**实验四 情感分类 作业报告**

1. 实验目的
   1. 进一步加深对卷积神经网络基本原理的理解。
   2. 掌握卷积神经网络处理文本的各项技术。
   3. 掌握文本分类模型Text-CNN的架构和原理
2. 实验要求
   1. 基于TensorFlow建立Text-CNN模型。
   2. 实现对中文电影评论的情感分类。
3. 实验原理

Text-CNN和传统的CNN结构类似，具有词嵌入层、卷积层、池化层和全连接层的四层结构。

其中，Text-CNN的词嵌入层使用二维矩阵来表示长文本。词嵌入（Word embedding）将输入文本的每个词语通过空间映射，将独热表示（One-Hot Representation）转换成分布式表示（Distributed Representation），进而可以使用低维的词向量来表示每一个词语。经过词嵌入，每个单词具有相同长度的词向量表示。将各个词语的向量表示连起来便可以得到二维矩阵。得到词向量的方式有多种，常用的是Word2vec方法。若使用预训练好的词向量，在训练模型的时候可以选择更新或不更新词向量，分别对应嵌入层状态为Non-static和Static。

Text-CNN的卷积层是主要部分，卷积核的宽度等于词向量的维度，经卷积后可以提取文本的特征向量。和在图像领域应用类似，Text-CNN可以设置多个卷积核以提取文本的多层特征，长度为N的卷积核可以提取文本中的N-gram特征。

Text-CNN的池化层一般采取Max-over-time pooling，输出最大值，从而判断词嵌入中是否含N-gram。

Text-CNN的全连接层采用了Dropout算法以防过拟合，并使用Softmax函数输出各个类别的概率。

算法具体原理可阅读：Kim Y .2014--《Convolutional Neural Networks for Sentence Classification》一文。

1. 实验所用工具及数据集
2. 主要工具

Python 3.5+、TensorFlow 1.3.0、Numpy 1.13.1、jieba 0.39

1. 数据集
2. 训练集。包含2W条左右中文电影评论，其中正负向评论各1W条左右。
3. 验证集。包含6K条左右中文电影评论，其中正负向评论各3K条左右。
4. 测试集。包含360条左右中文电影评论，其中正负向评论各180条左右。
5. 预训练词向量。中文维基百科词向量word2vec。
6. 实验步骤与方法



