Web Spider Project Framework

1. Project Part A (Input: search keyword; Output: csv file)

Create final\_keyword string (lowercase of search keyword) which will be used for file naming purposes and creating a dataframe

Use if statement to replace whitespace as “+”

No

Yes

Is the search keyword more than one word?

Type in search keyword (疾病名称) in English

Form search url by combining <http://www.chinapubmed.net/search/?q=> + search keyword + “&sort=year”. Set current Page # to 1 and create soup object with search url to read html

Start

Function F: create keyword series

Input: final\_keyword

Output: a series of the keyword

Function E: extract 期刊

Input: none

Output: A series of 期刊

Function D: extract pmid

Input: none

Output: A series of pmid

Function C: extract 英文标题

Input: none

Output: A series of 英文标题

Function B: extract article url(原始url)

Input: none

Output: A series of article urls

Function A: extract 影响因子

Input: none

Output: A series of 影响因子

Function G: create initial dataframe

Input: final\_keyword

Output: Runs Functions A through F to create a filled in Dataframe with columns Keyword, 原始url, 影响因子, 英文标题, pmid, and 期刊. Rows matching the criteria 影响因子 > 3 and with final\_keyword in the 英文标题 are kept

,

Global Variables

1. Final keyword
2. soup
3. Search url
4. Page #
5. Dataframe
6. Length of Dataframe

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Keyword | 原始url | 影响因子 | 英文标题 | pmid | 期刊 |
| ….. | …. | ….. | …… | …… | …… |

|  |  |  |
| --- | --- | --- |
| Language | IDE | Libraries used |
| Python | Python 3 Jupyter Notebooks | 1. Bs4 2. Requests 3. Pandas 4. re |
| Note: Path of csv file is currently set to same folder in which the python 3 script is run. This may be adjusted accordingly. | | |

Reset index of dataframe and export as csv file

Yes

Run Function I until 40 articles matching the criteria are found

No

Have 40 articles matching the criteria been found (len(Dataframe) = 40)?

Function I: Extract information from next page and append to Dataframe

Input: none

Output: Runs Function H and Function G. Updates Dataframe and Length of Dataframe

1. Set Dataframe equal to output of Function G
2. Set Length of Dataframe equal to len(Dataframe)

Function H: find next url/page

Input: none

Output: updates Search url and increases Page # by 1

*Test Run (Web Crawler Project Part A.ipynb)*

Search keyword = Breast cancer

Initial Search url = <http://www.chinapubmed.net/search/?q=Breast+cancer&sort=year>

Graphical user interface, text, application, email

Description automatically generated

A close up of text on a white background

Description automatically generated

1. Project Part B (Input: csv file, Output: csv file)

|  |  |  |
| --- | --- | --- |
| Language | IDE | Libraries used/to be used |
| Python | Python 3 Jupyter Notebooks | 1. Bs4 2. Requests 3. Pandas 4. re |
| Note: Path of csv file is currently set to same folder in which the python 3 script is run. This may be adjusted accordingly. | | |

Structure:

1. Create function to read csv file from Part A as dataframe1 and return dataframe1
2. Create functions to extract the following information
   1. Function A: 关键词
   2. Function A.2: 关键词当前网页的网址
   3. Function B: 中文标题
   4. Function C: 中文正文
   5. Function D: 作者
   6. Function D.2: 作者当前网页的网址
   7. Function E: 英文正文
   8. Function F: 发表日期
   9. Function G: 官网
   10. Function H: 免费下载链接
   11. Function I: 全文下载链接
3. Create a function (J) to create and fill new columns in dataframe (new columns = 关键词，中文标题，中文正文，作者，英文正文，发表日期，官网，免费下载链接，全文下载链接)
4. Use appropriate functions from Project Part A to crawl through up to five pages of urls extracted from Functions A.2 and D.2 Create a new dataframe2 for this extracted information.
5. Repeat function J to extract information for missing columns for this new dataframe2.
6. Append dataframe2 to dataframe1 to make dataframe\_final. Reorganise order of columns if necessary. Reset index of dataframe\_final and export as csv. Desired format is as shown:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 疾病名称 (keyword) | 原始url | 关键词 | 中文标题 | 中文正文 | 英文标题 | 作者 | 英文正文 | pmid | 期刊 | 影响因子 | 发表日期 | 官网 | 免费下载链接 | 全文下载链接 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |