**华东师范大学数据学院上机实践报告**

|  |  |  |
| --- | --- | --- |
| **课程名称**： 人工智能 | **年级**：大二 | **上机实践成绩**： |
| **指导教师**：罗轶凤 | **姓名**：史继林 |  |
| **上机实践名称**：NLP/CV大作业 | **学号**：10172100347 | **上机实践日期**：20190702 |
| **上机实践编号**： | **组号**： | **上机实践时间**： |

**基于LSTM的唐诗和歌词生成器**

1. **项目选题**

本次选题在自然语言处理、计算机视觉等备选项中选择了自然语言处理，这在前面的project中已有涉及，本次实验意在更深层次地了解算法和模型的数学原理。最终定题为projrct3中已有涉及的LSTM模型，用以生成古诗和歌词。

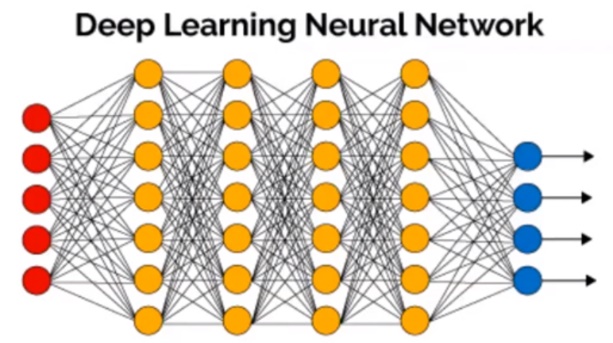
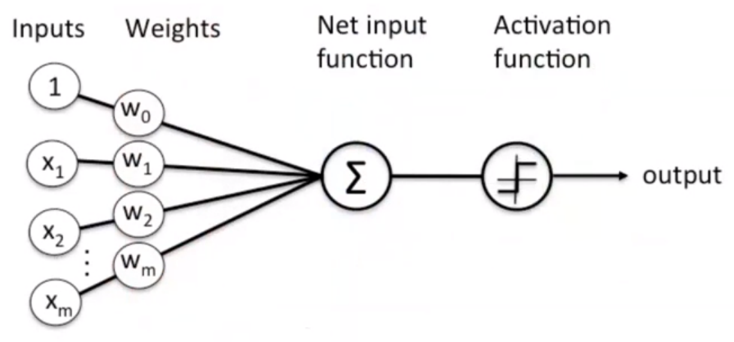
1. **项目思路**

古诗和歌词的特点是：当前字和前字之间是有关系的，对人来说，是词语之间的惯用组合，对于计算机来说，这代表具有序列型。样本间存在顺序关系，每个样本和它之前的样本存在关联。比如说，在文本中，一个词和它前面的词是有关联的；在气象数据中，一天的气温和前几天的气温是有关联的。古诗和歌词的这一特点，决定了在进行神经网络模型选择时， 应保证适用于序列模型。

1. **模型选择**

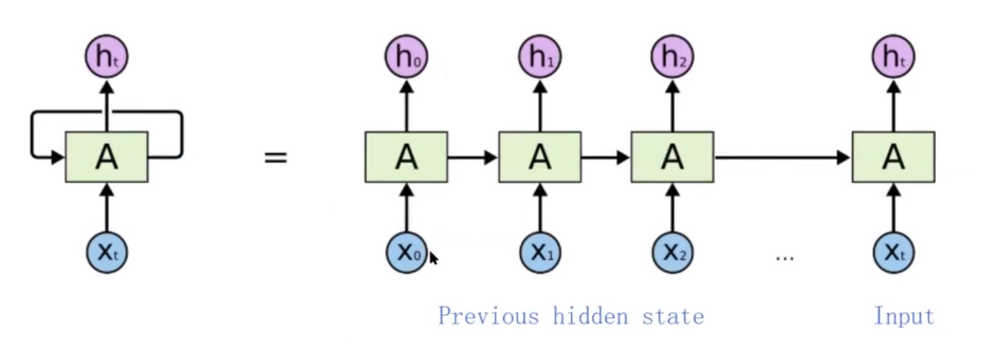
选择LSTM的思路：

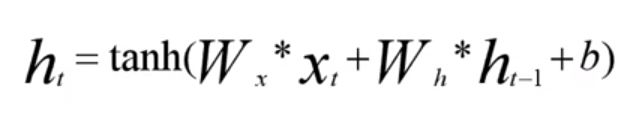
1. **普通深度神经网络**



神经网络本质上是一种通过梯度下降法进行优化的权值网络；CNN是在NN的基础上加入卷积层，自动对输入进行特征提取。通过输入层、隐藏层对输入权重求和计算，最终在输出层输出。

但其缺点很明显：不能处理有时间序列概念的输入，如语音识别、视频、文本生成、经济学指数等。

1. **循环神经网络**



对普通神经网络的初步改进是使用循环神经网络RNN，RNN是为了对序列数据进行建模而产生的。从输入到输出单一传输方向来看，将上一层的输入也作为下一层的一个输入，从而上一节点对下一节点的影响。对应于数学公式中，计算当前层的输出值ht时，除了将输入序列xt作为函数自变量，还应将上一层输入ht-1作为自变量，将二者乘对应处理矩阵并线性组合后，用tanh函数归一化为ht。

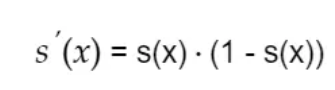
但RNN也有缺点：梯度弥散和梯度爆炸。

梯度弥散原因：

1、Ht是xt和ht-1的二元函数，梯度为一组连乘的积

2、梯度计算公式

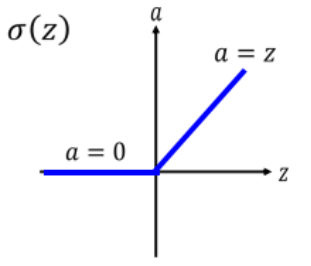
https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D214/sign=d75f12b3ecf81a4c2232ebc8e32b6029/77094b36acaf2edd78a0e92e8a1001e9380193a0.jpg



