作业三

冯昊 2016013255

爬虫使用

爬虫在目录`./Homework3/spider.py`，其执行会改变数据库文件`./Homework3/db.sqlite3`的内容。

- 直接执行：
  - 进入`./Homework3/`目录后执行：
    - `python3 spider.py [num=30]`可以从腾讯网首页上爬取`num`条未在数据库中的新闻页面并加入数据库中。
      - `num`可以省略（此时为默认值30），且不建议设置过大的数目，以免被封IP或者耗时过长。
    - `python3 spider.py -l`可以从腾讯网首页上爬取设置在文件`base_functions.py`中的默认10条网址，但如果这些网址已经存在数据库中，将不再重复爬取。
    - `python3 spider.py -f filename`可以打开文本文件`filename`并且读取其中网址数据（每行一条），并且爬取不在数据库中的网址。
    - `python3 spider.py -url URL`可以爬取网址为`URL`的新闻页面，并且加入数据库中
      - 直接输入URL的命令不会检查`URL`是否已在数据库中，并且不会替换原有在数据库中的相同网址的内容。
    - 输入不合法指令会提示正确输入方法。
- 在网页上进行操作：
  - 在新闻列表页面(`/news/`)有按钮`Get new data`，按下后会进行从腾讯网首页进行爬取10条新闻网页的操作。
  - 这个操作是使用全局多线程变量实现的，因此在这一操作过程中可以同时进行网页的访问等其他操作。
  - 同时执行多次爬取操作的避免：
    - 在代码逻辑中，会进行是否正在进行数据库爬取操作的判断，以避免同时爬取造成重复与性能下降，以及潜在的被封IP的可能。
    - 因此，无法在上一次爬取数据未结束时进行新的爬取数据操作。

网页部署

该工程由`PyCharm`创建，因此直接使用`PyCharm`打开项目目录运行即可。

若手动运行，则直接执行`python manage.py runserver`

管理员账号：

- 账号：
  - `admin`
- 密码：
  - `admin`

网页介绍

- 整体框架：
  - 该项目包含新闻列表、新闻详情、管理员登陆、管理员后台管理等几个页面。
  - 每个网页第一行右侧有`Home`（返回首页）、`Admin`（后台管理）、`Logout`（退出登录）、`Login`（管理员登陆）选项，当登陆后则登陆选项不显示，反之不显示后台管理与退出选项。
- 新闻列表网页：
  - 获取新数据按钮：`Get new data`，点击后会在后台从腾讯新闻首页查找10条未录入新闻加入数据库中。
  - 查找：输入查询词后点击`Search`以在全文查找，点击`Search(Only title)`以在标题查找。
  - 新闻列表：显示新闻标题，点击进入对应的详情页面。
  - 翻页：若可以翻页，则能够点击前一页`←`或者后一页`→`；同时会显示当前页数与总页数，右侧可以输入页面号后点击跳转。
- 新闻详情页：
  - 显示新闻文本内容，页面标题为新闻标题，新闻内容后有：阅读原文（`Go to original page`）选项与返回（`Back to news list`）选项。
- 登陆页面、后台管理页面：
  - 若未登录，则访问`/admin/`会自动跳转到登陆页面，密码经过md5加密后传输。
  - 登陆后访问`/admin/`则可进入后台管理页面，显示所有新闻并且可以选择删除。点击`recover deleted`按钮可进入恢复页面。
  - 在恢复页面和管理页面均可对单独的新闻或者多选后点击`recover selected`或者`delete selected`以对多个新闻操作。
- 删除相关页面：
  - 若登陆，访问`/del/[int:id]`即可删除；否则返回错误信息。

网页功能

- **登陆注销**：
  - 使用自建数据库操作，同时将登陆时随机生成的字符串作为`session`保存到数据库`news_user`里的`sess`字段内，以避免同一用户在多处登陆。
  - 在加载网页时传递的参数首先检查是否登陆，以确定页首显示的登陆、注销、管理等页面正常。
- **文章删除与恢复、批量操作**：
  - 通过在表内设置`deleted`字段来判断有无删除，通过修改该字段实现删除和恢复操作。
  - 使用多选框以实现多选批量删除/恢复的功能，传递信息时将`id`号连接成字符串后`POST`到后端进行解析。
  - 若网页已经删除，则只能被管理员看到，未登陆则打开会提示404。
  - 被删除的网页在列表里面不会出现（包括新闻列表与查询结果，无论自己是否为管理员）。
- **文章查找**：
  - 在新闻列表中可以输入关键词查找。
  - 支持对全文匹配查找，若搜索词整体完全连续出现在标题或者新闻中（正文或者标题）则会被显示出来。
- **内容分页**：
  - 新闻列表内容以20条为一页，可以向前、向后翻页与指定页面跳转。