**HƯỚNG DẪN GÁN NHÃN**

1. **Mục tiêu của bài toán**

* Bài toán có 2 mục tiêu:
  + Thứ nhất: Xác định chủ đề được đề cập đến trong bình luận (4 nhãn tương ứng với 4 chủ đề đặt ra)
  + Thứ hai: Nhận biết được bình luận thể hiện thái độ hài lòng, trung tính hay không hài lòng (3 nhãn)

1. **Ý nghĩa mỗi nhãn và trường hợp sử dụng**

**Đối với nhãn chủ đề:**

* Có 4 nhãn: *Service, Infrastructure, Sanitary, Location.*
  + *Service*: bình luận được gán nhãn *Service* khi có đề cập đến vấn đề về chất lượng dịch vụ của khách sạn như thái độ phục vụ, nhân viên, thức ăn, giải trí… (keyword: nhân viên, lễ tân, lịch sự, phục vụ, tư vấn, dịch vụ, tiện ích, giúp đỡ, hỗ trợ, phản hồi, di chuyển, ngon…)
  + *Infrastructure*: bình luận được gán nhãn *Infrastructure* khi có đề cập đến vấn đề về cơ sở vật chất của khách sạn như phòng ốc, nhà hàng, sân vườn, nội thất… (keyword: phòng, khách sạn, nhà hàng, bàn ghế, giường ngủ, phòng tắm, nhà vệ sinh, rộng, hẹp, đầy đủ, tiện nghi, hệ thống, sảnh…)
  + *Sanitary*: bình luận được gán nhãn *Sanitary* khi có đề cập đến vấn đề về vệ sinh của khách sạn (keyword: nhà vệ sinh, sạch sẽ, sạch, đẹp, thoáng mát, dọn, dọn dẹp, vệ sinh...)
  + *Location*: bình luận được gán nhãn *Location* khi có đề cập đến vấn đề về vị trí của khách sạn như view khách sạn, quan cảnh xung quanh, gần sông, hồ, núi, biển (keyword: view đẹp, trung tâm, vị trí, khung cảnh, biển, núi, sông, hồ, đường, …)
* Lưu ý:
  + Đối với các bình luận có các cụm từ có ý bao quát như *mọi thứ, tất cả* thì tất cả các nhãn Service, Infrastructure, Sanitary gán là 1.
  + Đối với các bình luận ngắn gọn như *Rất tốt, tuyệt* vời thì tất cả 4 nhãn này đều gán là 0.

**Đối với nhãn thái độ:**

* Có 3 nhãn là: *Satisfied, Normal, Unsatisfied* được gán các giá trị lần lượt tương ứng là *2, 1, 0* tương ứng với mức độ đánh giá của khách hàng là *hài lòng, bình thường* và *không hài lòng*.
  + *Satisfied* (2): bình luận được gán nhãn này nếu thế hiện được sự hài lòng của người đánh giá. Đối với các đánh giá vừa thể hiện sự hài lòng lẫn không hài lòng, nếu như đánh giá hài lòng chiếm ưu thế hơn thì gán nhãn này
  + *Normal* (1): bình luận được gán nhãn Normal khi không thể hiện thái độ gì với khách sạn, hoặc bình luận có các đánh giá hài lòng bằng các đánh giá không hài lòng. Với từ “như hơi, tạm ổn, cũng được, không quá” thì mang nghĩa trung tính
  + *Unsatisfied* (0): bình luận được gán nhãn Unsatisfied khi thể hiện thái độ không hài lòng với khách sạn, hoặc bình luận có các đánh giá không hài lòng chiếm ưu thế.

1. **Độ phủ của nhãn**

* Đối với nhãn chủ đề: Có 4 nhãn: *Service, Infrastructure, Sanitary, Location*. Một bình luận được gán 1 hoặc nhiều nhãn trong số 4 nhãn kể trên. Với các nhãn có trong bình luận thì tại cột nhãn đó được đánh dấu là 1 còn với các nhãn còn lại là 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Comment | Service | Infrastructure | Sanitary | Location |
| Phòng\_ốc sang\_trọng, tiện\_nghi sạch\_sẽ, nhân\_viên dễ\_thương | 1 | 1 | 1 | 0 |

* Đối với nhãn thái độ: Mỗi bình luận được gán 1 trong 3 nhãn tương ứng với mức độ hài lòng của người bình luận.

|  |  |
| --- | --- |
| Comment | Attitude |
| Phòng\_ốc đẹp, thoáng\_mát, nhân\_viên nhiệt\_tình, thức\_ăn ngon | 2 |
| Nhận phòng quá lâu, vệ\_sinh vô\_cùng kém | 0 |
| Dịch\_vụ tạm ổn, không có gì đặc biệt | 1 |

* Gán nhãn cho 2231 bình luận đầu tiên của bộ dữ liệu

1. **Thực hiện gán nhãn**

* Thực hiện gán nhãn bằng công cụ Google Sheet. Thực hiện gán trực tiếp trên bảng tính.