其值都小于1。

因而t较大，迭代次数较多时，梯度趋于0，这就是梯度弥散。当梯度弥散后，参数更新慢，效率极低，效果甚至类似随机输出。

而梯度爆炸的一个原因是，为了解决梯度弥散，我们使用ReLU等激活函数，但其值可到无穷大，因此梯度也可能过大。

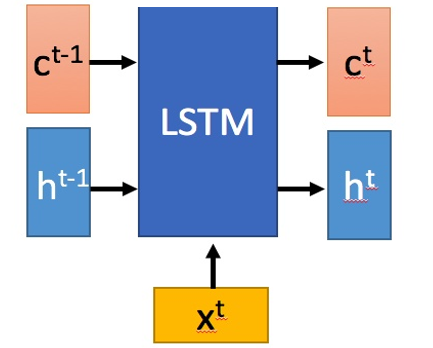
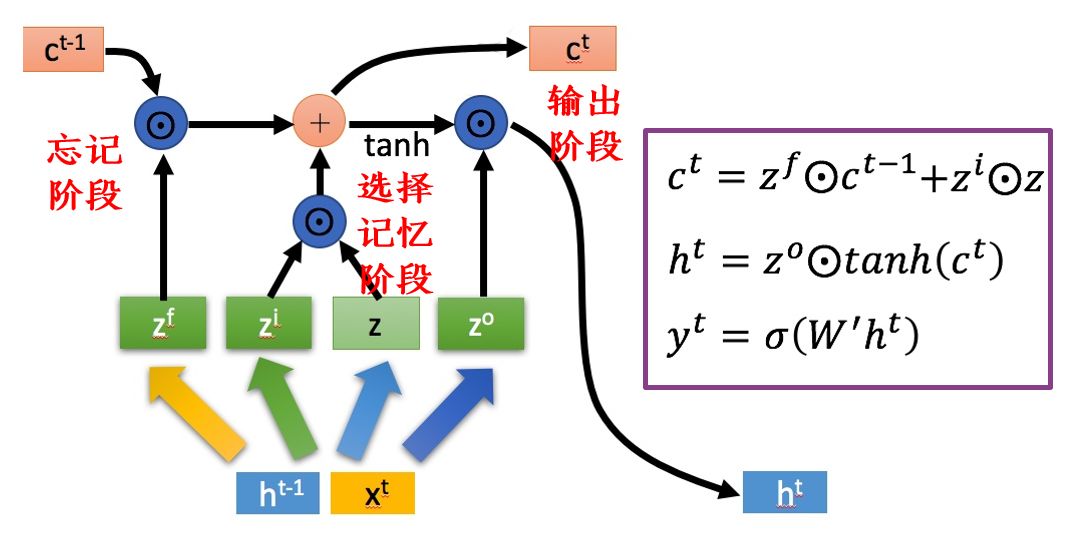


所以只是理论上可以记忆任意长的序列。不能处理long sequence/timeseries问题。网络几乎不可训练。

1. LSTM

LSTM 具有类似的链式结构，但重复模块具有不同的结构。不是一个单独的神经网络层，而是四个，并且以非常特殊的方式进行交互。

ct是cell state中的内容，用来保存节点传递下来的数据。每次传递会对某些维度进行“忘记”并且会加入当前节点所包含的内容，改变相对较小。

ht用来和当前输入组合以获得门控信号，对于不同的当前输入，传递给下一个状态的ht改变较大。

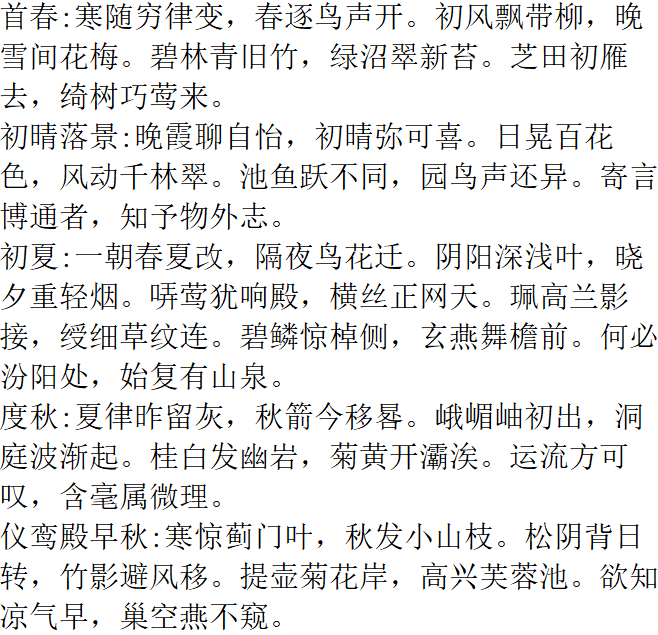
上一阶段的节点数据输出ct-1和当前输入矩阵z共同经变换矩阵运算相加后得到当前层的输出ct，分别对应忘记阶段和选择记忆阶段。