加载数据集

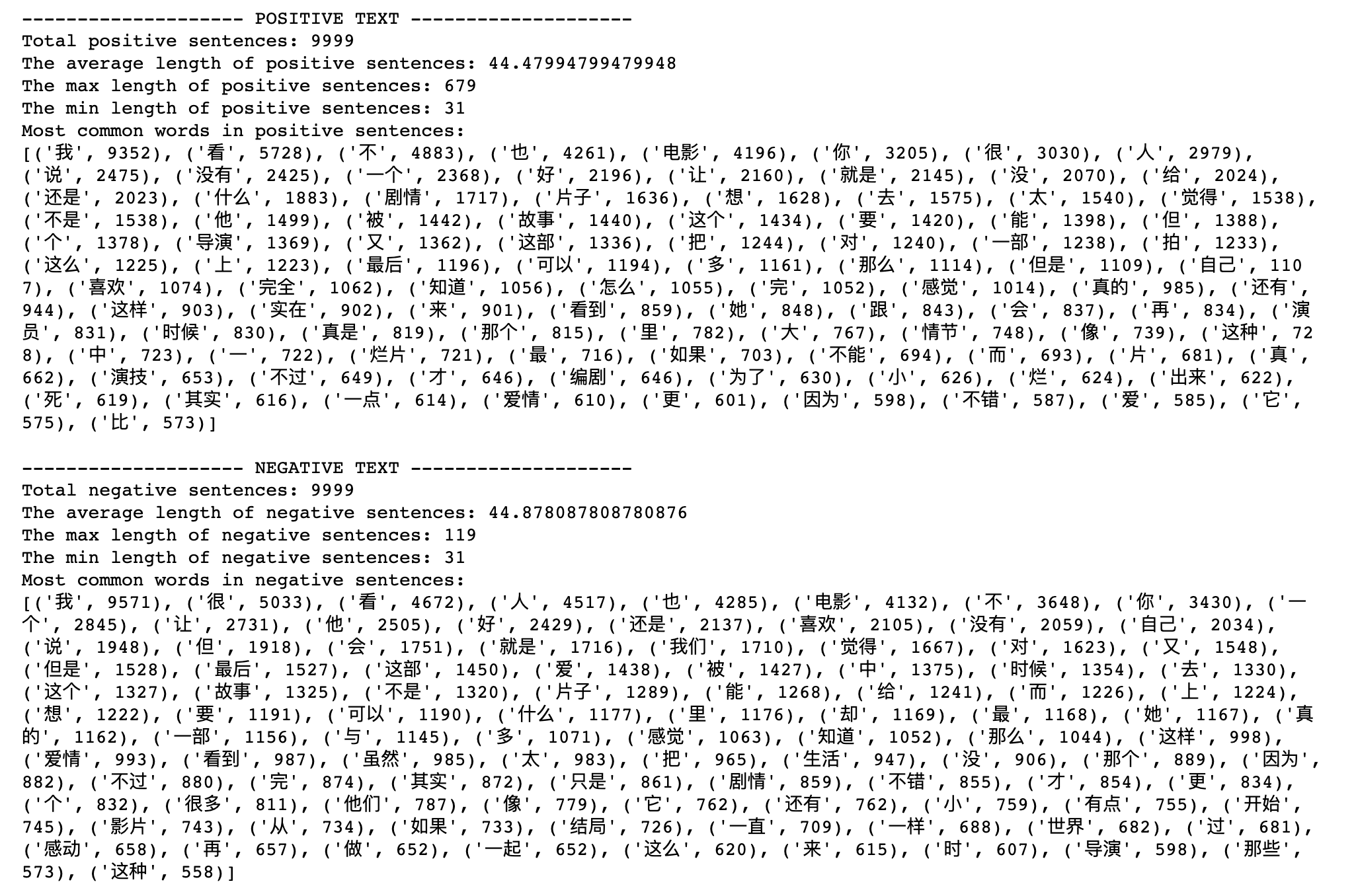


统计文本信息，输出高频词

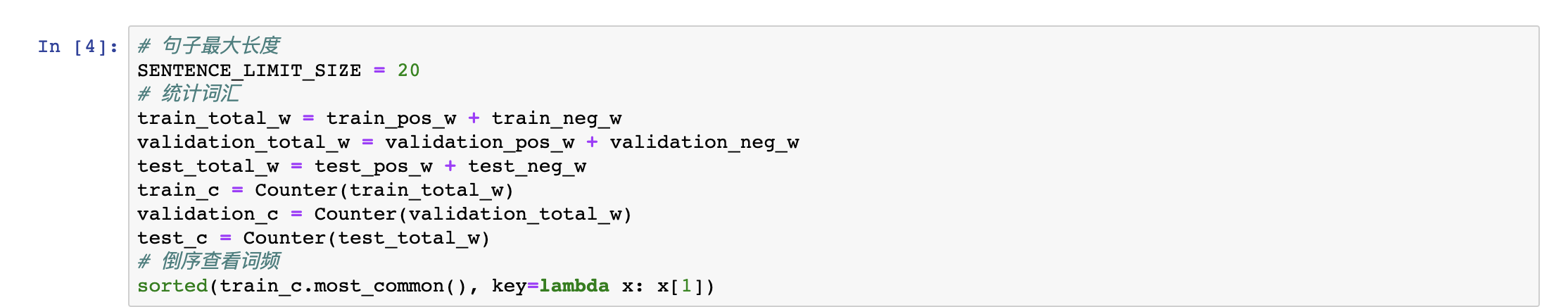




