

# GPT2-Chinese 模型训练

## 1. 训练过程

### 1.1 小说部分

#### 1.1.1 语料

- [斗破苍穹](#)
- 试了一下古诗词，但是语料过少，导致生成结果乱码
- 训练集大小：8w+ 行，共16.1M

#### 1.1.2 训练

- 预训练 - 使用默认循环次数

```
1 # 开始训练命令
2 python train.py --raw --raw_data_path='./data/doupo.txt'
```

- 默认循环次数为 5
- 修改 raw\_data\_path 路径为其他 默认 json 文件 "不能用" - 我也不知道为什么

```
1 # 模型中给出的代码：
2 lines = json.load(f)
3 # 修改后的代码：
4 lines = f.readlines()
```

- 生成结果：

```
1 # 生成文本命令
2 python ./generate.py --length=500 --nsamples=4 --prefix=开始 --
  fast_pattern --save_samples --save_samples_path=./generate
```

```
===== SAMPLE 1 =====
xxx, 风雨无情。天上云台人未老，一醉春雨洗愁颜。
一曲春光，千条绿柳，春色如花。柳絮莺啼，柳条花影，莺啼柳外，人倚东西。东风吹梦醒来。问年华过了，何事匆匆。芳草萋萋，青山寂寂，无计芳菲。无聊可、花开又落，不管芳时。无限春愁，春归何处，有人肠断，有谁知道，何处春风。
天道路，问花落花无主。春到也无情，春去无踪，只在春愁无语。芳草碧，正一抹红楼，春归无路。花落春江，柳垂春住，不是春光何处。春去去来
100%|===== 200/200 [00:01:00:00, 146.15it/s] =====
===== SAMPLE 2 =====
xxx, 天下有佳气，[UNK][UNK]如可人。不如天地间，万事一销伸。[UNK][UNK]有如此，不可一日真。不然天下无，何必不吾身。
我本山中，[UNK]然有一事。一日无一言，不知何与累。今晨出山去，已复见其次。不见山亦[UNK]，不觉山可爱。有田无不种，有田亦种菜。人生无足道，此是无所用。不如种禾者，种麦亦已种。不说田家人，不如田会上。有田不可种，可种禾与黍。不见田家田，不如耕与耨。田夫耕且耕，不耕亦已耕。田田种麦田，种禾不可耨。田夫耕，
100%|===== 200/200 [00:01:00:00, 126.49it/s] =====
===== SAMPLE 3 =====
xxx, 我欲求之[UNK]，何如问之天。
一从[UNK][UNK]兹，十载不自由。君如不见君，君心如秋水。一身如[UNK][UNK]，一身如浮[UNK]。一死如死灰，三年如一日。我身本不知，我身安足恃。君心本无营，人死亦有几。不知有心人，自古皆知是。我心本如水，君子如相似。君子慎其然，我心亦何厚。君看汝之子，君其汝其寿。我心如有何，不如汝之后。
吾儒学孔明，有志不可期。有如三年交，有如不可移。不如一日闲，不如千里驰。有时复不见，有如一日期。今
100%|===== 200/200 [00:01:00:00, 146.61it/s] =====
===== SAMPLE 4 =====
xxx, 不能[UNK]其。一朝不来，千万斯千。
春来江南春草长，江南花落后春光。一年一度花已去，不堪回首东流光。春风杨柳春未归，江城春草今何在。春花春去不复知，桃花李花不如此。江南杨柳多情情，江北春风无限情。春来春去春还去，花落花前不可寻。春风吹尽春愁去，江南江北多春光，春来不去无归航。江南水北花如雪，杨花落尽江南色。
春风昨夜春花发，江上花枝不可说。今年春色满楼台，花下莺啼花落尽。不知春去无多情，无限
```

## 2. 问题汇总

### 2.1 文件读取

```
1 # 模型中给出的代码:
2 lines = json.load(f)
3 # 修改后的代码:
4 lines = f.readlines()
```

- **问题描述:** 原模型中使用 train.json 文件进行多个文件的读取, 但是实际只读取了 json 文件本身。
- **解决方法:** 修改默认读取文件
- **出现原因:** 待解决

### 2.2 模型构造参数

- ./GPT2-Chinese/config/model\_config.json
- ./GPT2-Chinese/config/model\_config\_small.json

```
1 # 原始值
2 {
3     "initializer_range": 0.02,
4     "layer_norm_epsilon": 1e-05,
5     "n_ctx": 1024,
6     "n_embd": 768,
7     "n_head": 12,
8     "n_layer": 12,
9     "n_positions": 1024,
10    "vocab_size": 21128
11 }
12
13 #修改值
14 {
15     "initializer_range": 0.02,
16     "layer_norm_epsilon": 1e-05,
17     "n_ctx": 1024,xunlian
18     "n_embd": 768,
19     "n_head": 12,
20     "n_layer": 12,
21     "n_positions": 1024,
22     "vocab_size": 21128
23 }
```

- **问题描述:** 使用原模型中参数, 预训练过程中内存分配不足, 卡死
- **解决方法:** 缩小构造参数值
- **出现原因:** 原模型中构造参数过大, 更深一点的原因正在看一些 GPT2 的文章

## 3. 本周要做的事

- 查找一些 GPT2 相关的文章和论文进行阅读;
- 看一下模型源码;

- 重新找一些其他语料进行训练；
- 继续学习 bert 和 pytorch。

## 补充部分

# 1.2 古诗词部分

## 1.2.1 语料

- [古诗词集](#)
- 训练集大小： 共85w+ 行， 187M

## 1.2.2 训练

- 预训练 -- 使用默认循环次数-5
  - 同上
- 生成结果

```
===== SAMPLE 1 =====
xxx, 风雨无情。天上云台人未老，一杯春雨洗愁醒。

一曲春光，千条绿柳，春色如花。柳絮莺啼，柳条花影，莺啼柳外，人倚东西。东风吹梦醒来。问年华过了，何事匆匆。芳草萋萋，青山寂寞，无计芳菲。无聊可、花开又落，不管芳时。无限春愁，春归何处，有人肠断，有谁知道，何处春风。

天遥路，问花落花无主。春到也无情，春去无踪，只在春愁无语。芳草碧，正一抹红楼，春归无路。花落春江，柳垂春住，不是春光何处。春去无踪
100%| 200/200 [00:01<00:00, 146.151t/s]
===== SAMPLE 2 =====

xxx. 天下有佳气，[UNK][UNK]如可人。不如天地间，万事一朝伸。[UNK][UNK]有如此，不可一日真。不然天下无，何必不吾身。

我本山中人，[UNK]然有一事。一日无一言，不知何与累。今晨出山去，已复见其次。不见山亦[UNK]，不觉山可爱。有田无不种，有田亦种菜。人生无足道，此是无所用。不如种禾者，种麦亦已种。不识田家人，不如田舍上。有田不可锄，可种禾与麦。不见田家田，不如耕与耨。田夫耕且耕，不耕亦已耕。田田种麦田，种禾不可耨。田夫耕，
100%| 200/200 [00:01<00:00, 126.491t/s]
===== SAMPLE 3 =====

xxx. 我欲求之[UNK]，何如问之天。

一从[UNK][UNK]去，十载不自由。君如不见君，君心如秋水。一身如[UNK][UNK]，一身如浮[UNK]。一死如死灰，三年如一日。我身本不知，我身安足恃。君心本无苦，人死亦有几。不如有心人，自古皆知是。我心本如水，君子如相似。君子俱其然，我心亦何厚。君看汝之子，君其汝其寿。我心知有何，不如汝之后。

吾尝学孔明，有志不可期。有如三年交，有如不可移。不如一日闲，不如千里驰。有时复不见，有如一日期。今
100%| 200/200 [00:01<00:00, 146.611t/s]
===== SAMPLE 4 =====

xxx, 不能[UNK]其。一朝不来，千万斯千。

我来江南春草长，江南花落春光长。一年一度花已去，不堪回首东流长。春风杨柳春未归，江城春草今何在。春花春去不复知，桃花李花不如此。江南杨柳多情情，江北春风无限情。春来春去春还去。花落花前不可寻，春风吹尽春愁去。江南江北多春光，春来不去无归航。江南水北花如雪，梅花落尽江南色。

春风昨夜春花发，江上花枝不可说。今年春色满楼台，花下莺啼花落尽。不知春去无多情，无限
```

## 附录：

```
1 # data_processing.py -- 古诗词数据处理函数
2 import os
3
4
5 def getfilelist(rlist, path):
6     # 获取 path 目录下所有文件
7     for dir, floder, file in os.walk(path):
8         for i in file:
9             t = "%s/%s" % (dir, i)
10            rlist.append(t)
11    return rlist
12
13
14 def save2file(path):
15     # 读取 && 过滤 && 保存 数据
16     print('----- Processing start -----')
17     s = '' # 存放文件中读取的数据
18     number = 0
19     for filename in getfilelist([], './data'):
20         print('Current file number : %d <--> Current file : %s' %
21             (number, filename))
22         number += 1
```

```
23         with open(filename, 'r', encoding='utf8') as f:
24             next(f) # 过滤掉第一行无用数据
25             for line in f.readlines():
26                 line = line.replace('"', '') # 过滤掉双引号 "
27                 s += line.split(',')[0] # 获取古诗句 过滤掉取用数据
28         with open(path, 'w') as f:
29             f.write('%s' % s) # 读取数据写入文件中
30         print('----- Processing completed -----')
31
32
33 def main():
34     save2file('./poetory.txt')
35
36 if __name__ == "__main__":
37     main()
38
```