当前层的门控信号由正则化的输出矩阵得到。

最终输出则由ht变化而来。

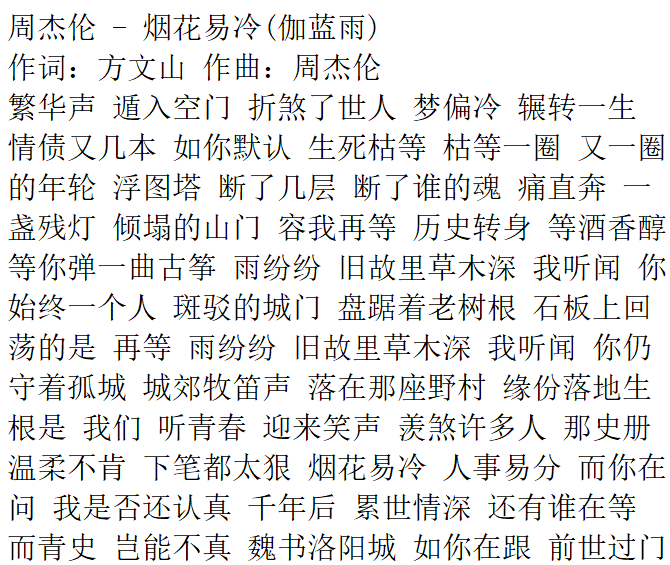
为了让不同时序的记忆对当前影响变得可控，LSTM引入了输入门和输出门，之后又有人对LSTM进行了扩展，引入了遗忘门，通过几个门的结果叠加，LSTM对记忆的操作是相加的，线性的，因此在求梯度也就是求导时，会涉及到几个积的相加，这在数学意义上解决了连乘趋于0的窘境，因而解决了梯度弥散。

