使用transformer进行写对联

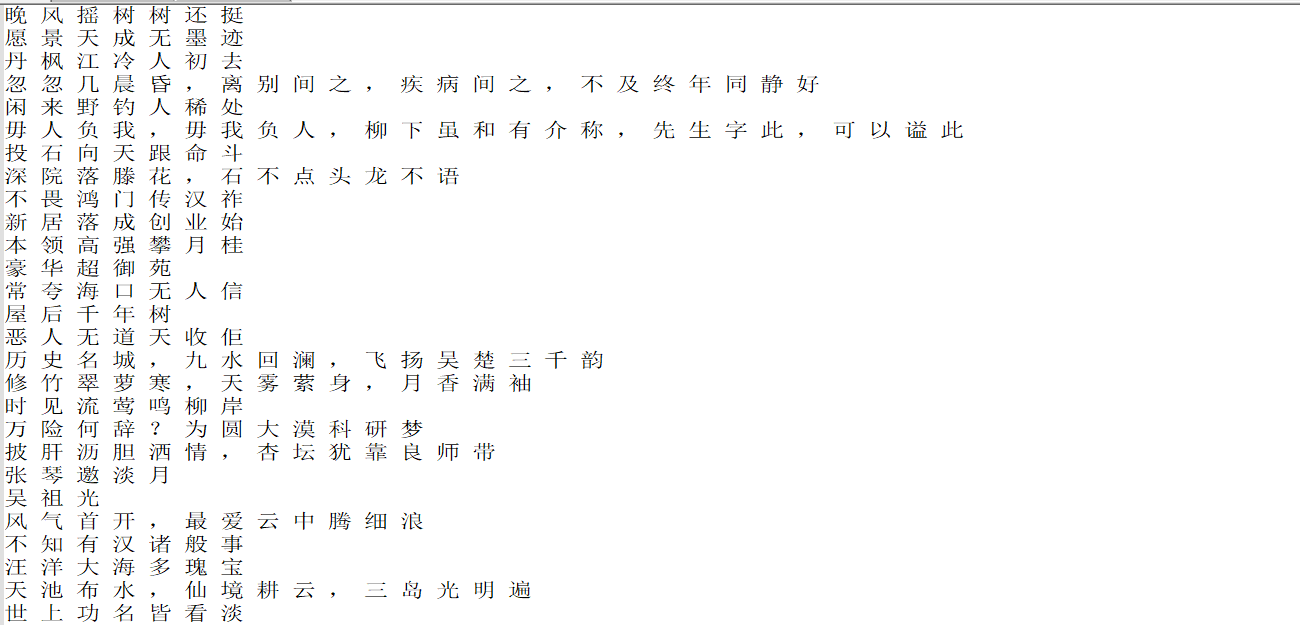
2201785 傅世康

1. 搜集数据：

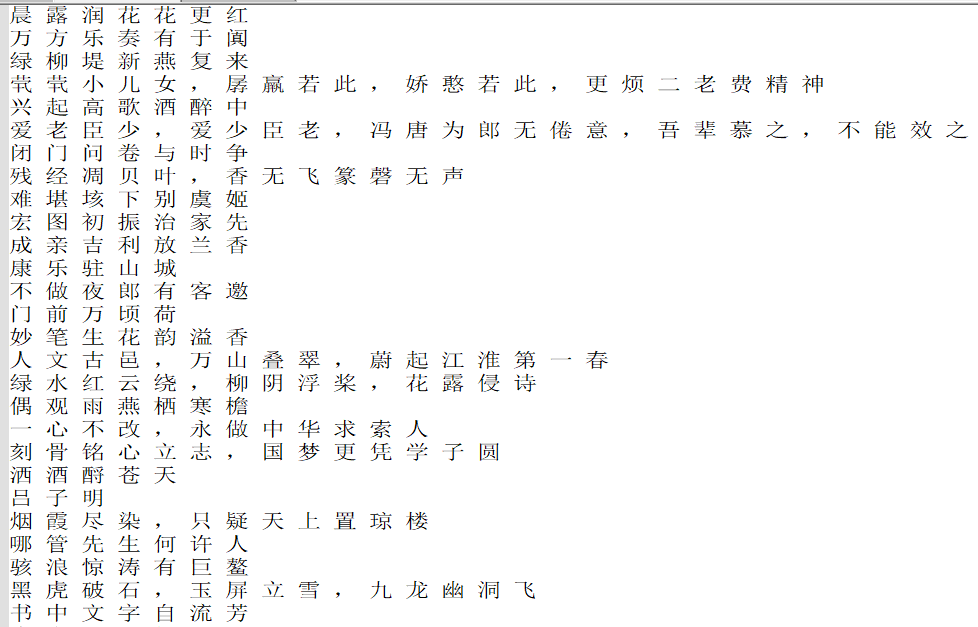
从<https://github.com/wb14123/couplet-dataset>下载的数据：

