

本周进度

1. 使用古诗词语料对 GPT - 2 模型进行训练；
2. 看 GPT - 2 模型源码 -- 部分完成；
3. 看 [NeteaseCloudMusic](#) 代码 -- 爬取网易云音乐数据 -- 部分完成；
4. 看论文 -- 部分完成。

1. GPT - 2 模型训练

1.1 古诗词部分

1.2.1 语料

- [古诗词集](#)
- 训练集大小：共85w+ 行， 187M
 - 预处理

1.2.2 训练

- 预训练 -- 使用默认循环次数-5
 - 同小说预训练
- 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。

天涯路，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%|===== SAMPLE 2 =====| 200/200 [00:01<00:00, 146.151t/s]

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。有田不可确，可种禾与麦。不见田家田，不如耕与稼。田夫耕且耕，不耕亦已耕。田田种麦田，种禾不可植。田夫耕，
100%|===== SAMPLE 3 =====| 200/200 [00:01<00:00, 126.491t/s]

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。君心本无营，人死亦有几。不如有心人，自古皆如是。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。不如一日问，不如千里驰。有时复不见，有如一日期。今
100%|===== SAMPLE 4 =====| 200/200 [00:01<00:00, 146.611t/s]

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，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
```

2. 问题总结

- 爬取网易云音乐数据所使用项目可能爬取不成功 -- 还没来得及实践
 - 项目未给出使用方式，需要看一下源代码；
 - 项目两年前更新，可能会爬取不成功

3. 本周任务

- 总结 GPT - 2；
- 实践测试爬取网易云音乐数据 -- 尝试查找一些其他爬虫方法；
- 看论文。