AI古诗创作初探——从有监督到无监督

Ⅰ、上交文件说明：

我们的项目分为几个部分，在上交文件中，您将看到：

1. __pycache__：python缓存文件
2. models文件夹：保存下来的模型参数
   a. 内含wuyan文件夹，其中有encoder和decoder的模型参数
3. word_vector文件夹：预训练词向量相关代码
4. rnn文件夹：第一篇论文(Zhang Xingxing & Lapata)的相关代码实现
5. 风格迁移论文相关代码（我们项目的重点）
   a. data.py：数据处理
   b. generate.py：生成诗歌
   c. hypterdata_funct.py：超参数模块
   d. models.py：网络模型
   e. train.py：训练模块
6. GUI部分
   a. picture_view.py
   b. get_topk.py
   c. timg.png
7. 预训练词向量部分
   a. word_vector.pths
   b. word_vocab.pkl
8. readme&overview.pdf：此文件，团队项目概览
9. Pre_ai2020.pdf：项目报告文件，内含详细的原理分析
10. 风格论文.pdf：2018年 清华大学的一篇无监督风格迁移的论文《Stylistic Chinese Poetry Generation via Unsupervised Style Disentanglement》

备注：

1. 预训练词向量部分、rnn部分，我们只是呈现实现过的代码，没有写详尽的注释，另外，考虑到可能存在数据路径、运行环境等问题，并不保证其能正常运行。
2. 我们的重点是风格迁移论文相关代码。在路径设置正确的情况下，只要在终端运行train.py即可运行。原本generate可通过终端运行，后经GUI封装后需在GUI中运行。
3. 最终成果检验，在所需第三方库齐全的情况下，只要在终端运行picture_view.py即可。（风格生成部分需要加载模型参数，较慢，请耐心等待）
4. 本文件只是项目概览，详细的原理分析、感想总结，以及实践中有感而发的深刻观察，请见报告—Pre_ai2020.pdf

训练数据说明：

- 我们的训练使用了助教提供的数据集中的qtrain，内含5言诗一万余首，7言诗六万余首。

运行环境：

1. 神经网络框架：pytorch 1.4.0
2. 语言版本：python 3.7+
3. 模型训练：服务器，GPU为 GTX1080ti

项目链接：https://github.com/dromniscience/nlp_ai_2020

Ⅱ、项目介绍与组员分工

我们的项目分为四个阶段：

1. 预训练词向量
2. 直接用古诗语料训练语言模型，然后用beam search的方式生成
3. 实现2015年 Zhang Xingxing & Lapata的论文（也即助教推荐的论文）
4. 实现2018年 清华大学的一篇无监督风格迁移的论文《Stylistic Chinese Poetry Generation via Unsupervised Style Disentanglement》

在我们的项目展示中，实现了三个功能

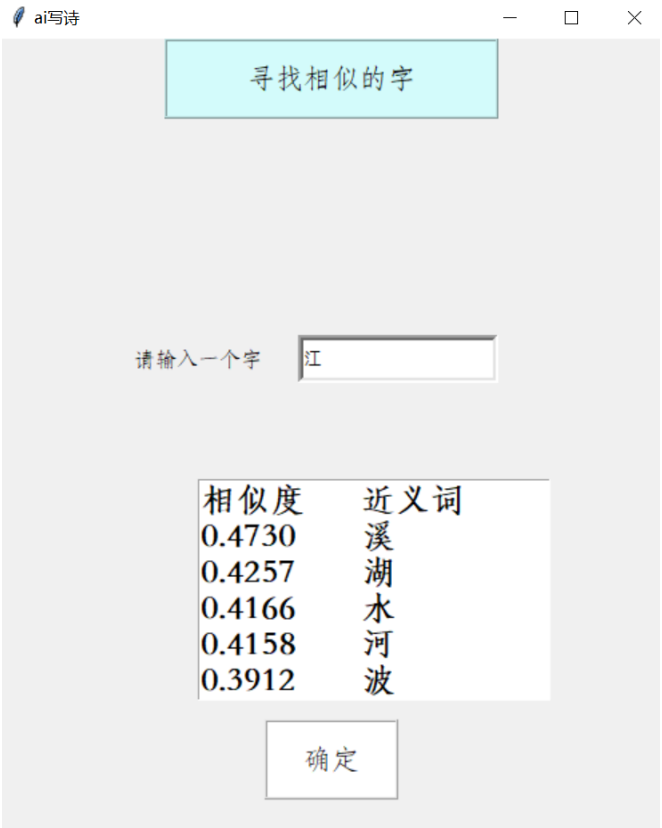
1. 利用预训练词向量，计算余弦相似度，寻找一个词的近义词
2. 利用Zhang Xingxing & Lapata的论文（下记为“第一篇论文”），生成藏头诗（考虑到上交文件的大小，这一部分我们预先找好了一些作品，在GUI交互中只作呈现，没有真实地进行神经网络运算）
3. 利用风格迁移的论文（下记为“第二篇论文”），通过保存的神经网络数据，生成给定首句后不同风格的诗句（注：这部分训练很麻烦，周期很长，到截至此刻，效果仍未做到十全十美，故我们的展示是挑选了一部分优秀的结果，我们后续会在github项目中继续更新模型参数，敬请关注~）