1. **数据集介绍**
2. **全唐诗**



|  |  |
| --- | --- |
| 数量 | 4.8w |
| 格式 | 标题：内容 |
| 来源 | Github开源  爬取自郑州大学图书馆 |

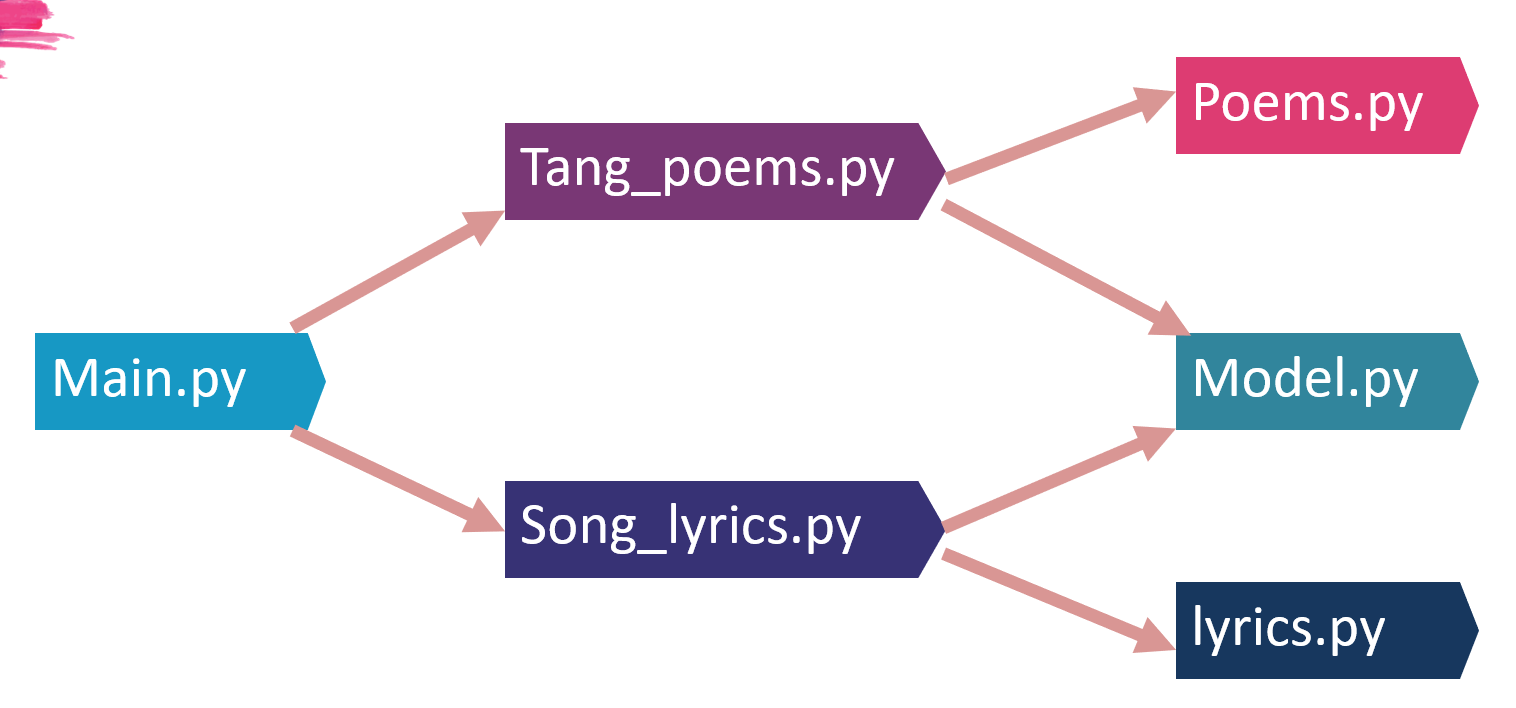
**2、歌词**



|  |  |
| --- | --- |
| 数量 | 300首 |
| 格式 | 歌手-歌名-词曲-歌词 |
| 来源 | Github开源 |



1. **代码实现**



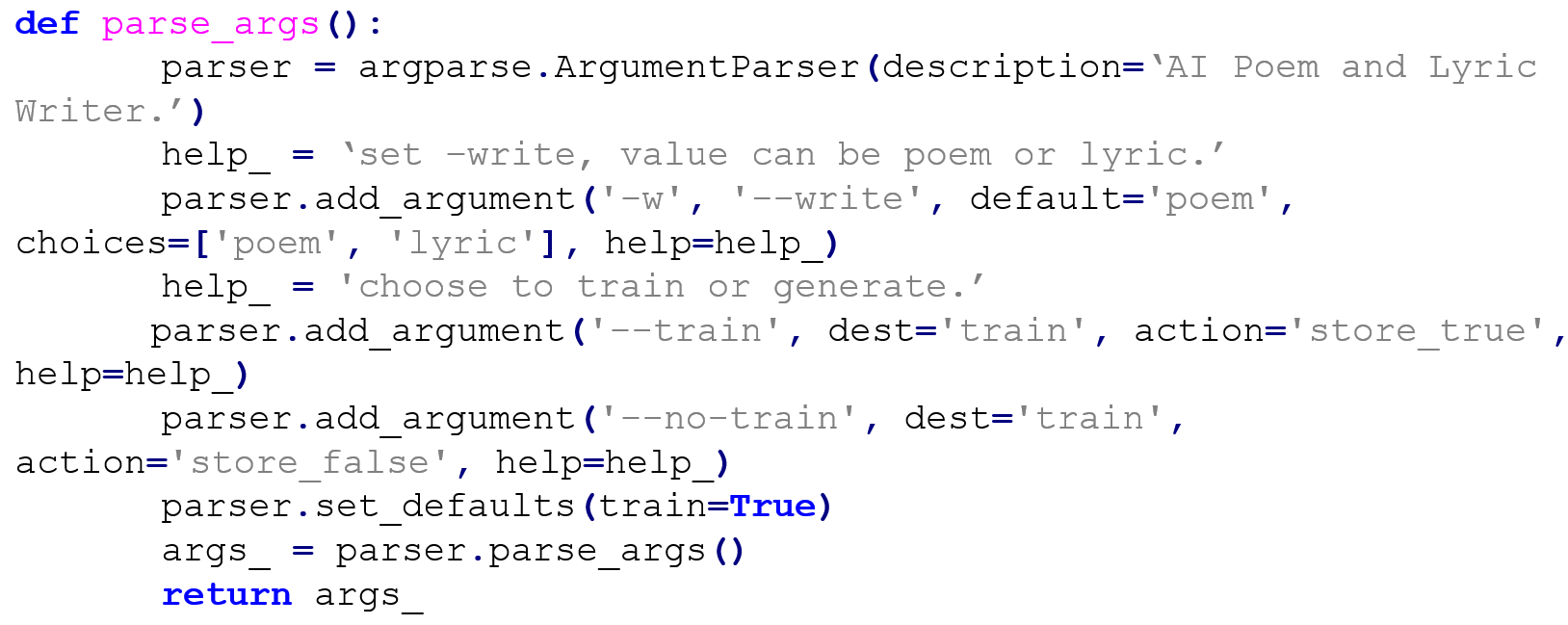
1、Main.py可以视为具有导航性的脚本文件。在main中实现了对训练或测试文本对象类别的选择和训练与测试的功能选择。

2、Tang\_poems.py和Song\_lyrics.py实现了训练模型和应用模型进行文本生成。另外，在训练过程中对中间模型进行了保存。

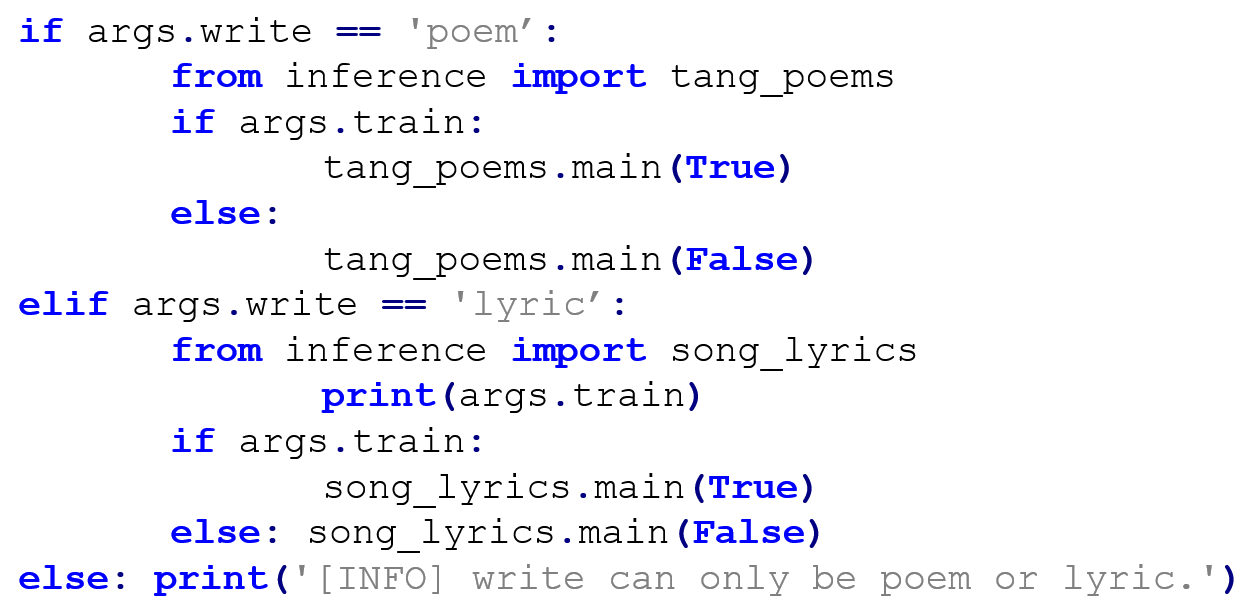
3、Poems.py和Lyrics.py实现对数据集的预处理和对batch的制作生成。

