**# Tong-Music by using LSTM and GPT2 from Morizeyao/GPT2-Chinese**

**## 多少个深夜无法入眠，思念，是那么的甜，那么的温暖。想你的每一次偶然，泪流满面，思念如潮水漫过心田。**

**## Description**

- 中文的GPT-2训练代码（感谢[Morizeyao]（ https://github.com/Morizeyao ）的贡献，可以写诗，新闻，小说，或是训练通用语言模型）。

- 十分感谢[hughqiu](https://github.com/hughqiu "hughqiu")的散文模型。

- 中文的LSTM押韵处理器，其中配有马尔科夫链可自主生成文本。

- 本程序主用于歌词和说唱歌词的生成。

**## 项目状态**

- 本人第一次使用github，此项目也为高中时期练手所用，难免会有很多暴力（朴素）的地方，请谅解。

- 一年内会持续更新填坑。

**## demo：《我忘不了》**

我忘不了她，

忘不了她对我的爱，

她在我的心里生根发芽，

开出一朵美丽的花来。

我想，

我真的不是这样，

我们之间没有什么可以沟通彼此心的桥梁，

那美丽的传说真是令人难忘。

我忘不了她的存在，

忘不了她在的日子里，

我的心里从此刻沉重起来，

我不能接受这个现实。

我是有点无奈的，

我只是希望这样的日子能有个可以倾诉的角色，

让你知道我的存在是多么有压抑的选择，

也许我的努力就够了。

**## 使用方法**

**### LSTM部分**

- 若使用自己的歌曲库可修改chinese\_lyrics.txt。\*\*chinese\_lyrics.txt是个txt文件，歌词一句一行以'\n'为分割。

- 若使用自己的歌曲库需清空demo.txt和chinese\_rappers.rap两个文件。

- Chinese\_lyrics\_flow.py训练时要将程序第19行train\_mode改为True，生成时要改为False，直接运行即可。

- 然后打开flow.py可自动匹配出押韵方式————也可自己设置长度————然后复制print出的ans结果。

**### GPT-2部分**

- 可参照 https://github.com/Morizeyao/GPT2-Chinese 自行配置，这里使用的是网友训练好的散文模型，可使课词更富有诗情画意。

- 这里已经不需要再重新训练。

- 根据复制的print出ans的结果修改generate\_with\_flow.py第10行style \*\*style是指接下来的文本押韵方式。

- 也可以根据自己喜欢的风格修改style。

- 修改--prefix为自己想以之为开头的主题直接运行，即可自行生成文本。

**## 文件结构**

- pycache文件夹为马尔科夫链生成的模型参数。

- cache文件夹为GPT-2语料库。

- config文件夹中贮存GPT-2模型基本参数。

- scripts, tokenizations文件为GPT-2配套文件。

- chinese\_lyrics.txt为中文说唱曲库。

- Chinese\_lyrics\_flow.py为使用keras搭建的LSTM的中文说唱押韵模型，也可以直接生成说唱，只不过前后文内容衔接效果欠佳。

- chinese\_rappers.rap储存LSTM参数

- demo.txt为Chinese\_lyrics\_flow.py的说唱曲目

- demo\_我忘不了.txt为使用generate\_with\_flow.py生成的说唱歌曲这里使用了调押和单押的方式，将几个生成文件合并之后的展示文件。

- eval.py为GPT-2配套文件

- flow.py用于使demo文件生成对应的押韵style

- generate.py为原先GPT-2生成文件

- generate\_texts.py也为原先GPT-2生成文件

- generate\_with\_flow.py为正式的修改后的生成文件

- markov\_speaking.py为马尔科夫链生成文本文件

- train.json为GPT-2训练文本的样式

- train.py用于训练GPT-2

- train\_single.py也用于训练GPT-2

**## 注意**

- Chinese\_lyrics\_flow.py训练时时间可能较长，若时间紧迫可修改其中epoch和长度参数减少训练和生成时间。

- 若要自己尝试训练GPT-2数据量较少时可以考虑自己建立语料库建立方法即点击cache里的py文件即可。

- style长度建议不长于12。

- generate\_with\_flow.py生成可能会较慢，这是因为电脑词穷了，可以通过更改topk参数来改变随机性，改变生成的速度。

- 若自己准备歌曲数据转换为chinese\_lyrics.txt时注意每一行最后一个字后除了'\n'不要有任何字符。

**## model文件下载地址**

|  模型名称 |   模型介绍|   分享者|  链接地址1 |  链接地址2 |

| ------------ | ------------ | ------------ | ------------ | ------------ |

| 散文模型  | 使用130MB的名家散文、情感散文和散文诗歌训练所得 。  |  [hughqiu](https://github.com/hughqiu "hughqiu") | [百度网盘【fpyu】](https://pan.baidu.com/s/1nbrW5iw34GRhoTin8uU2tQ)   | [GDrive](https://drive.google.com/drive/folders/1rJC4niJKMVwixUQkuL9k5teLRnEYTmUf?usp=sharing "GDrive") |

**## 联系作者**

微信：158109871775

邮箱：jianyq2003@sina.com