成员及分工

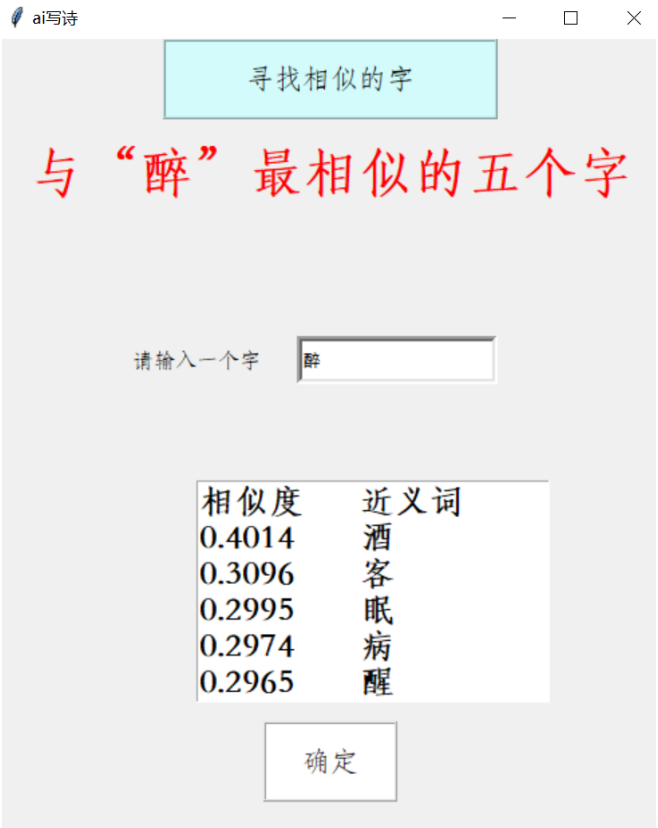
- 丁睿：第一篇论文的RCM模块、第二篇论文的风格loss项、第二篇论文的模块化、报告总主讲人
- 岳鹏云：第一篇论文的RGM模块、第二篇论文的风格loss项，风格模型网络训练
- 郑书泓：第一篇论文的模型整合、第一篇论文的ppt介绍、GUI界面
- 朱家祺：预训练词向量及相关ppt、第一篇论文的CSM部分、第一篇论文的模型整合、GUI界面
- 朱大卫：第一篇论文的数据处理模块、第二篇论文的seq2seq+attn框架，风格模型网络训练。