文本统计信息展示

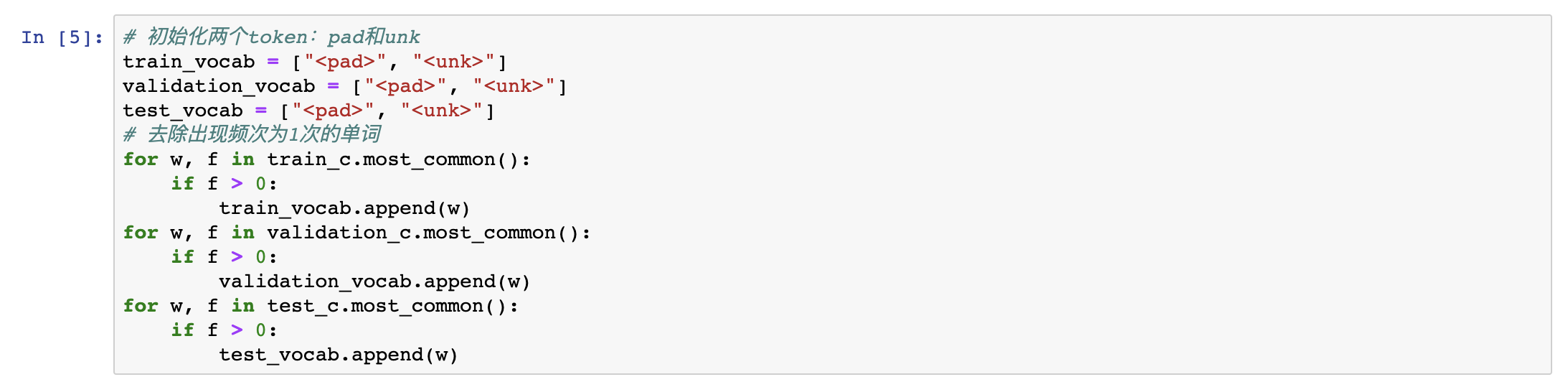


倒叙查看词频

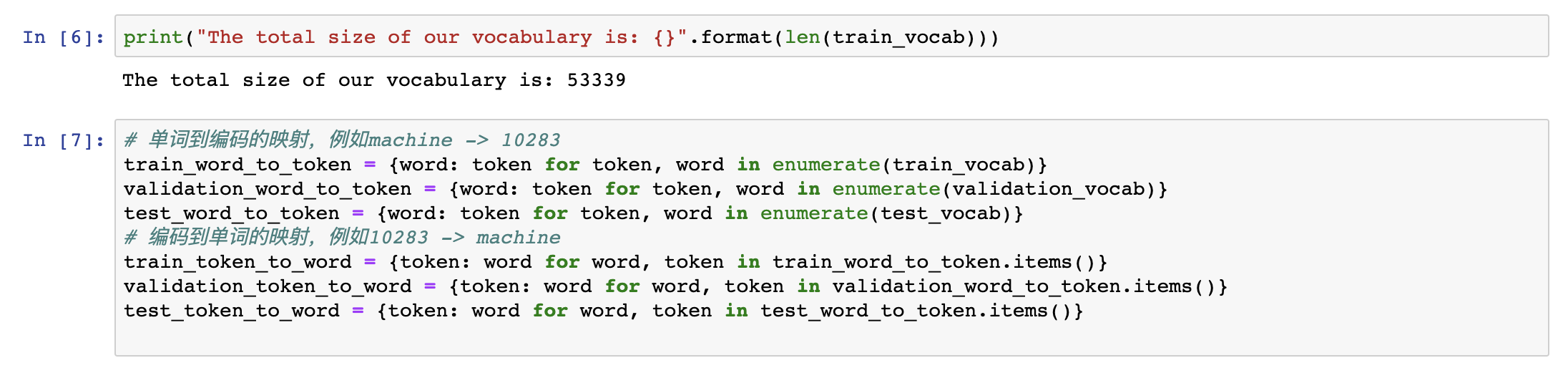
  


初始化两个token：<pad>和<unk>

生成词汇表



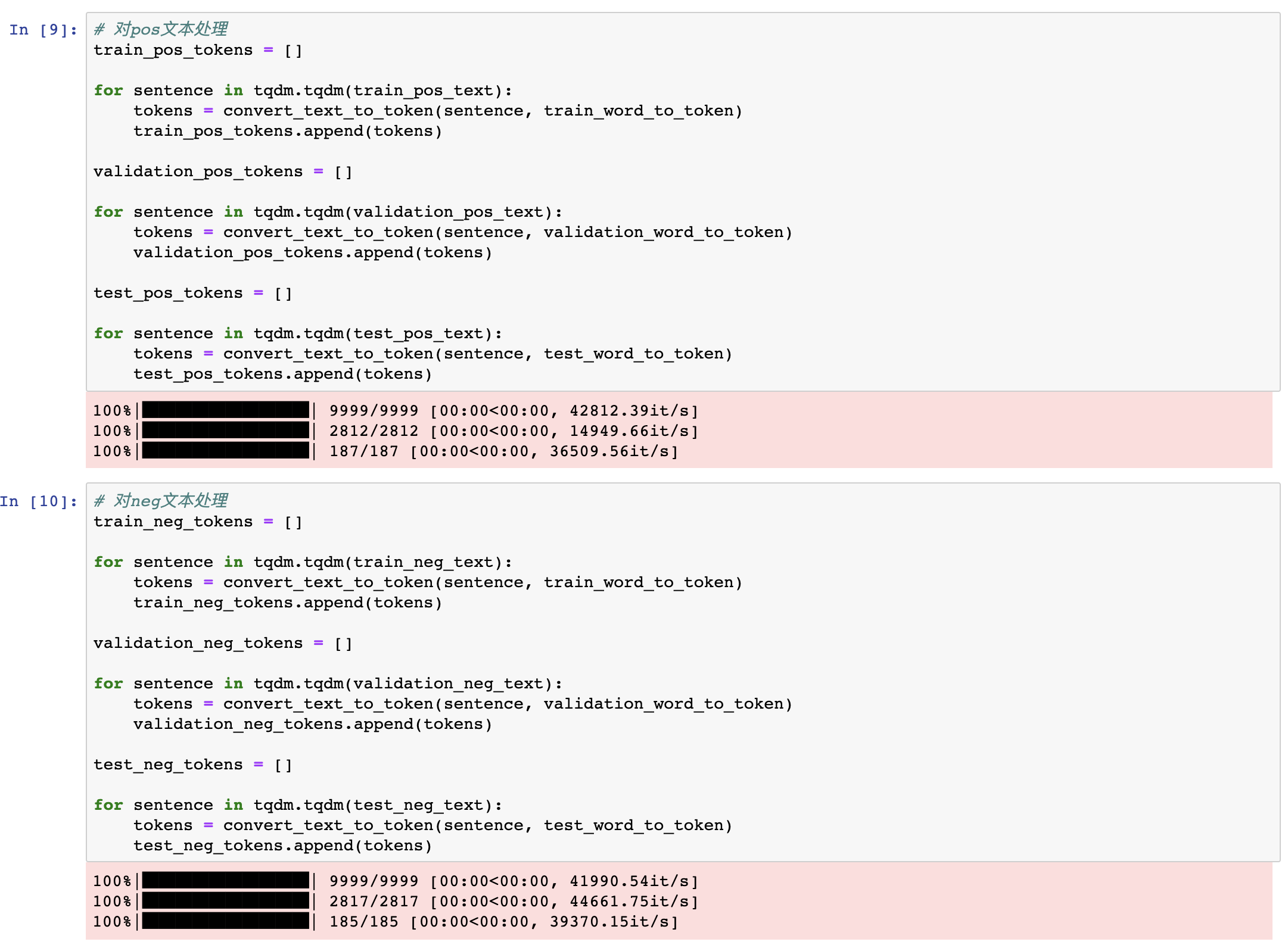
