百科搜索聚合

概述

在大学学习过程中，常常会需要搜索一些遗忘的概念。但是从单方面，如百度百科获取信息，有时候不足以完全理解某些概念，需要再次从维基百科中查询相关内容。

鉴于大陆对维基百科中文的不友好，我们还需要把概念翻译成英文后再进行搜索。

这个基于Python3.5的**百科搜索聚合**便是为此而设计，你只需给出关键词的中文或英文，程序便会自动爬取所需概念的概要，打印到屏幕上。

程序使用

运行Python脚本

1. 利用集成开发环境

如Pycharm, 导入项目后运行，即可使用

2. 在命令行窗口使用Python解释器

`Ctrl+R`输入`cmd`呼出命令行窗口，使用命令

```
cd 文件夹根目录(baikeWiki的上一级)
py -3.5 -m baikeWiki.spiderMain
```

查询方法

当运行成功后，直接输入要搜索的关键词（中文英文均可），回车，便会返回结果。

图示：

搜索`离散数学`

程序设计

技术路线

有道翻译API + 第三方库：

1. requests
2. BeautifulSoup
3. re (后续版本)

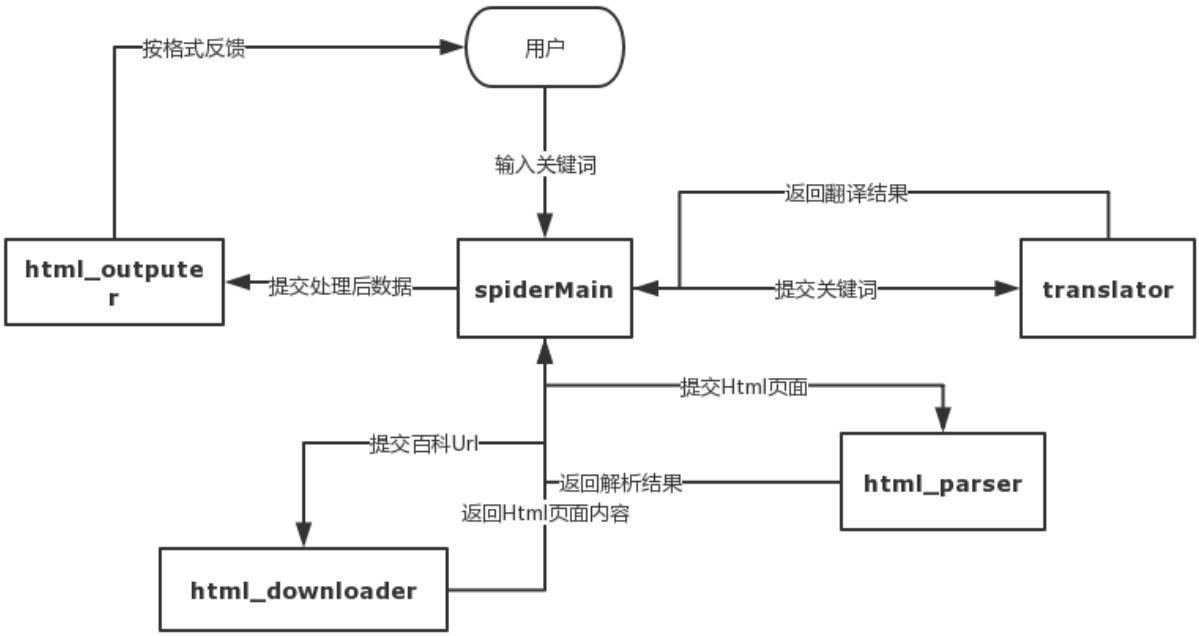
基本框架

根据最终程序所需，程序分为五大部分，分别为：

1. `spiderMain` 主函数，也可说是调度器，程序的入口，调度其他类
2. `translator` 翻译器，利用有道API翻译关键词
3. `html_downloader` 下载器，下载HTML网页
4. `html_parser` 解析器，解析HTML网页
5. `html_outputer` 输出器，输出解析后内容

程序从`spiderMain`进入，用户输入关键词，提交给`translator` 翻译成英文， 中文英文提交给`Url_generator` 生成百度百科、维基百科链接提交给`html_downloader` ，下载下来的网页提交给`html_parser`解析，解析出百科内容后，提交给`html_outputer` ，一个循环结束。

流程图如下：



SpiderMain

主要完成整个程序的调度，Url生成器在这个类中，后期可以考虑整合到`html_downloader`里面。

异常的处理：利用有道API翻译时，网络连接失败会导致程序异常

YouDao(translator)

- 有道翻译官网申请API接口
- 网上找的使用方法
- 利用中文`utf-8`编码特性，判断用户输入类型

HtmlDownloader

- 使用第三方库`requests`下载HTML网页

HtmlParser

- 使用第三方库`BeautifulSoup`解析HTML网页

百度百科：使用浏览器打开一个百科网页，分析源代码，发现百度百科的标题保存在一个`<dd class='lemmaWgt-lemmaTitle-title'>`标签下的`h1`标签内，而摘要保存在一个`<div class='lemma-summary'>`的标签内。

利用`BeautifulSoup`提供的`find()`方法找到这两个标签，再用`.get_text()`方法获取标签内的字符串。最后将两个信息存到一个字典中。

类似的，对

维基百科：我们访问一个维基百科页面，分析它的源代码，发现标题保存在`id='  firstHeading'`的标签下，摘要保存在`id=mw-content-text`的第一个子标签`<p>`中。

利用同样的办法，获取到维基百科的标题和摘要存到字典中。

HtmlOutputer

由于我们从`Parser`接受的数据是个字典，因此我们可以很方便的将数据输出成我们喜爱的格式。

打印形式后期可以根据内容的增减进行优化。

后续版本

修复已知Bug

个例：

1. math ,wiki异常
2. cat 异常

通病：

1. wiki同义词页面处理
2. 异常字符打印错误

改进，增加功能

1. **增加对同义词页面的选择支持**

分析同义词页面，展现一个菜单，为用户提供选择的途径。

2. **对百科内容的二次加工，自动生成更全面的摘要**
3. **对有道翻译提供的内容进行处理，整合进**translator

在测试中，发现翻译诸如`猫`这类词时，有道会返回`the cat`这种带冠词的结果，而利用这个结果进行搜索是无法搜到百科内容的。

改进方案：

- 后期去掉`the`
- 改进获取百科页面的途径
- ...

4. **打包程序，生成无序**Python**解释器的`exe`可执行文件**
5. **更友好的**GUI
6. ...