4、Model.py是对模型参数的计算定义以及模型训练后参数的保存。

①main

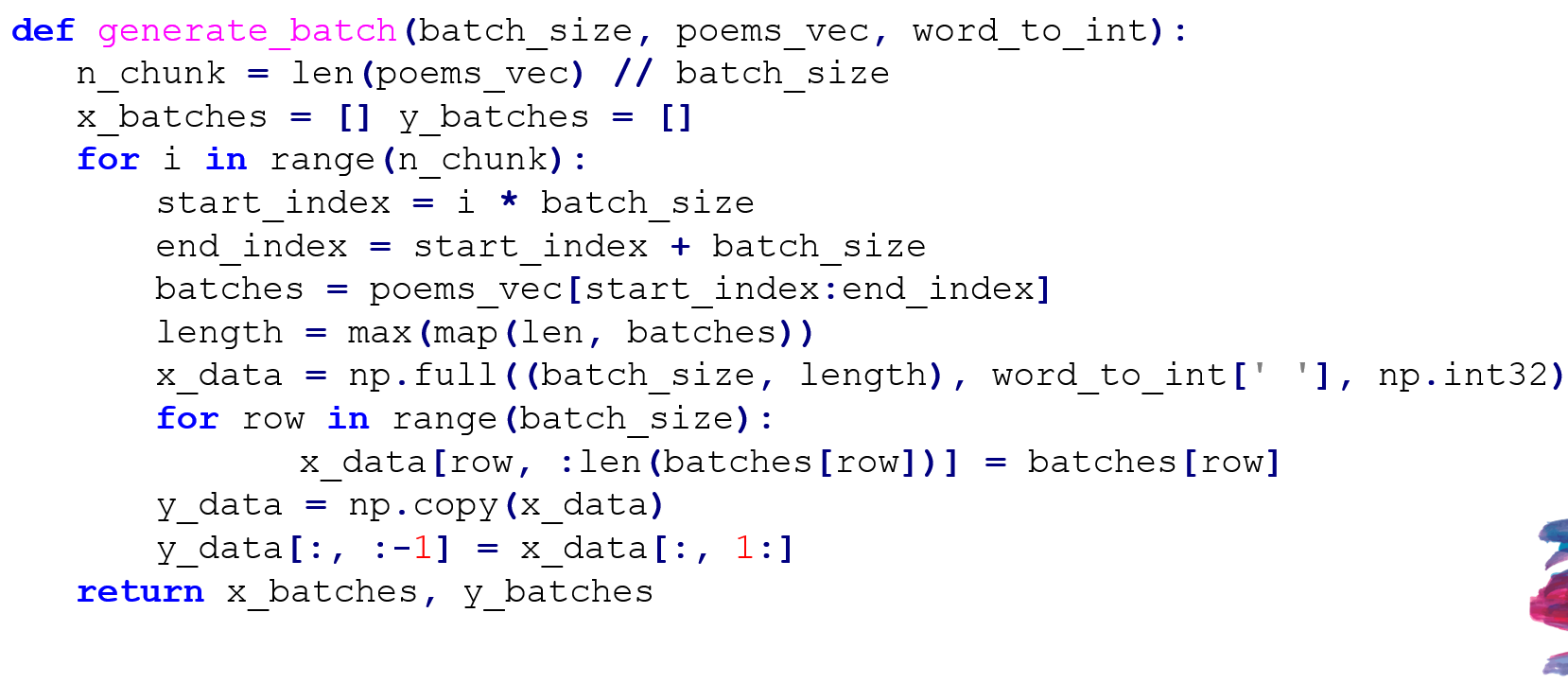
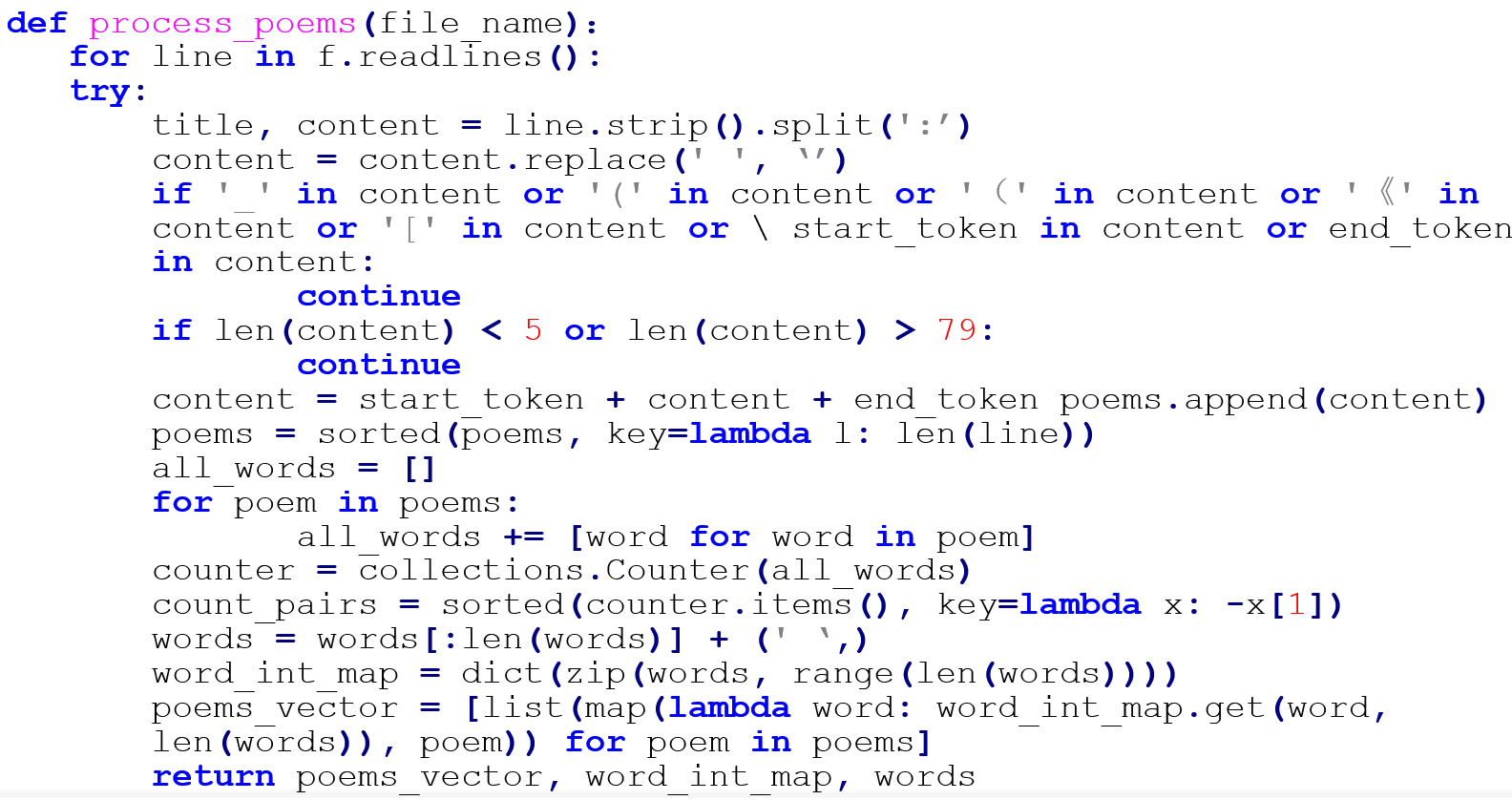


