**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MÌNH**

**ĐẠI HỌC KHOA HỌC TỰ NHIỀN**

**KHOA: CNTT**



**ỨNG DỤNG PHÂN TÍCH DỮ LIỆU THÔNG MINH – 19\_21**

**PROPOSAL ĐỒ ÁN MÔN HỌC**

|  |  |
| --- | --- |
| Nhóm K18:  18120203 - Vũ Đăng Hoàng Long  18120212 - Huỳnh Long Nam  18120249 - Hồ Minh Trí  18120462 - Nguyễn Thị Mận  18120505 - Đào Quốc Phong  18120544 - Trương Minh Tài  18120579 - Đặng Minh Thọ | Giáo viên hướng dẫn:  Nguyễn Tiến Huy |

Với sự bùng nổ của thương mại điện tử, rất nhiều hình thức kinh doanh truyền thống đã thay đổi, thích nghi với công nghệ. Trong đó, không thể không kể đến các ứng dụng đặt đồ ăn nhanh: Grabfood, Shopeefood, Baemin, Gofood,… Mang lại sự tiện lợi cho cuộc sống, cũng như tạo ra nhiều việc làm.

Nhóm K18 nhận thấy đề tài này có nhiều thông tin hữu ích có thể khai thác, đồng thời đòi hỏi nhiều kỹ năng như thu thập và xử lý dữ liệu, tìm kiếm và phân tích các đặc trưng, cuối cùng là trực quan, minh họa các đặc trưng tìm được. Vì những lý do trên, nhóm xin lựa chọn đề tài này cho đồ án môn học.

# **Các ý tưởng sẽ khai thác**

Dựa trên thông tin nhóm thu thập được trên các web đặt thức ăn:

* Vị trí quán (quận, thành phố,…)
* Danh mục quán ăn (đồ ăn nhanh, ăn vặt, trà sữa,…)
* Loại quán ăn(lề đường, quán ăn, nhà hàng,…)
* Thời gian hoạt động (8h - 22h, 9h - 23h59,…)
* Menu và biểu giá
* Các tiêu chí đánh giá (giá cả, không gian, chất lượng,…)
* Nhận xét của người dùng

…

Nhóm nhận thấy có thể trả lời được các câu hỏi sau:

* Phân bố của các loại món ăn theo khu vực => sự đa dạng món ăn theo khu vực
* Mặt bằng chung giá cả của từng khu vực => phần nào phản ánh mức sống của khu vực
* Giá cả của cùng 1 món ăn giữa các sàn; Giá gốc/Giá khuyến mãi => chính sách bán hàng, kích cầu của các sàn
* Số lượng đánh giá, số người đánh giá, các bình luận ảo, spam => phân loại các quán ăn tin cậy
* Tương quan giữa giá cả và đánh giá => tương quan giữa giá cả và chất lượng dịch vụ

# **Các công cụ, thư viên sử dụng**

Để trả lời các câu hỏi trên, nhóm đề ra quy trình và các công cụ như sau

* Thu thập và tiền xử lý dữ liệu: python(request, beautifulsoup, selenium, pandas,…), javascript
* Lưu trữ dữ liệu: csv, excel hoặc có thể là mysql
* Khám phá, khai thác, phân tích dữ liệu: python (numpy, pandas, matplotlib, seaborn,…) nếu có thể sẽ khai thác các bình luận (các model NLP)
* Trực quan kết quả đạt được: Power BI

Các công cụ được nhóm lựa chọn bởi sự phổ biến, nhiều nguồn tài liệu để tham khảo, cũng như được miễn phí, open-source dễ tiếp cận.

# **Kế hoạch theo tuần**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | | 12 | 13 | 14 | | 15 | |
| Overall | - Lập nhóm - Tìm insights - Đưa proposal | | - Thu thập data của các trang web đặt đồ ăn phổ biến - Làm sạch dữ liệu - Khám phá dữ liệu - Đề xuất insights dựa trên dữ liệu | | | | - Xử lý dữ liệu - Mô hình hóa dữ liệu (nếu cần) - Phân tích, trả lời câu hỏi đặt ra | | | | | Minh họa dữ liệu | | | | Review | | Báo cáo | |
| Long |  | | Đề xuất các insights |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Mận | Crawl Shopeefood |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Nam | Support Crawl data |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Phong | Plan B: Bất động sản |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Tài | Crawl Foody |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Thọ | Plan C: Tuyển dụng |  |  |  |  |  |  |  |  | |  |  |  | |  | |
| Trí | Crawl Grabfood |  |  |  |  |  |  |  |  | |  |  |  | |  | |

Kế hoạch dự kiến của nhóm gồm 13 tuần (tuần 3 -> tuần 15) như sau:

Thời gian được chia làm 4 phần chính:

* Thu thập dữ liệu (tuần 3 -> tuần 6): đây là giai đoạn khó khăn nhất của đồ án, dùng nhiều công để thu thập dữ liệu từ web.
  + Nếu có thể, nhóm sẽ thử thu thập dữ liệu từ mobile app, vì đối tượng của các ứng dụng đặt đồ ăn là người dùng mobile, do đó, dữ liệu trên mobile app cũng đầy đủ, chi tiết hơn.
  + Vì dữ liệu khá lớn(~300k hàng quán, theo đó các món ăn, biểu giá và bình luận người dùng), nhóm sẽ tìm công cụ lưu trữ phù hợp, có thể là mysql, cũng như mô hình hóa dữ liệu cho phù hợp.
* Xử lý, phân tích (tuần 7 -> tuần 10):
  + Từ dữ liệu thu thập được, nhóm đặt ra các câu hỏi, giả thuyết, và dùng python để khám phá, phân tích, chứng minh các giả thuyết.
  + Có thể sử dụng một số mô hình học sâu để rút trích đặc trưng từ các bình luận
* Trực quan hóa (tuần 11 -> tuần 13):
  + Hoàn thiện quy trình thu thập, xử lý, phân tích
  + Minh họa các insights tìm được bằng Power BI