生成词汇和id映射的字典



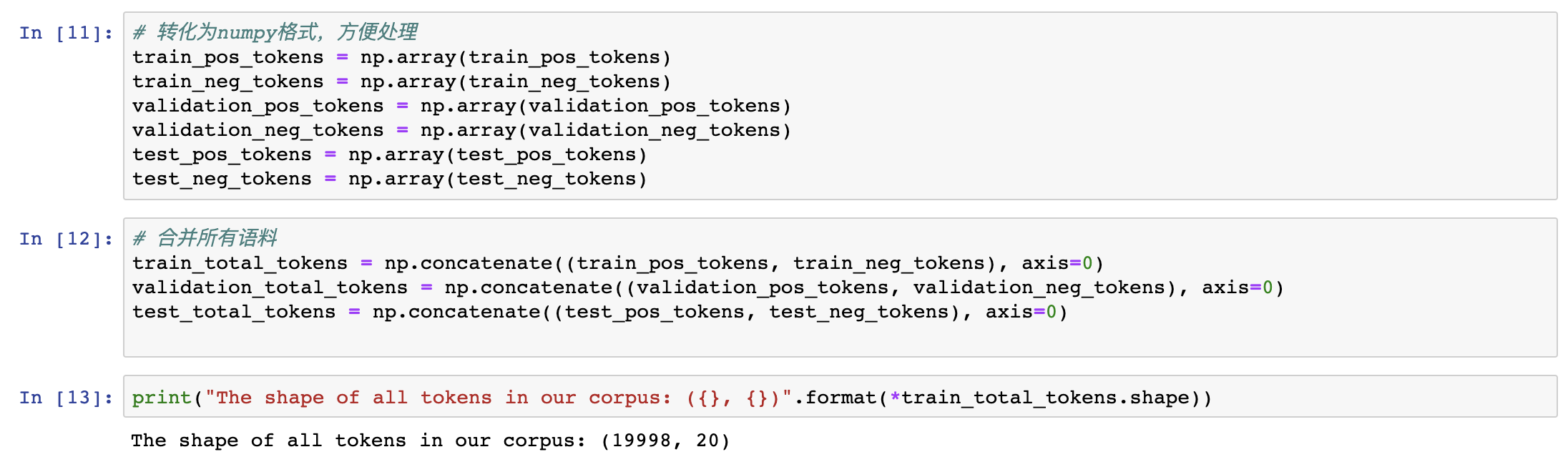
定义转换函数



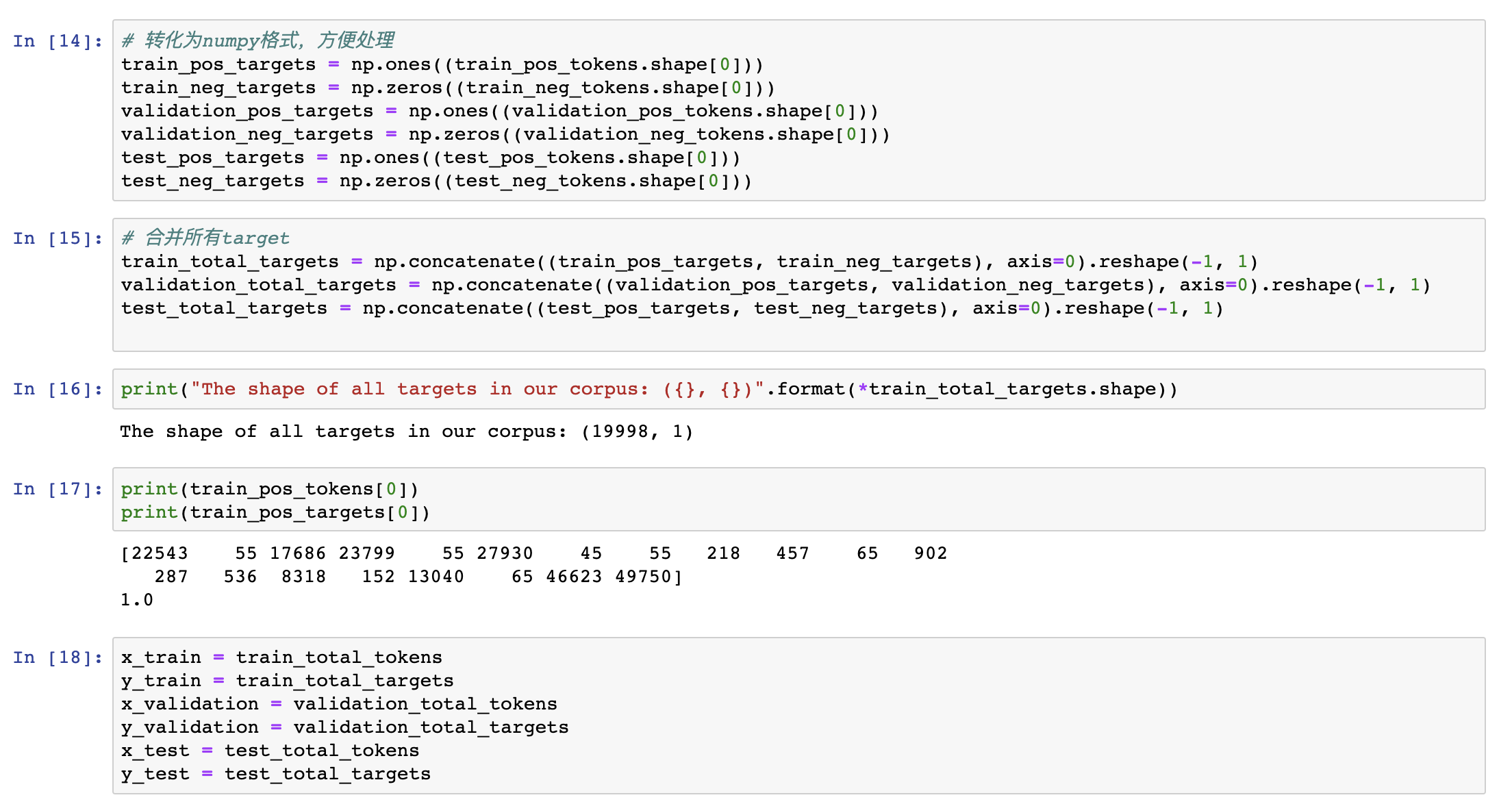