这个数据集包含超过700,000 条对联。

这是上联：

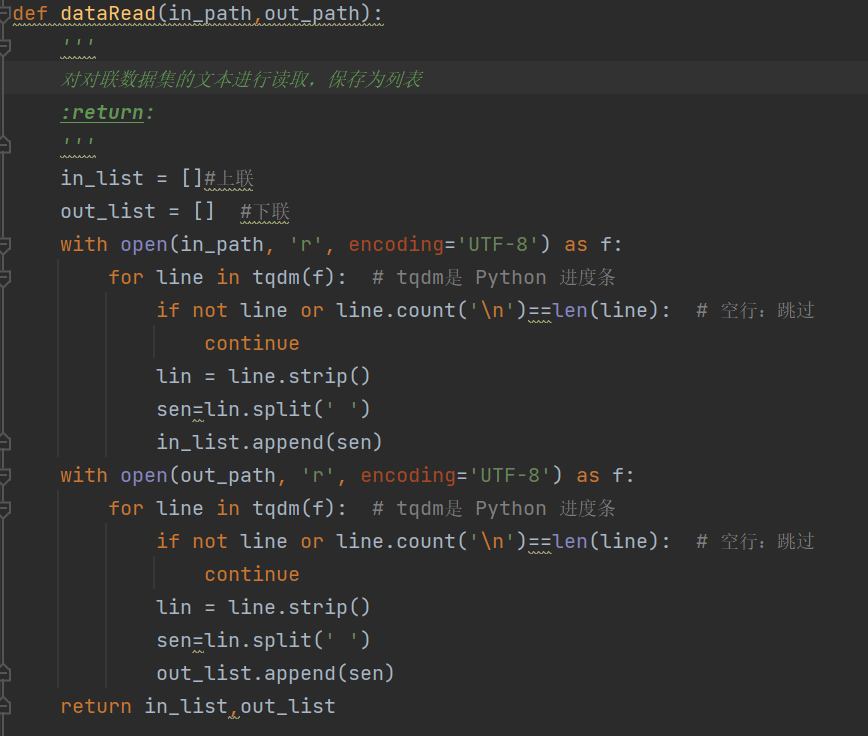


这是下联：

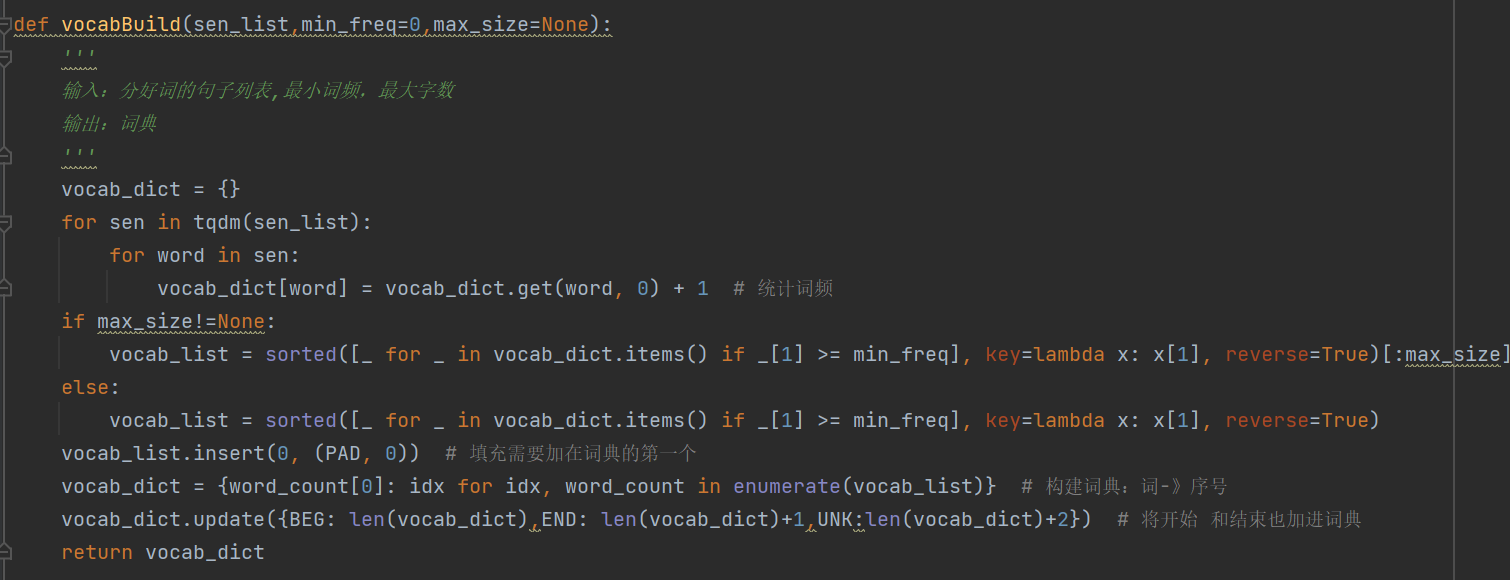


二.数据处理

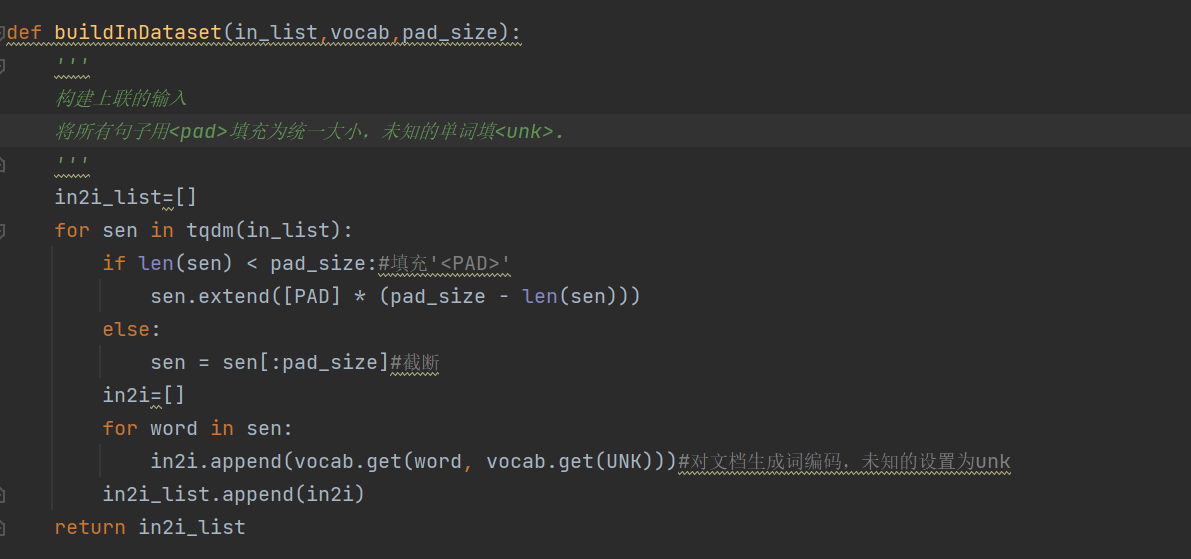
观察数据集发现，数据集由中文汉字和标点符号构成，所以不需要去停用词。



首先读取对联数据集，每一行的对联存到一个列表中，每个字转化为列表中的一个元素，所有的对联组成一个大列表。



然后是构建字典，键是单词，值是单词对应的索引。设置了最小词频和最大单词数，如果词表太长可以去除出现频率很少的单词减少词表长度。而本项目中词表长度不超过10000个单词，而且为了最后的效果更好，因此无需设置最小词频和最大单词数。



构建上联的输入。Transformer模型要求输入是统一的大小，这样才能成batch训练加快训练速度。将所有句子用<pad>填充为统一大小，未知的单词填<unk>.通过统计最长的对联长32，因此这里的pad\_size设置为32。



