# Data Combine on Python

_ 达 ___

2025 年 8 月 24 日

## 1 文档介绍：

- 从 zotero 导出的医学数据 csv 文件一共有五个（按照年份）
- 数据清洗时，合并为一个文件更便于处理，这就是本文件所做的事情
- 这个文档处理完后的数据会到 `data_process.ipynb` 中继续后续处理

[6]:
```
import pandas as pd
import glob
import os
```

### 1.1 定义文件路径

- 意思让电脑知道文件的位置

[7]:
```
# --- 1. 定义文件路径 ---
# 定义数据文件夹的相对路径
# . 表示当前 Notebook 文件所在的目录
# .. 表示上一级目录
# 假设你的 Notebook 在 'notebooks/' 文件夹下，而数据文件夹 'data/raw/'
# 在仓库的根目录下，所以需要先回退一级 (..) 再进入 'data/raw/'
data_path = os.path.join(os.getcwd(), '..', 'data', 'raw')
processed_data_path = os.path.join(os.getcwd(), '..', 'data', 'processed')
```

[8]:
```
# 使用 glob 找到所有以.csv 结尾的文件
# 这个路径是相对于当前 Notebook 文件执行时的工作目录而言的
all_csv_files = glob.glob(os.path.join(data_path, "*.csv"))
# 打印出来检查一下，确保文件都被找到了
print("以下 CSV 文件将被读取：")
if not all_csv_files:
```

```
    print("--- 警告：未找到任何 CSV 文件！请检查：---")
    print(f"1. notebook 的位置是否正确？ (当前工作目录：{os.getcwd()})")
    print(f"2. 'data/raw/' 文件夹是否在你的仓库根目录下？")
    print(f"3. 'data/raw/' 文件夹内是否包含您要读取的 .csv 文件？")
else:
    for f in sorted(all_csv_files): # 使用 sorted 保证输出顺序一致
        print(f)
```

以下 CSV 文件将被读取：

d:\MEDscience_map_of_XJTU\notebook\..\data\raw\pubmed-XianJiaoto-set2021.csv
d:\MEDscience_map_of_XJTU\notebook\..\data\raw\pubmed-XianJiaoto-set2022.csv
d:\MEDscience_map_of_XJTU\notebook\..\data\raw\pubmed-XianJiaoto-set2023.csv
d:\MEDscience_map_of_XJTU\notebook\..\data\raw\pubmed-XianJiaoto-set2024.csv
d:\MEDscience_map_of_XJTU\notebook\..\data\raw\pubmed-XianJiaoto-set2025.csv

### 1.2 关键步骤：合并操作

[9]:
```
# --- 2. 循环读取并合并 ---
# 创建一个空列表，用来存放每个文件读取后的 DataFrame
df_list = []

for file in all_csv_files:
    # 循环读取每一个 csv 文件
    df_temp = pd.read_csv(file)
    # 将读取的 DataFrame 添加到列表中
    df_list.append(df_temp)

# 使用 pd.concat() 函数将列表中的所有 DataFrame 一次性合并
# ignore_index=True 是非常重要的一步，它会重新生成一套连续的索引，避免索引重复
df_combined = pd.concat(df_list, ignore_index=True)
```

### 1.3 保存合并后的文件

[12]:
```
# --- 3. 保存合并后的文件 ---
# 定义新文件的保存路径，使用 processed_data_path
output_filename = os.path.join(processed_data_path, 'zotero_data_combined.csv')
# 将合并后的大文件保存起来
try:
```

```
    df_combined.to_csv(str(output_filename), index=False) # to_csv 需要字符串路
径
    print(f"\n合并完成！总共包含 {len(df_combined)} 条记录。")
    print(f"数据已保存到: {output_filename}")
except Exception as e:
    print(f"保存文件 {output_filename} 时出错: {e}")
```

合并完成！总共包含 25105 条记录。
数据已保存到:
d:\MEDscience_map_of_XJTU\notebook\..\data\processed\zotero_data_combined.csv

### 1.4 检查合并后的数据

[13]:
```
# --- 4. 检查合并后的数据 ---
# 显示前 5 行，快速预览
print("\n合并后数据预览: ")
print(df_combined.head())

# 显示数据信息，检查列名、非空值和数据类型
print("\n合并后数据信息: ")
df_combined.info()
```

合并后数据预览:

| | Key | Item Type | Publication Year | \ |
|---|---|---|---|---|
| 0 | 4WPZ4WZA | journalArticle | 2022 | |
| 1 | 8BVI224Y | journalArticle | 2021 | |
| 2 | XUSNP58T | journalArticle | 2021 | |
| 3 | H4XWF2NY | journalArticle | 2021 | |
| 4 | EKJAB2RB | journalArticle | 2021 | |

