性能。最后绘制训练损失曲线。

## 结果

下面所附两张图分别是结果输出图像和loss梯度下降图像。具体数值为Accuracy on the test set: 45344.92Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756。

loss图像从最高1.4285下降到1.3051并趋于稳定。

最后的这张图是我实现指代消解所经历的过程，其中包括很多东西，包括生成出来的文件啊，或者是我一次接一次的更改代码产生的我也不知道多少版本的代码等等，最后完成了这一任务还是满感慨的。



```
Epoch 474/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 475/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 476/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 477/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 478/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 479/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 480/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 481/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 482/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 483/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 484/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 485/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 486/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 487/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 488/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 489/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 490/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 491/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 492/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 493/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 494/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 495/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 496/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 497/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 498/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 499/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
Epoch 500/500, Loss: 1.3051, Accuracy on the test set: 45344.92%
Precision: 0.0454, Recall: 0.2260, F1 Score: 0.0756
```

Figure 1: 结果输出

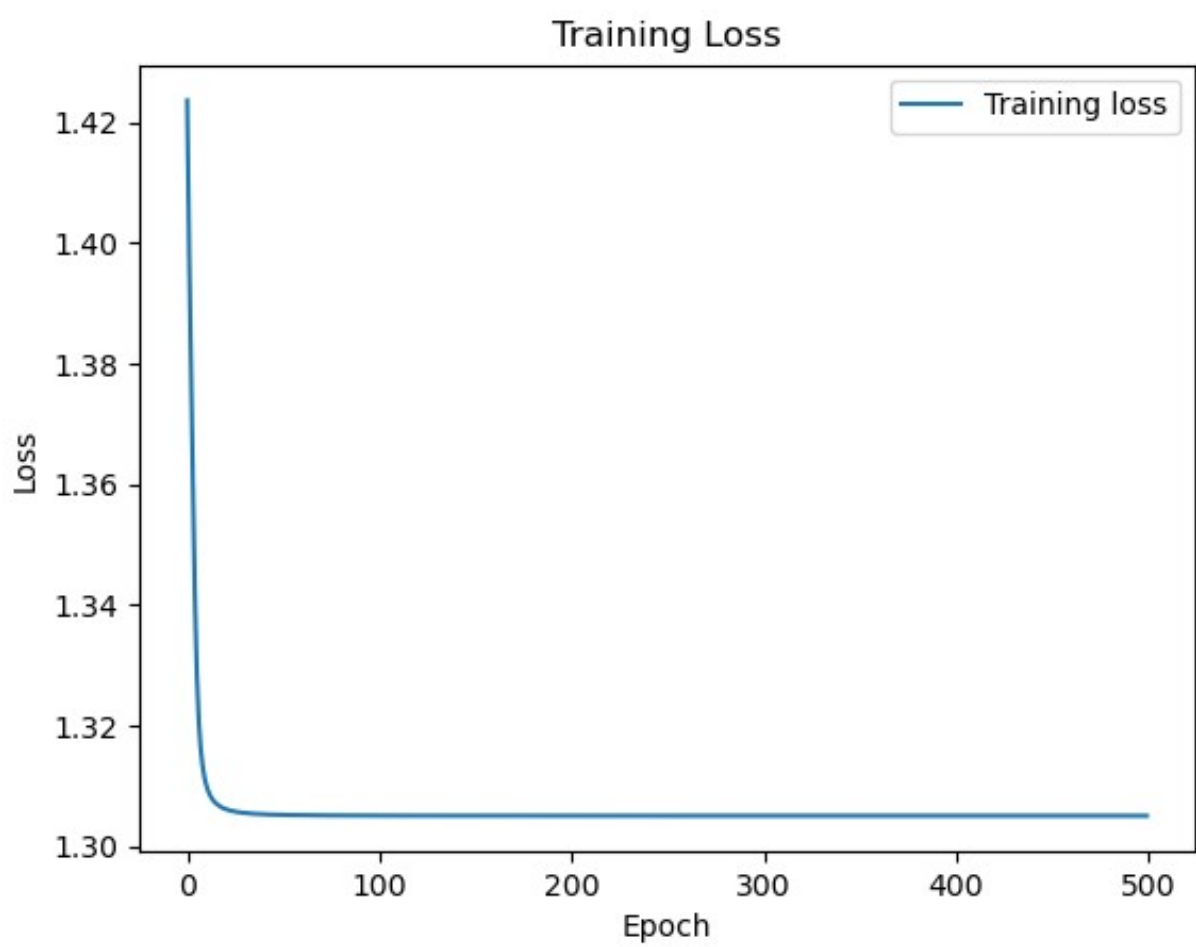


Figure 2: loss图

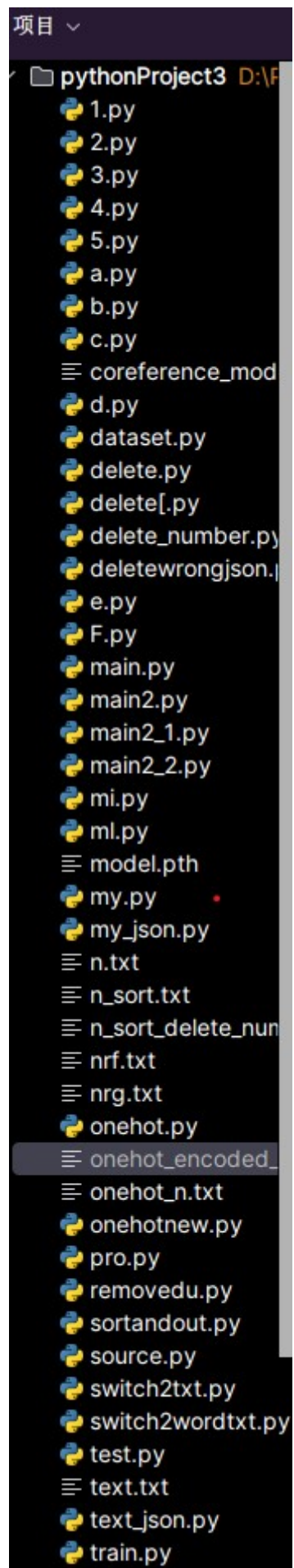


Figure 3: 实现过程