Tue, Dec 22, 2015

Python Project (http://odinaryk.github.io/blog/2015/12/22/python-project/)

Python个人文档报告

5140379046 樊泽坤

负责工作一览

- 1. 项目全局架构
- 2. 后端：京东数据抓取
- 3. 前端：GUI部分，QtItemLayout和QtItemBox的实现
- 4. 数据分析：排序、存储以及刷新显示
- 5. 前后端交互，代码整合

具体实现

1. 项目全局架构

- 前后端分离，定义前端GUI的设计目标，定义后端数据get方法输入输出接口，定义item项
- 小组分工，4人各自负责一个网站的抓取，以及部分前端设计或团队文档书写工作

2. 后端：京东数据抓取(jd.py)

首先尝试直接从电脑网页端抓取。
京东的html数据大多由js动态生成div项构造，直接通过requests的get请求从网页源代码抓取不可行。
仔细查看京东电脑网页端源代码，发现所需要的多项数据缺项，若是任由其空缺未免可惜。
然后尝试从手机网页端抓取。
手机网页端通过bs4解析抓取数据，缺项仅价格一项，为动态生成，然而比价软件，价格是第一要务决不可空缺。
仔细查看网页源代码，发现一个json项warelist，内含所有所需信息。
最后，通过 `requests` 模拟请求，构造正则表达式，利用 `re` 匹配出**warelist**的内容，然后调用 `json` 解析器，解析为**python**对象，再构造出**item**，并 `dump` 为**json**在磁盘上保存。

3. 前端：GUI部分，QtItemLayout和QtItemBox的实现(item_show.py)

- `QtItemLayout` 继承 `QGridLayout` ， 整体排版一系列QWidget的组合
  绑定多个 `QLabel` ，每个 `QLabel` 对应一个数据项，将 `QPixmap` 与 `QLabel` 绑定实现图片显示
  绑定一个 `QPushButton` ，事件绑定 `webbrowser.open` ，实现调用外部浏览器
- 一个 `QtItemBox` 为多个 `QtItemLayout` 的组合，其整体作为一个 `QWidget` ，以便于被上层 `QWindow` 管理
  一个 `QtItemBox` 默认由12个 `QtItemLayout` 为4*3排布，内部通过一个 `QGridLayout` 的 `addLayout` 方法对 `QtItemLayout` 进行整体排版

4. 数据分析：排序、存储以及刷新显示(gui.py部分内容)

- 排序使用 `list.sort(key=lambda x:x['key'])` 进行排序
- 存储使用 `json.dump` ，命名规则为商品名(gbk)+时间(s)+.json
- 刷新显示采用删除 `QtItemBox` 这个Widget并根据排序后的itemlist重新构造 `QtItemBox` 并绑定在 `QWindow` 上
- 事件驱动，捕捉相应菜单项的信号，通过自定函数处理，使用connect绑定

5. 前后端交互，代码整合(item_get.py,item_show.py,gui.py)

- item_get.py中，统一调用预先设计的 `get(name)` 接口，调用jd,amazon,tmall,one这四个模块获取json数据，再从文件读取
- 由于爬虫限制等原因，部分模块有时无法成功抓取数据，此时捕捉到exception，相应的get方法会返回空列表
- item_show.py中，调用item_get中的 `get_items(name)` 方法，得到itemList
  并通过itemList构造 `QtItemBox`
- gui.py中，捕捉 `SearchButton` 的事件，同时读取 `SearchTable` (QLineEdit)的unicode串数据，并调用item_show中的 `QtItemBox` 类构造函数以构造itemBox，并刷新显示

问题及解决过程

1. 原定义item为自定义class，后来发现内建 `dict` 更适合，修改定义
2. 代码架构时，未设计统一的get方法，后来各自返工添加
3. 后端设计抓取方案时问题不断，解决过程见上
4. GUI部分，原本Item也设计为 `QWidget` ，后来排版重叠，修改为更合适的 `QGridLayout`
5. QtItemBox的数据，原本在类定义中书写，然而由于类static性质，刷新显示失败，因此改为在 `__init__()` 中初始化
6. 前后端交互时，中文编码是一大问题，后来手工统一为 `unicode` 串，解决
7. 实现排序时，由于各人所抓取的数据格式不一致，尤其是缺项格式，因此细化定义统一格式

收获及感想

- 学习了 `requests` 、 `BeautifulSoup` (虽然最后没用到)， `re` 的运用及正则表达式书写，对于运用 `PyQt4` 写python GUI有了一点经验，同时为其他语言使用Qt打下一定基础。
- 写爬虫技术进一步提高，绕过了js动态生成div的问题
- 项目架构的具体化真的很重要，否则需要小组成员不断返工，切记切记
- 项目管理和分工也并不是随随便便的事情，从小组各成员身上学到了很多。
- 深刻体会到python相比于C++在开发效率上的优势和语法的自然性

dev (http://odinaryk.github.io//tags/dev)

Related Post

← Older

All Posts (http://odinaryk.github.io//post)

Newer → (http://odinaryk.github.io/blog/2015/12/20/a-strike-in-cet-6/)

Recent Posts

Python Project (/blog/2015/12/22/python-project/)