🚀智能创作平台使用方法

前置需求

- Node.js >= 12.0.0
- Yarn
- 所需要的Python包（见back-end/mysite/requirements.txt)
    - ```
      pip install -r requirements.txt
      ```

安装

前端

```
cd Front-end
yarn install
yarn start
```

现在你可以在浏览器中访问 http://localhost:3000/

后端

```
cd Back-end/mysite
python manage.py runserver
```

在summary/utli中存放摘要生成相关代码和模型在summary文件夹下的视图层view.py相关位置调用生成函数

🚀功能简介

自动标题生成

项目描述

根据文章内容生成文章标题

环境配置

Python 版本：3.8
PyTorch 版本：1.10.0
CUDA 版本：11.3

所需环境在 `requirements.txt` 中定义。
```

数据

- 软件杯官方数据 http://www.cnsoftbei.com/plus/view.php?aid=729
- 开源摘要数据 https://zhuanlan.zhihu.com/p/341398288

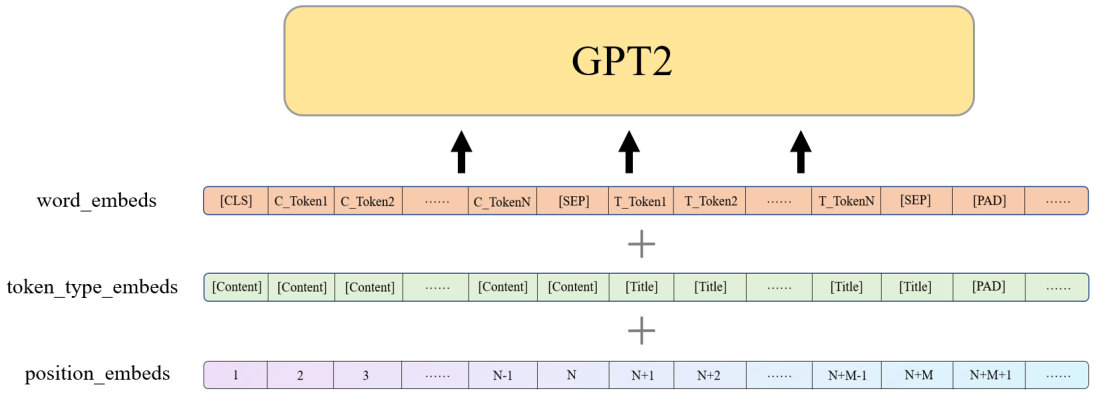
目录结构

```
./
├── README.md
├── requirements.txt         # Python包依赖文件
├── check/                   # 保存模型
├── logs/                    # 保存tfevent文件
├── data/                    # 保存训练数据
|   ├── train.json
|   ├── dev.json
|   ├── ...
├── vocab/                   # 词表目录
|   ├── vovab.txt      # 用于构建tokenizer的词表
├── config/
|   ├── config.json          # 模型配置
├── src/                     # 核心代码
|   ├── config.py      # 模型、训练参数
|   ├── dataset.py     # 数据集
|   ├── finetune.py    # 微调
|   ├── generate.py    # 测试生成
|   ├── model.py       # 模型定义
|   ├── train.py       # 训练
|   ├── utils.py       # 工具函数
```

注：vocab.txt来自 https://huggingface.co/hfl/chinese-macbert-base/blob/main/vocab.txt

模型

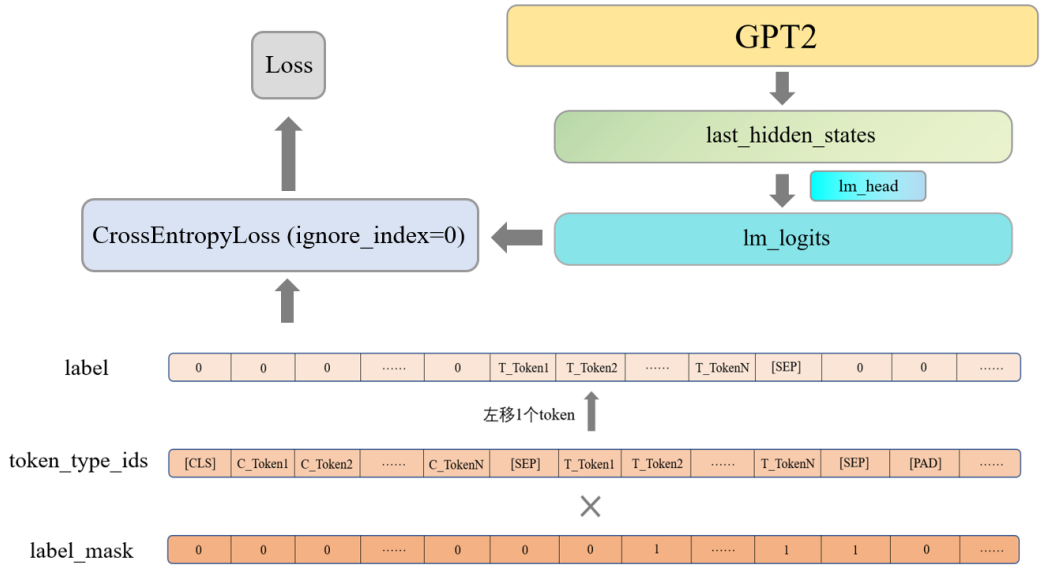
- GPT2



注：

1. N、M分别为内容和标题的最大长度。若超过最大长度，标题直接按最大长度截断，内容则分别截取最大长度一半的开头和结尾。
2. [Conetent] 和 [Title] 为添加的特殊字符，用于区分内容和标题片段。

- Loss



生成测试