原文本转换成id



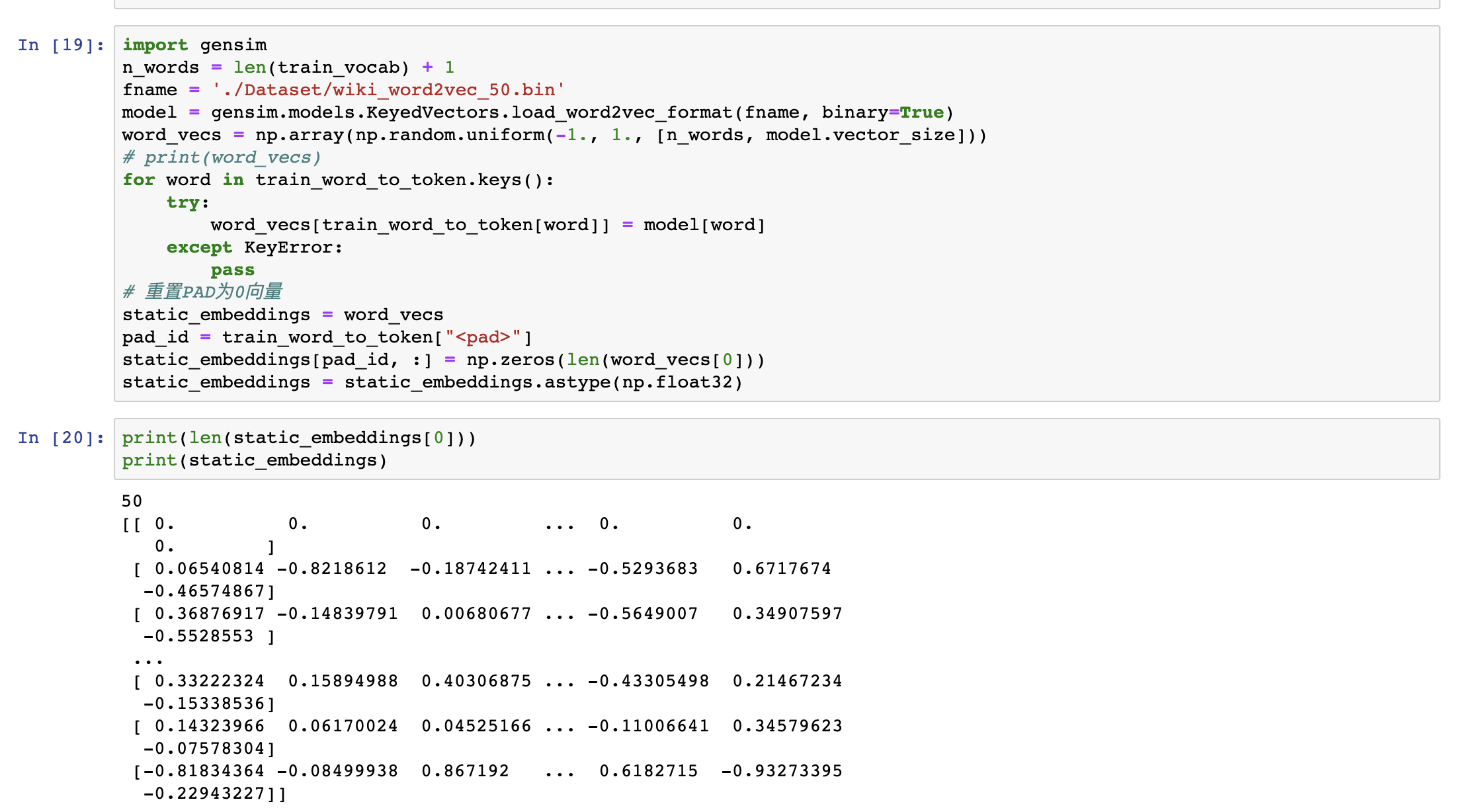
生成训练数据