在parse中定义argument为可选参数，分别是唐诗/歌词，训练/测试。



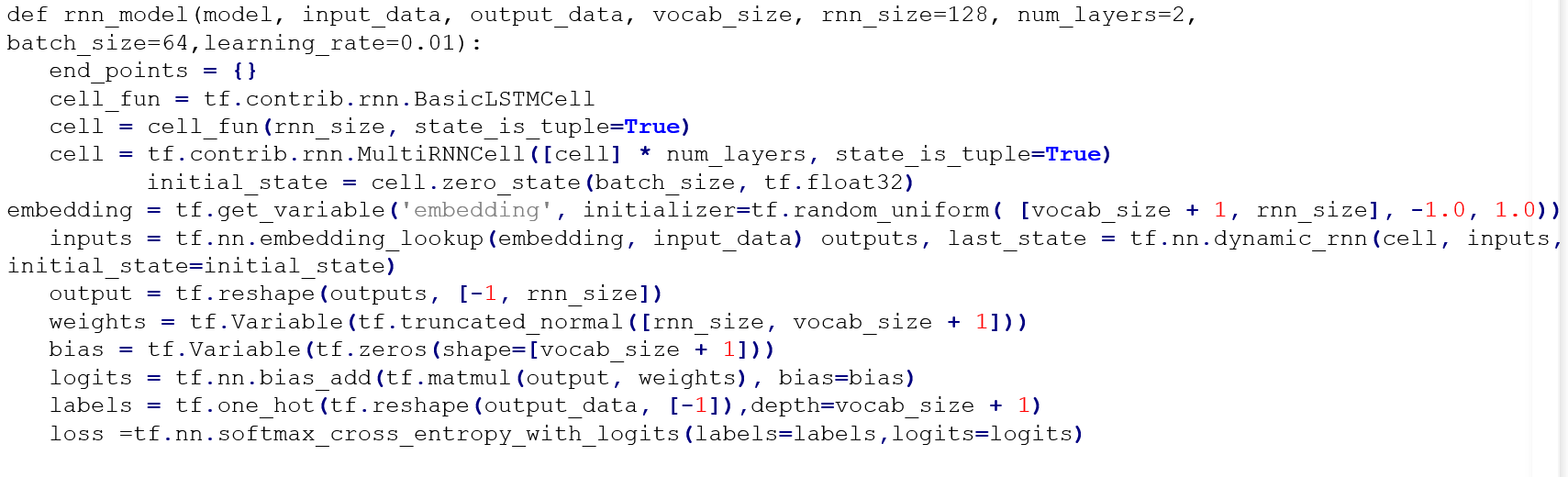
对于不同的参数，调用tang\_poems和song\_lyrics。

②poem



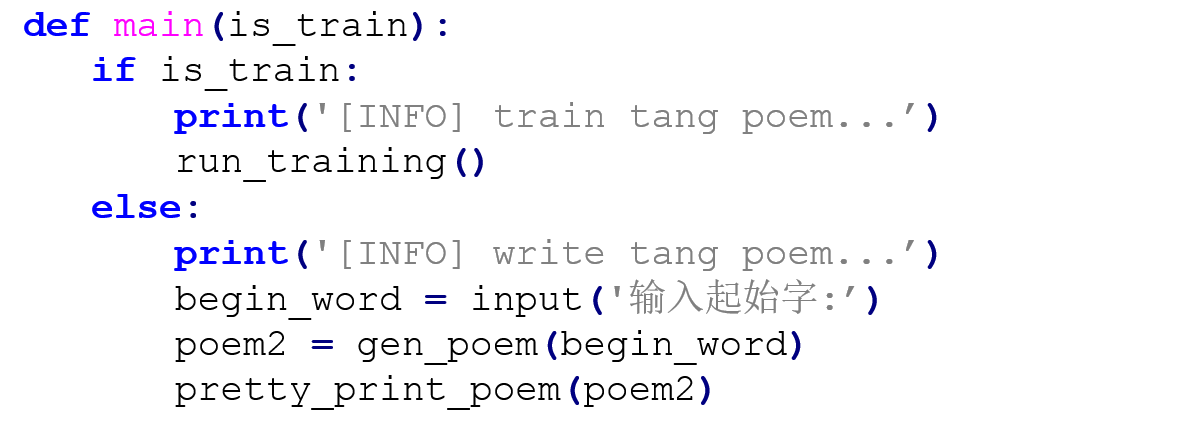
1. 处理数据，包括删除标题，删除字符，删除过长或过短的数据，对单字向量化
2. 生成batch，包括计算每个开始结束下标值，计算长度，生成X和Y数据