| | Author | \ |
|---|---|---|
| 0 | Sun, Lei; Wang, Hua; Yu, Shanshan; Zhang, Lin;… | |
| 1 | Zhou, Lina; Ma, Xiancang; Wang, Wei | |
| 2 | Jin, Xuting; Ren, Jiajia; Li, Ruohan; Gao, Ya;… | |
| 3 | Feng, Wei; Wang, Jian; Yan, Xin; Zhang, Qianqi… | |

```
4  Shao, Can; Wang, Xiaomeng; Ma, Qingyan; Zhao, …

                                                 Title  \
0  Herceptin induces ferroptosis and mitochondria…
1  Relationship between Cognitive Performance and…
2  Global burden of upper respiratory infections …
3  ERK/Drp1-dependent mitochondrial fission contr…
4  Analysis of risk factors of non-suicidal self-…

                             Publication Title  ISBN                ISSN  \
0  International journal of molecular medicine   NaN  1791-244X 1107-3756
1               Journal of affective disorders   NaN  1573-2517 0165-0327
2                           EClinicalMedicine   NaN            2589-5370
3                          Cell proliferation   NaN  1365-2184 0960-7722
4                Annals of palliative medicine   NaN  2224-5839 2224-5820

                           DOI  Url  … Programming Language Version System  \
0        10.3892/ijmm.2021.5072  NaN  …                  NaN     NaN    NaN
1     10.1016/j.jad.2020.12.059  NaN  …                  NaN     NaN    NaN
2  10.1016/j.eclinm.2021.100986  NaN  …                  NaN     NaN    NaN
3             10.1111/cpr.13048  NaN  …                  NaN     NaN    NaN
4          10.21037/apm-21-1951  NaN  …                  NaN     NaN    NaN

  Code  Code Number Section  Session Committee History  Legislative Body
0  NaN          NaN     NaN      NaN       NaN     NaN               NaN
1  NaN          NaN     NaN      NaN       NaN     NaN               NaN
2  NaN          NaN     NaN      NaN       NaN     NaN               NaN
3  NaN          NaN     NaN      NaN       NaN     NaN               NaN
4  NaN          NaN     NaN      NaN       NaN     NaN               NaN

[5 rows x 87 columns]
```

合并后数据信息:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25105 entries, 0 to 25104
Data columns (total 87 columns):
 #   Column                Non-Null Count  Dtype
```

```
---  ------                --------------  -----
 0   Key                   25105 non-null  object
 1   Item Type             25105 non-null  object
 2   Publication Year      25105 non-null  int64
 3   Author                25105 non-null  object
 4   Title                 25105 non-null  object
 5   Publication Title     25064 non-null  object
 6   ISBN                  0 non-null      float64
 7   ISSN                  25063 non-null  object
 8   DOI                   24969 non-null  object
 9   Url                   0 non-null      float64
 10  Abstract Note         24183 non-null  object
 11  Date                  25105 non-null  object
 12  Date Added            25105 non-null  object
 13  Date Modified         25105 non-null  object
 14  Access Date           0 non-null      float64
 15  Pages                 22971 non-null  object
 16  Num Pages             0 non-null      float64
 17  Issue                 17668 non-null  object
 18  Volume                24594 non-null  object
 19  Number Of Volumes     0 non-null      float64
 20  Journal Abbreviation  25064 non-null  object
 21  Short Title           0 non-null      float64
 22  Series                0 non-null      float64
 23  Series Number         0 non-null      float64
 24  Series Text           0 non-null      float64
 25  Series Title          0 non-null      float64
 26  Publisher             0 non-null      float64
 27  Place                 41 non-null     object
 28  Language              25105 non-null  object
 29  Rights                17986 non-null  object
 30  Type                  0 non-null      float64
 31  Archive               0 non-null      float64
 32  Archive Location      0 non-null      float64
 33  Library Catalog       0 non-null      float64
 34  Call Number           0 non-null      float64
 35  Extra                 25105 non-null  object
```

```
36  Notes                0 non-null      float64
37  File Attachments     0 non-null      float64
38  Link Attachments     0 non-null      float64
39  Manual Tags          22640 non-null  object
40  Automatic Tags       0 non-null      float64
41  Editor               0 non-null      float64
42  Series Editor        0 non-null      float64
43  Translator           0 non-null      float64
44  Contributor          0 non-null      float64
45  Attorney Agent       0 non-null      float64
46  Book Author          0 non-null      float64
47  Cast Member          0 non-null      float64
48  Commenter            0 non-null      float64
49  Composer             0 non-null      float64
50  Cosponsor            0 non-null      float64
51  Counsel              0 non-null      float64
52  Interviewer          0 non-null      float64
53  Producer             0 non-null      float64
54  Recipient            0 non-null      float64
55  Reviewed Author      0 non-null      float64
56  Scriptwriter         0 non-null      float64
57  Words By             0 non-null      float64
58  Guest                0 non-null      float64
59  Number               0 non-null      float64
60  Edition              0 non-null      float64
61  Running Time         0 non-null      float64
62  Scale                0 non-null      float64
63  Medium               0 non-null      float64
64  Artwork Size         0 non-null      float64
65  Filing Date          0 non-null      float64
66  Application Number   0 non-null      float64
67  Assignee             0 non-null      float64
68  Issuing Authority    0 non-null      float64
69  Country              0 non-null      float64
70  Meeting Name         0 non-null      float64
71  Conference Name      0 non-null      float64
72  Court                0 non-null      float64
```

```
 73  References            0 non-null      float64
 74  Reporter              0 non-null      float64
 75  Legal Status          0 non-null      float64
 76  Priority Numbers      0 non-null      float64
 77  Programming Language  0 non-null      float64
 78  Version               0 non-null      float64
 79  System                0 non-null      float64
 80  Code                  0 non-null      float64
 81  Code Number           0 non-null      float64
 82  Section               0 non-null      float64
 83  Session               0 non-null      float64
 84  Committee             0 non-null      float64
 85  History               0 non-null      float64
 86  Legislative Body      0 non-null      float64
dtypes: float64(66), int64(1), object(20)
memory usage: 16.7+ MB
```

### 1.5 下一步

- 进入 data_process.ipynb 继续下一步数据处理