构建下联的输入。Transformer模型要求输入是统一的大小，这样才能成batch训练加快训练速度。将所有句子用<pad>填充为统一大小，其中下联部分有两个输入：decoder输入，以<beg>开头，loss的输入，以<end>结尾。



保存数据：将处理好的数据通过pickle模块保存到文件方便读取。

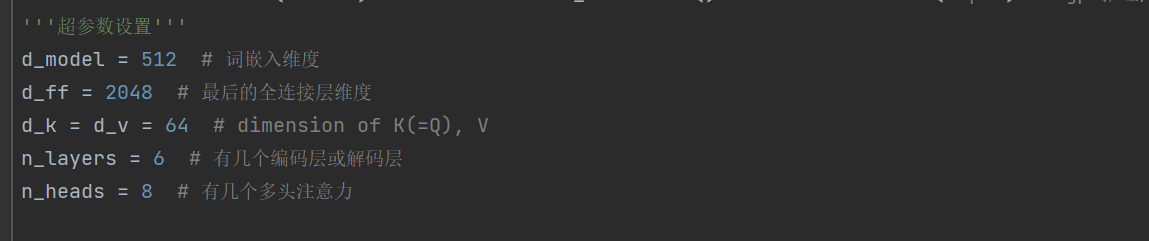


定义了数据集类方便批量读取。

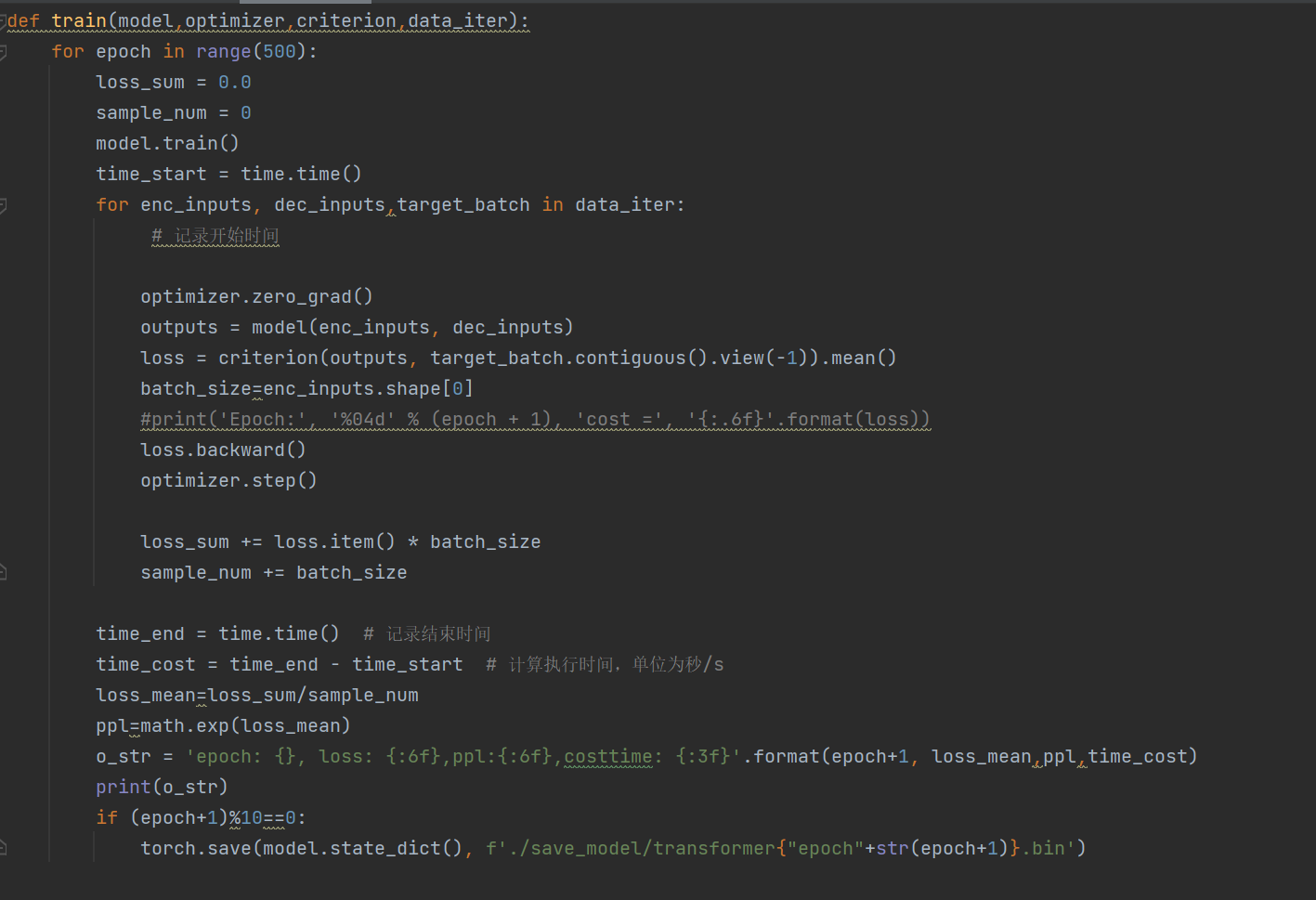
三,构建训练模型

Transformer的代码参照了<https://github.com/graykode/nlp-tutorial>中的Transformer并做了注释和修改，代码详见我的github。

