**BÁO CÁO KẾT QUẢ**

**Tên đề tài:** WEB CRAWLING CHO CÁC KẾT QUẢ TRẬN ĐẤU CHAMPIONS LEAGUE

1. **Tổng quan**

* **Mục tiêu:** Mục tiêu chính của dự án là thu thập và trích xuất các kết quả trận đấu trong khuôn khổ Champions League (UCL) từ trang [**thethao247.vn**](https://thethao247.vn/) **(Thể thao 247)**. Dữ liệu cụ thể cho việc trích xuất bao gồm thời gian, tên đội nhà, tỉ số và tên đội khách.

1. **Phương pháp**

* **Các công nghệ được sử dụng trong dự án:**
  + Scrapy: 1 framework cho việc thu thập dữ liệu từ website cho Python.
  + Beautiful Soup: 1 thư viện dùng cho việc phân tích các cú pháp trong các tài liệu HTML và XML.
  + Python: Ngôn ngữ lập trình được dùng để phát triển.
* **Quá trình thu thập dữ liệu:**
  + Khởi tạo: 1 Scrapy spider được khởi tạo để bắt đầu việc thu thập dữ liệu từ phần kết quả của trang Thể thao 247.
  + Trích xuất dữ liệu: Spider sẽ phân tích cú pháp ở trang chính để trích xuất thời gian, tên đội nhà, tỉ số và tên đội khách.
  + Lưu trữ dữ liệu: Dữ liệu được trích xuất sẽ được lưu trữ trong một định dạng có cấu trúc (JSON, CSV,..) cho việc phân tích sau này.

1. **Kết quả**

* Tổng số kết quả đã thu thập: Spider đã thu thập thành công và trích xuất được 34 kết quả trận đấu.
* Mẫu dữ liệu trích xuất:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Thời gian** | **Đội nhà** | **Tỉ số** | **Đội khách** |
| 1 | FT | Sporting CP | 0 – 3 | Dortmund |
| 2 | FT | Manchester City | 2 – 3 | Real Madrid |
| … | … | … | … | … |

1. **Những lỗi đã gặp**

* 503 Errors: Spider thỉnh thoảng gặp phải lỗi “503 Services Unavailable”, thứ yêu cầu triển khai logic thử lại và xoay vòng user-agent để khắc phục.

1. **Kết luận**

* Dự án thu thập dữ liệu đã thành công trích xuất một bộ dữ liệu gồm các kết quả trong khuôn khổ Champions League từ trang Thể thao 247. Dữ liệu được thu thập có thể được sử dụng cho nhiều mục đích khác nhau như cung cấp thông tin chuyên sâu về trận đấu.

1. **Đề xuất**

* Tiếp tục theo dõi: Triển khai một lịch trình cho việc thu thập dữ liệu để giữ cho bộ dữ liệu luôn được cập nhật.
* Làm giàu dữ liệu: Kết hợp dữ liệu đã thu thập được với các bộ dữ liệu khác, có thể phân tích sâu hơn về chúng (ví dụ: thông số trận đấu)

1. **Phụ lục**

* Code repository:
* File dữ liệu mẫu: