新闻摘要与分类系统

Created @Apr 23, 2021 7:42 PM

开发、运行环境准备

> Python + PaddlePaddle + Django

1. Python 3.8，需要安装下列库，一般直接 `pip install` + 包名即可
   - django 3.1+
   - jieba 中文分词
   - requests + beautifulsoap4 + lxml
   - SnowNLP
   - paddlepaddle， numpy，
   - 其他如果有报错说缺什么，就装什么
2. PyCharm 尽量用 Professional 版本，Community 也行
3. 网页 UI 需要的 css/js 都在项目文件中了，不需要额外下载

项目结构

1. `data/*.txt` 这些是爬虫获取到的数据文件，按照新闻的类别，把对应的摘要写入对应的文件中，`data_test/*.txt` 里面是模型测试的数据
2. `NewsSummaryClassification` 这里面是网页项目运行、启动相关的配置，一般不需要修改
3. `news_classification` 是网页展示模块，主要关注下列文件
   - `static`：里面包含网页静态文件， css 样式、js 脚本、img 图片素材等
   - `templates`：网页组织 HTML 文件模板，基本一个 HTML 对应一个页面
   - `views.py`：实现前后端数据交互
4. `paddlemodel`：通过 paddlepaddle 实现的深度神经网络的分类模型
5. `spider`：新闻爬虫相关，通过 `sina_news.py` 启动

6. `model.pdopt` 和 `model.pdparams`：深度模型的模型参数保存，可以快速加载、使用模型
7. `manage.py`：网页项目的运行入口
8. `embeddings.txt` 和 `vocabulary.txt`：文本特征提取相关，预训练的词向量数据

运行方式

1. 新闻爬虫
   - 启动 `spider/sina_news.py` 脚本，通过参数指定起始页和中止页数
   - 启动过程中会在控制台输出获取到的新闻的基本信息
   - 结束后，会写入 `data` 文件夹下对应类别的文件中
2. paddlepaddle 分类模型的训练与测试
   - 从 https://github.com/Embedding/Chinese-Word-Vectors 下载的中文词向量文件
   - 修改 `paddlepaddle/embedding.py` main 函数里的文件路径
   - 执行 `embedding.py` 从而得到 `embeddings.txt` 和 `vocabulary.txt`，作为提取特征的预训练数据
   - 运行 `network.py` 开始训练模型，并且保存下来
3. 网页展示
   - PyCharm 专业版可以直接启动 Django 项目，或者命令行输入 `python manage.py runserver 8000`，等命令行出现

     ```
     April 23, 2021 - 12:04:58
     Django version 3.1.7, using settings 'NewsSummaryClassification.settings'
     Starting development server at http://127.0.0.1:8000/
     Quit the server with CONTROL-C.
     ```

   - 表示成功启动，一些 warning 可以忽略
   - 浏览器打开访问 127.0.0.1:8000