Ⅲ、成果展示

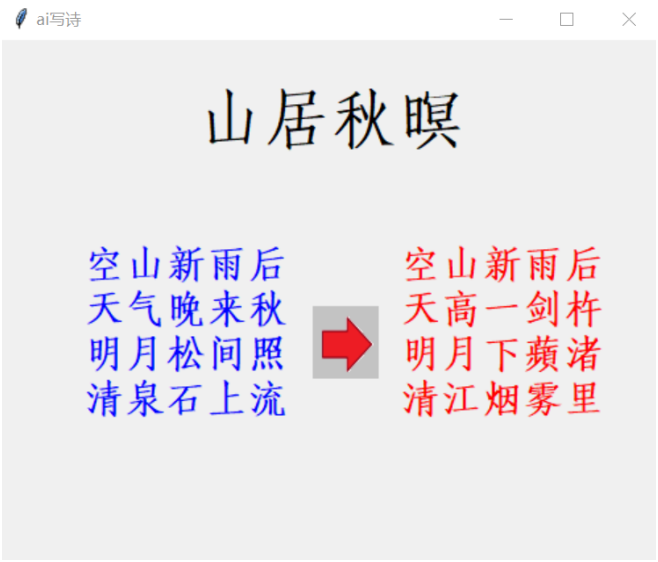
预训练的词向量：与“江”最相似的五个字



与“醉”最相似的五个字

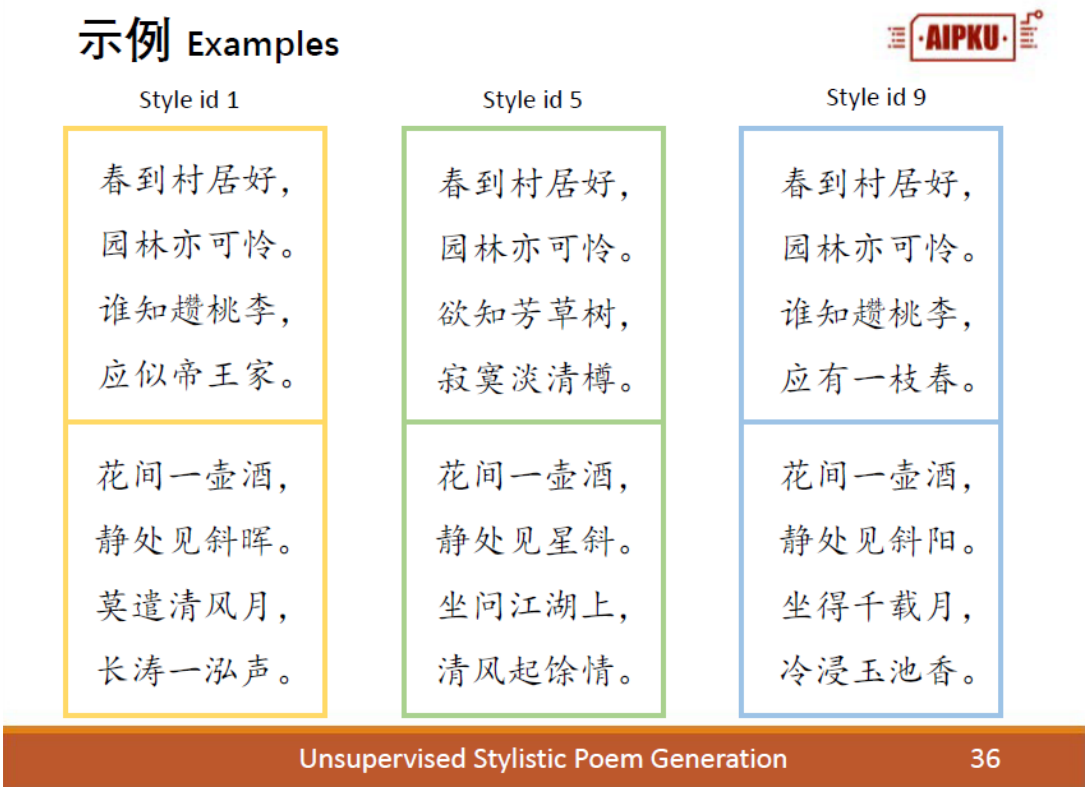


助教给的论文：老诗新赋（《山居秋暝》）



藏头诗（“我喜欢你”）

风格迁移：



Ⅳ、总结

- 详见课程报告Pre_ai2020.pdf

Ⅴ、后续工作：

1. 完善GUI界面
2. 进一步训练模型，达到更好的效果
3. 增加限制，完善诗的格律、韵脚等
4. 文本风格迁移至今仍然是nlp领域的难题，风格迁移这篇文章为我们提供了一个很好的思路，期待后续能在它的基础上实现其它语料的风格迁移。