**Biên bản họp nhóm**

**Đồ án môn: Nhập môn Khoa học Dữ liệu – 18\_21**

* **Danh sách thành viên**

|  |  |
| --- | --- |
| Họ tên | MSSV |
| Trần Xuân Quý | 18120231 |
| Trương Minh Tài | 18120544 |

* **Tóm tắt các buổi họp**
* **Buổi 1:**
* Thời gian: thứ 7, 19/12/2020
* Nội dung:
* Tạo thùng chứa đồ án trên Github.
* Tìm hiểu các nguồn dữ liệu và các thuộc tính của dữ có thể thu thập
* Thực hiện bước thu thập dữ liệu.
* Kết quả:
* Hoàn thành việc tạo thùng chứa trên Github.
* Lựa chọn trang web kickstarter làm nguồn thu thập dữ liệu.
* Xác định phương thức thu thập dữ liệu ban đầu: parse html sử dụng thư viện Selenium.
* Công việc:
* Phần chung:
  + Tìm hiểu dữ liệu được định dạng như thế nào trên trang web.
  + Xác định cụ thể hơn về dữ liệu sẽ thu thập, các thuộc tính cần thu thập, số mẫu, …
* Phần riêng
* Trần Xuân Quý: Cài đặt chức năng sử dụng webdriver và thư viện Selenium kết nối tới các trang web cần thiết.
* Trương Minh Tài: Cài đặt chức năng lấy ra những dữ liệu cần thiết và lưu lại kết quả.
* **Buổi 2:**
* Thời gian: thứ 7, 26/12/2020
* Nội dung:
* Trình bày về việc lựa chọn các thuộc tính và cài đặt chức năng đã được giao.
* Thảo luận việc đặt câu hỏi cần trả lời và ý nghĩa câu hỏi.
* Tìm hiểu, góp ý về cài đặt chức năng của các thành viên.
* Kết quả:
* Xác định đối tượng thu thập dữ liệu là thông tin về các dự án trên kickstarter.
* Xác định các thuộc tính cần thu thập:
  + Id
  + Tên
  + Lời giới thiệu
  + Số tiền mục tiêu
  + Số tiền đã kêu gọi
  + Quốc gia
  + Hạn chót
  + Ngày tạo
  + Ngày khởi động
  + Số người tham gia hỗ trợ
  + Tỉ giá ngoại tê
  + Thể loại
  + Thể loại chính
  + Phần trăm mục tiêu đạt được
  + Kết quả gọi vốn của dự án.
* Công việc:
* Phần chung:

Tìm hiểu các bước tiền xử lý có thể có với các thuộc tính.

* Phần riêng:
* Trần Xuân Quý: Hoàn thiện chức năng sử dụng webdriver và thư viện Selenium kết nối tới các trang web cần thiết.
* Trương Minh Tài: Hoàn thiện chức năng lấy ra những dữ liệu cần thiết và lưu lại kết quả.
* **Buổi 3**
* Thời gian: thứ 7, 02/01/2021
* Nội dung:
* Thảo luận khám phá dữ liệu và thử nghiệm các phương thức tiền xử lý.
* Xác định câu hỏi cần trả lời và ý nghĩa của câu hỏi.
* Kết quả:
* Xác định câu hỏi cần trả lời:

Câu hỏi đặt ra là: kết quả gọi vốn của một dự án góp vốn cộng đồng có thể được tính như thế nào từ các thông tin đả thu thập được.

Trả lời được câu hỏi này có thể phần nào giúp cho một hoặc một nhóm những người phát triển sản phẩm độc lập đang cần yêu cầu góp vốn cộng đồng cho sản phẩm của mình có thể dự đoán được kết quả của việc gọi vốn và từ đó có những quyết định có lợi nhất cho dự án của mình.

* Xác định các yếu tố cần khám phá:

1. Các dòng bị trùng, bị thiếu.
2. Kiểu dữ liệu, phân bố giá trị của các cột.
3. Tỷ lệ các giá trị phân lớp.
4. Các cột không có nhiều ý nghĩa.
5. Các giá trị có thể tính từ các cột.
6. Rời rạc hóa một số cột có giá trị liên tục (nếu có).

* Phân công:
* Trần Xuân Quý: công việc 1, 2, 4
* Trương Minh Tài: công việc 3, 5, 6
* **Buổi 4**
* Thời gian: thứ 7, 09/01/2021
* Nội dung:
* Xác định các phương pháp tiền xử lý và mô hình hóa cuối cùng.
* Xây dựng pipeline.
* Chọn các siêu tham số để thử nghiệm.
* Áp dụng mô hình tốt nhất với tập kiểm tra.