③model

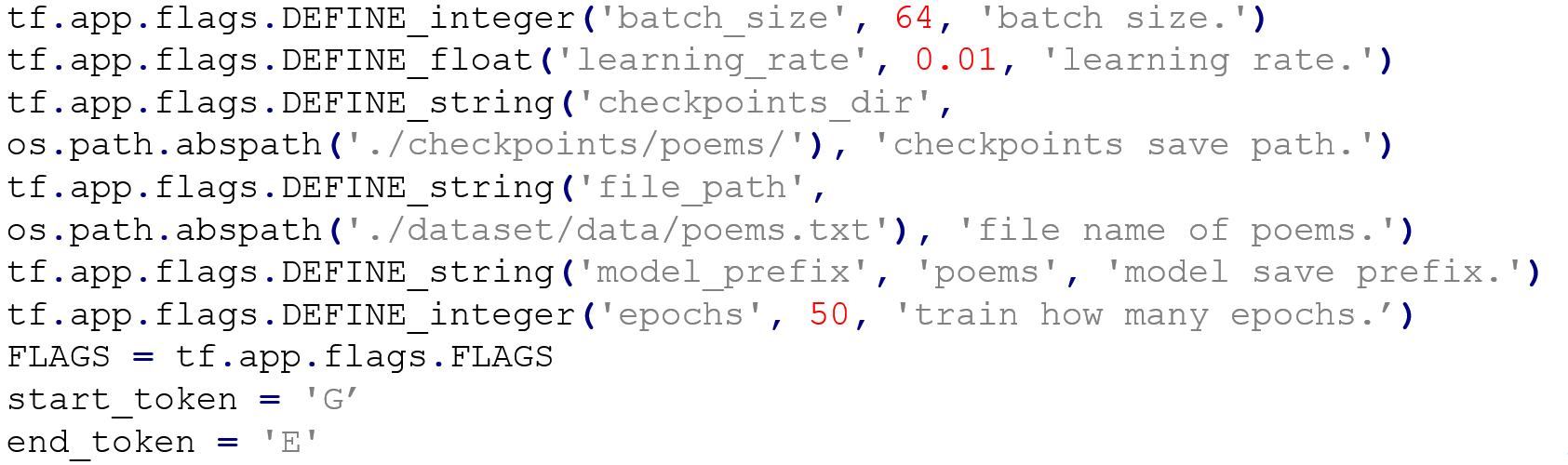


定义cell，定义网络，把输入向量转换成128维向量，当前输入是词的一部分。lookup找到属于哪个词，Output和last是输出 dynamic是官方推荐的执行方式。

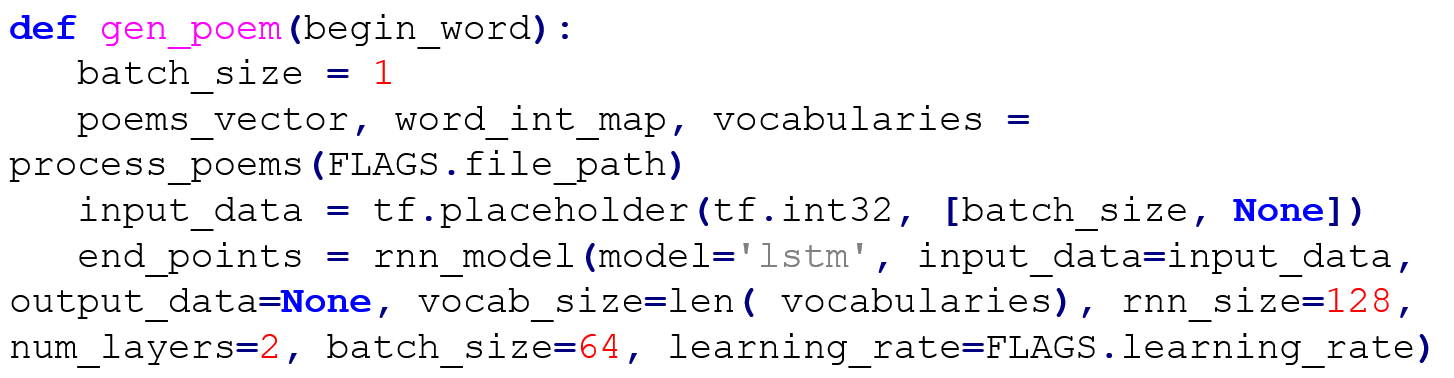
④tang\_poem



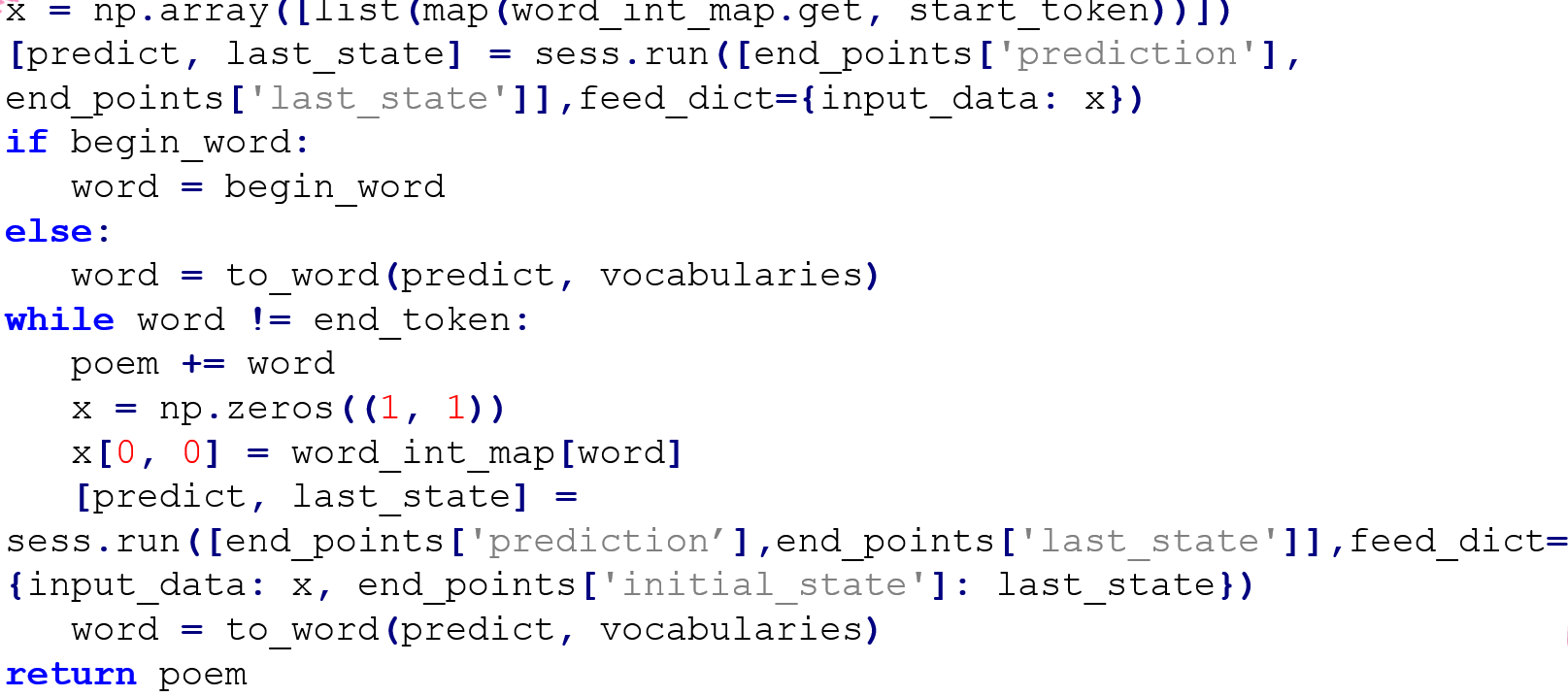
总体框架为：若训练，则调用run\_training，否则调用gen\_poem生成唐诗