```
generate:让初心和使命照亮初心使命                    truth:在学思践悟中牢记初心使命
generate:坚定“四个自信”，增强“四个自信”              truth:把我国制度优势更好转化为国家治理效能
generate:人民满意的公务员集体                truth:向人民满意而行
generate:用奋斗践行初心和使命                truth:在时间的坐标里感受共和国的光荣与梦想
generate:疫情防控每个干部的责任与使命                truth:谁守土有责，谁失职渎职，百姓看得一清二楚
generate:做官做人做事做官                    truth:为人不正为官必邪
generate:应对气候变化中国行动坚定有力                truth:中国气候行动展现重承诺勇担当的大国风范
generate:坚定不移走改革开放之路              truth:中国将坚定不移走改革开放之路
generate:以全面从严治党营造新时代的朗乾坤                    truth:标本兼治反腐败
generate:为创新发展提供战略支撑              truth:把科技自立自强作为国家发展的战略支撑
generate:以更高水平对外开放推动形成共商共享共赢              truth:以更高水平对外开放开创美好未来
generate:爱国主义情感让我们热泪盈眶                  truth:爱国主义精神构筑起民族的脊梁
generate:大数据助力国家治理现代化                    truth:以大数据促进国家治理现代化
generate:落实立德树人根本任务的关键课程              truth:理直气壮办好思政课
generate:从英雄模范中感悟奋进新时代先锋              truth:见贤思齐是对榜样最好的回应
generate:严格监管严格遵守法治、敬畏法治、敬畏法治                    truth:呵护好资本市场理性投资氛围
```

运行流程

```
训练 -> 微调 -> 测试
```

```
pip install -r requirements.txt
python src/train.py
python src/finetune.py
python src/generate.py
```

自动摘要生成

需要的资源下载

由于使用的词向量表示和训练数据集过大，并没有作为提交文件的一部分提交，故如需在本地进行部署，请完成如下资源的下载：

- 基于微博语料库训练的300维词向量
- NLPCC2017摘要数据

下载

- 基于微博语料库训练的300维词向量300维词向量，来源于https://github.com/Embedding/Chinese-Word-Vectors
    - 为加快下载，请从此处下载。
- `NLPCC2017` 摘要数据，来源于https://github.com/liucongg/GPT2-NewsTitle
    - 为加快下载，请从此处下载。

下载完成将如上资源放在目录 `backend/textrank/data` 中即可

- [x] textRank计算关键词
- [x] textRank计算关键句
- [x] web前端HTML界面
- [x] Django后端
- [ ] transformer
- [x] 检查一个词在停用词表中 用字典树优化
- [x] 计算关键句中 用字典树优化
- [ ] 更细粒度的分句
- [x] textrank可以进一步改进，如加入句子长度的惩罚，或者使用句向量判断相似性

参考资料来源

- 停用词表来源：https://github.com/goto456/stopwords
- 清华数据集来源：http://thuctc.thunlp.org/
- 清华新闻数据集来源：https://thunlp.oss-cn-qingdao.aliyuncs.com/THUCNews.zip