爬取PTT貼文資訊

chatgpt.com

以下是符合你需求的 Python 程式，使用 `requests`、`BeautifulSoup` 和 `lxml` 來爬取 PTT NBA 板的最新三頁貼文，並儲存成 UTF-8 編碼的 JSON 檔案 `ptt_NBA.json`。程式使用 PEP8 命名與格式規範，並加上詳細的 Step 中文註解：

```
import requests
from bs4 import BeautifulSoup
import json
import time

# Step 1：使用 requests 模組取得 HTML 網頁內容
BASE_URL = "https://www.ptt.cc"
START_URL = f"{BASE_URL}/bbs/NBA/index.html"

# 儲存所有貼文資料的列表
all_posts = []

# 設定要爬取的頁數
pages_to_crawl = 3
current_url = START_URL

# 模擬一般瀏覽器的 headers
HEADERS = {
    'User-Agent': 'Mozilla/5.0'
}

# 建立 Session 並通過 PTT 的年齡確認
session = requests.Session()
session.cookies.set('over18', '1')


for page inrange(pages_to_crawl):
    # Step 1：取得當前頁面 HTML
    response = session.get(current_url, headers=HEADERS)
    soup = BeautifulSoup(response.text, 'lxml')

    # Step 2：使用 CSS 選擇器取得所有貼文的 <div> 標籤（class 為 .r-ent）
    entries = soup.select('.r-ent')

    # Step 3：走訪每一筆貼文資料for entry in entries:
        post_data = {}

        # 1. 推文數：.nrec .hl.f1（有些是空白）
        push_tag = entry.select_one('.nrec .hl')
        post_data['推文數'] = push_tag.text.strip() if push_tag else'0'# 2.
文章標題：.title a（有些被刪除沒有 a）
        title_tag = entry.select_one('.title a')
        if title_tag:
            post_data['文章標題'] = title_tag.text.strip()
            # 3. 文章網址：從 <a> 標籤 href 屬性取得
            post_data['文章網址'] = BASE_URL + title_tag['href']
        else:
            post_data['文章標題'] = '(本文已被刪除)'
            post_data['文章網址'] = None# 4. 日期：.meta .date
        date_tag = entry.select_one('.meta .date')
        post_data['日期'] = date_tag.text.strip() if date_tag else''# 5. 作
者：.meta .author
        author_tag = entry.select_one('.meta .author')
        post_data['作者'] = author_tag.text.strip() if author_tag else''# 加
入到列表中
```

```
            all_posts.append(post_data)

        # Step 4：取得上一頁的 URL，繼續爬取
        prev_link = soup.select_one('.btn-group-paging a:nth-child(2)')
        if prev_link and'href'in prev_link.attrs:
            current_url = BASE_URL + prev_link['href']
            time.sleep(0.5)  # 禮貌性延遲else:
            break# Step 5：將結果寫入 JSON 檔案，使用 UTF-8 編碼
withopen('ptt_NBA.json', 'w', encoding='utf-8') as f:
    json.dump(all_posts, f, ensure_ascii=False, indent=4)

print('已成功將資料儲存為 ptt_NBA.json')
```

✔️ 程式功能總結：

- ✔️ 爬取 PTT NBA 板共 **3 頁貼文資訊**
- ✔️ 擷取項目：推文數、標題、網址、日期、作者
- ✔️ 處理被刪除的貼文（無 `<a>` 標籤）
- ✔️ 將所有貼文以 `UTF-8` 編碼儲存成 `ptt_NBA.json`

如果你還需要額外欄位（例如內文、IP、留言數等）或爬取更多頁數，可以進一步擴充功能。需要我幫忙嗎？