Batch大小取为64，学习率取为常规的0.01，设定中间结果保存路径，设置epoch为50。同时以G表示数据开始，以E表示数据结束，在预测中预测到E则停止预测。



Rnn\_model用以运行模型，向量变量和词变量由预处理得到。



通过输入的第一个字，每次生成下一个字，知道预测值为E时退出。

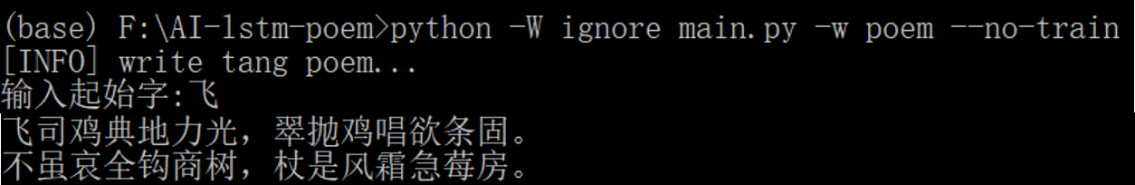
1. **项目效果**

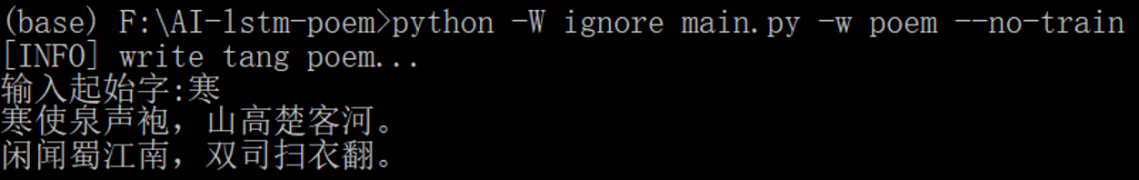
可在命令行中进行功能选择：

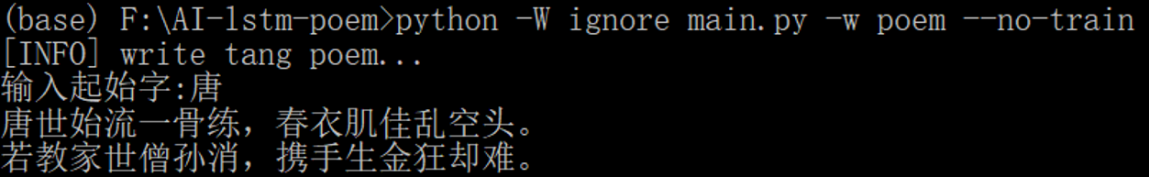
1、-W ignore忽略warning

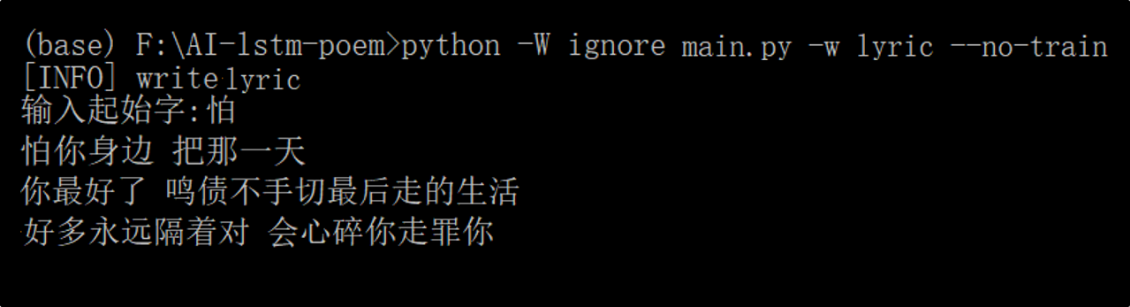
2、poem/lyric选择生成唐诗或歌词

3、train和no-train选择训练模型或测试模型

****

****

****



1. **未来展望**
2. 生成古诗没有用seq2seq，因此没有押韵和词语之间的常用搭配。
3. 生成歌词长短不确定，且含义不明，效果劣于古诗生成。
4. 比较有意思的点可以用于生成一定风格的代码，以假乱真。
5. 主要目的在于理解数学原理，而不是简单的clone项目简单调库。