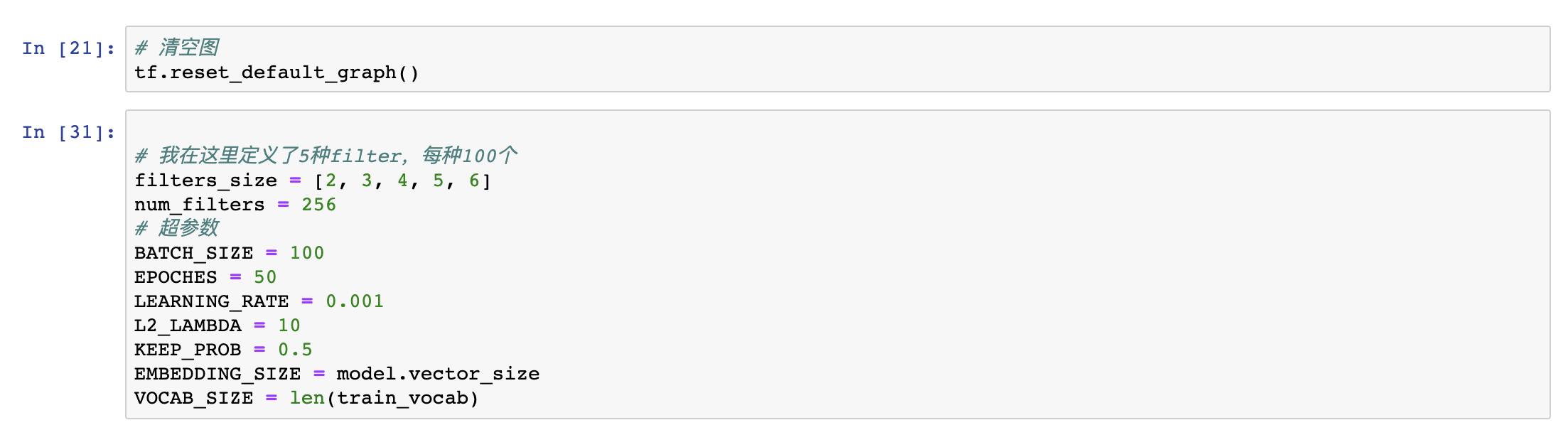


生成训练标签

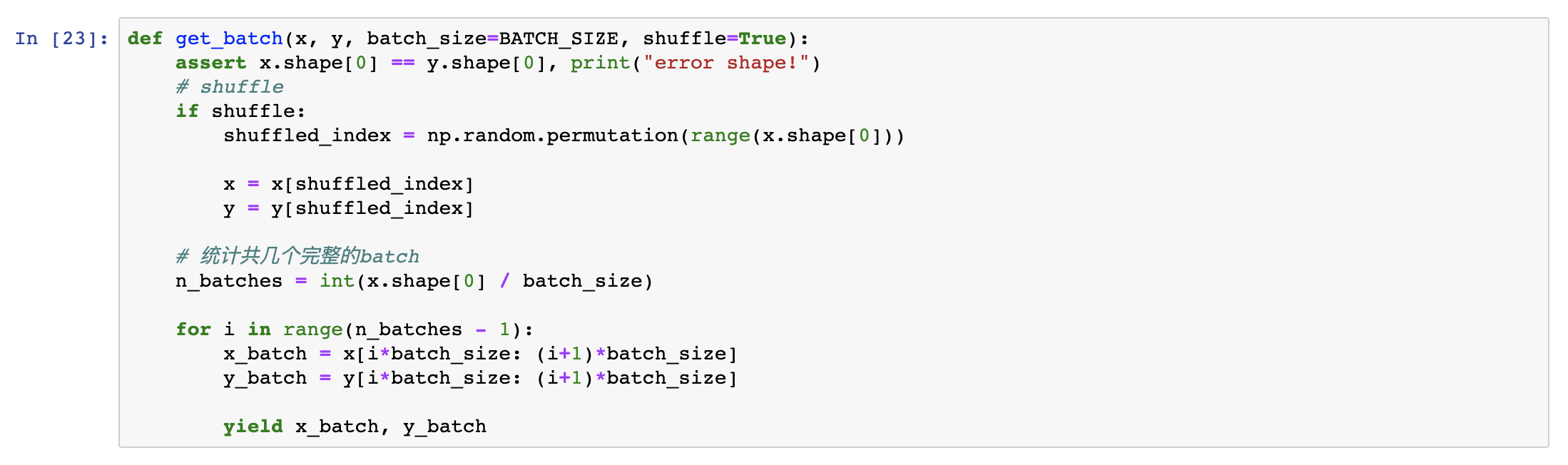


读取词向量

  
设置超参数