超参数设置如下：

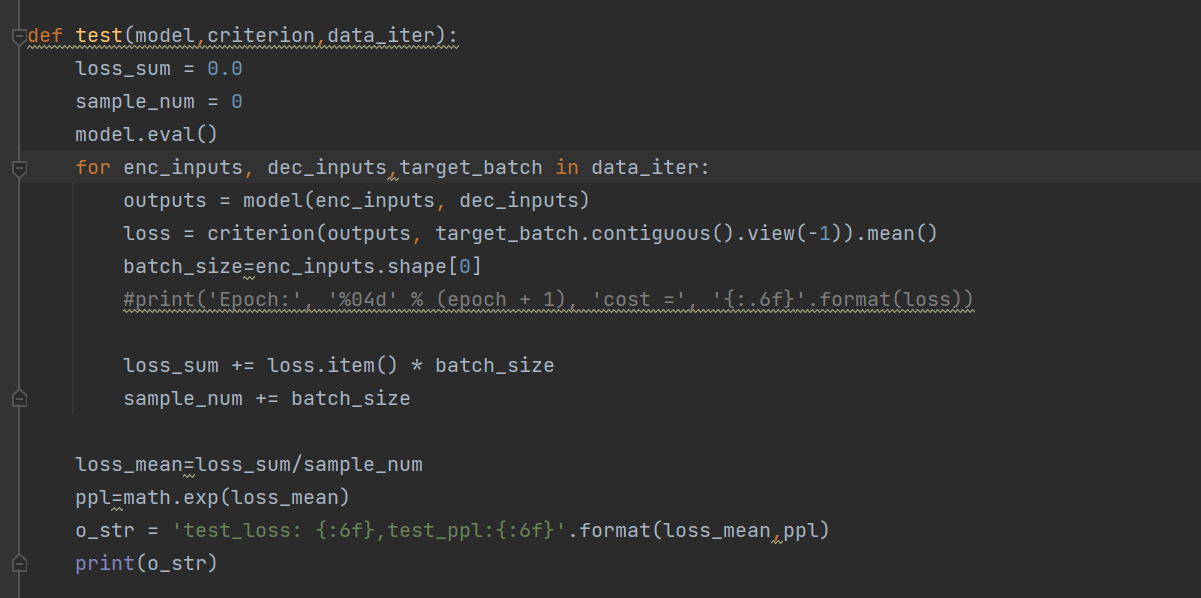


四．模型训练

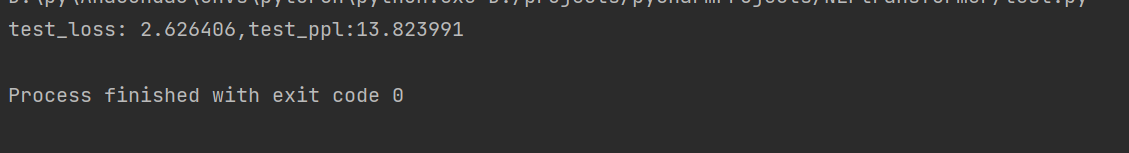


因为机器性能有限，只选取了前10000条对联进行训练，训练了500个周期，花了大约6个小时。

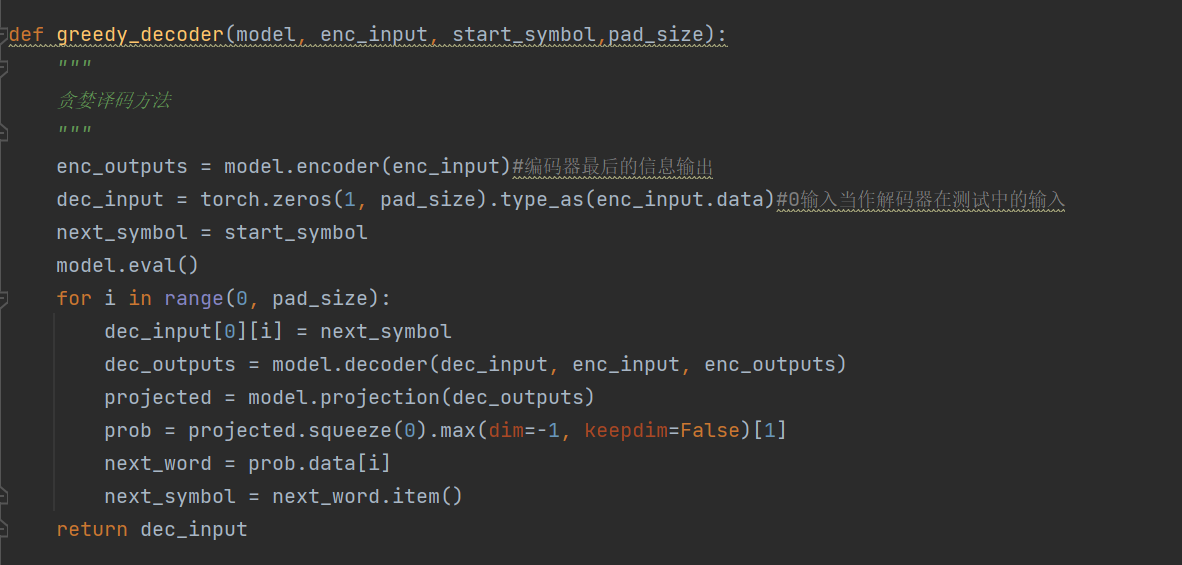
六.模型测试



在测试集中计算loss和困惑度，结果如下：

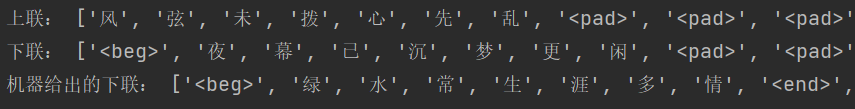


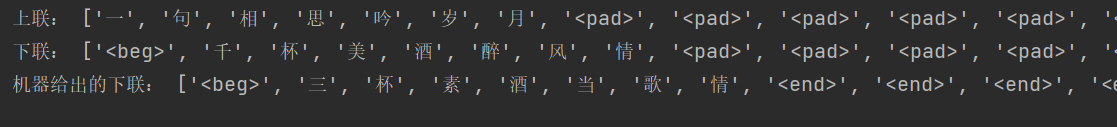
Ppl较高是13，效果似乎不太好。

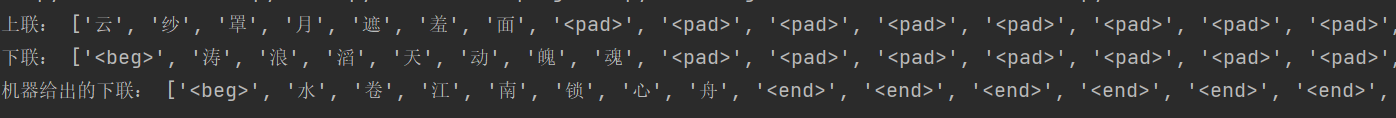


使用贪婪译码方法，

给出几个句子做测试：







可以看出虽然机器给出的对联和下联相差较大，但是从对联的效果来说，比如字数，名词对名词，数词对数词等方面对的还可以，但是少数字对的不够好。

7.总结：

通过这次作业，我学会了transformer的使用，阅读了transformer模型代码深入理解了transformer的原理，实现了transformer对对联的应用。这次实验也有不足之处：比如由于机器性能的原因，只训练了前10000条对联，训练集数量太少可能导致过拟合。