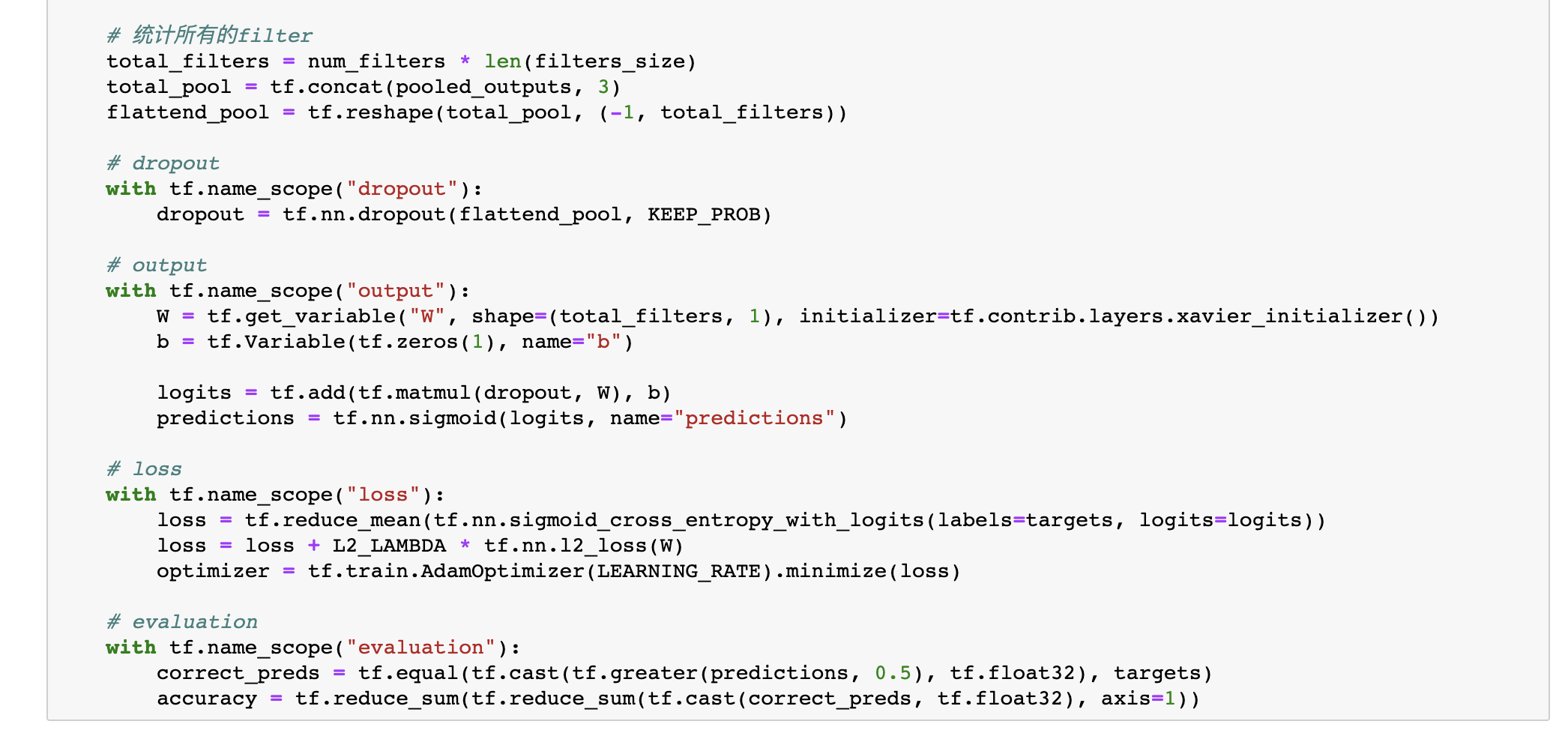


设置批量读取函数



定义TextCNN模型





模型训练和存储，并进行测试



实验结果

