

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



Nhập môn khoa học dữ liệu

Đề Tài: Dự đoán số sao khách sạn

Giảng viên hướng dẫn: PGS.TS Thân Quang Khoát

Sinh viên: Đào Minh Tuấn – 20173437

Nguyễn Quý Phúc - 20173302

Đào Minh Đức - 20173029

Nguyễn Thế Tùng Dương - 20173060

Đồng Văn Hiệp - 20173104

Hà Nội - 12/2020

MỤC LỤC

1. Mở đầu	3
2. Giới thiệu về đề tài.....	3
2.1. Các vấn đề của dự án	3
➤ Thu thập dữ liệu.....	3
➤ Xử lý dữ liệu nhiễu.....	3
➤ Xây dựng mô hình dự đoán.....	3
➤ Thực hiện dự đoán thực tế.....	3
2.2. Tính khả thi của dự án.....	3
2.3. Tài liệu tham khảo.....	4
3. Mục tiêu của đề tài.....	4
4. Giới thiệu về Selenium.....	4
5. Thuật toán Random Forests	6
6. Thuật toán KNN	7
7. Các bước thực hiện đề tài	9
8. Thực nghiệm.....	19

1. Mở đầu

Trí tuệ nhân tạo (Artificial Intelligent) hay học máy (Machine Learning) là một lĩnh vực đang được nhắc đến khá nhiều trong thời gian gần đây bởi tính ứng dụng của nó trong thực tiễn. Có rất nhiều ứng dụng của học máy đã được áp dụng trong cuộc sống hàng ngày như: google dịch, xe ô tô tự lái, hệ thống gợi ý mua hàng, mô hình dự đoán giá nhà đất... Trong chương trình của môn Khoa học dữ liệu, chúng em đã được tìm hiểu rất nhiều kiến thức về những mô hình học máy cũng như xử lý dữ liệu và dựa trên kiến thức đã được học, nhóm đã thống nhất chọn đề tài Dự đoán số sao của khách sạn. Đây là 1 đề tài đòi hỏi kỹ năng thu thập, xử lý dữ liệu và xây dựng mô hình học máy.

2. Giới thiệu về đề tài

2.1. Các vấn đề của dự án

- Thu thập dữ liệu.
- Xử lý dữ liệu nhiễu.
- Xây dựng mô hình dự đoán.
- Thực hiện dự đoán thực tế.

2.2. Tính khả thi của dự án

a, Tính khả thi về công nghệ

Hiện nay có rất nhiều thư viện hỗ trợ cho việc crawl data, tuy nhiên thời gian crawl là rất lâu, nếu không có nhiều máy tính sẽ thực hiện khó khăn.

b, Tính thực tiễn của dự án

Do dữ liệu được lấy trên trang Agoda.com, những thông tin trên trang web này có thể chưa thực sự đầy đủ và chính xác, thiếu 1 số trường liên quan trực tiếp đến việc dự đoán số sao khách sạn, hơn nữa tiêu chuẩn đặt số sao của khách sạn ở Việt Nam cũng chưa thực sự chuẩn xác và nghiêm ngặt nên project mang tính tham khảo và học hỏi thêm, nếu muốn đưa vào thực tế thì cần có dữ liệu tốt hơn.

2.3. Tài liệu tham khảo

- Slide Nhập môn khoa học dữ liệu – PSG.TS Thân Quang Khoát.
- <https://towardsdatascience.com/web-scraping-using-selenium-python-8a60f4cf40ab>
- <https://viblo.asia/p/gioi-thieu-ve-matplotlib-mot-thu-vien-rat-huu-ich-cua-python-dung-de-ve-do-thi-yMnKMN6gZ7P>
- Trong bản báo cáo của nhóm em, ở phần các bước thuật toán Random Forests và KNN do em trích phần thuật toán trong slide của Thầy nên em xin phép không dịch phần thuật toán sang tiếng Việt để giữ nguyên nghĩa gốc theo ý tác giả.

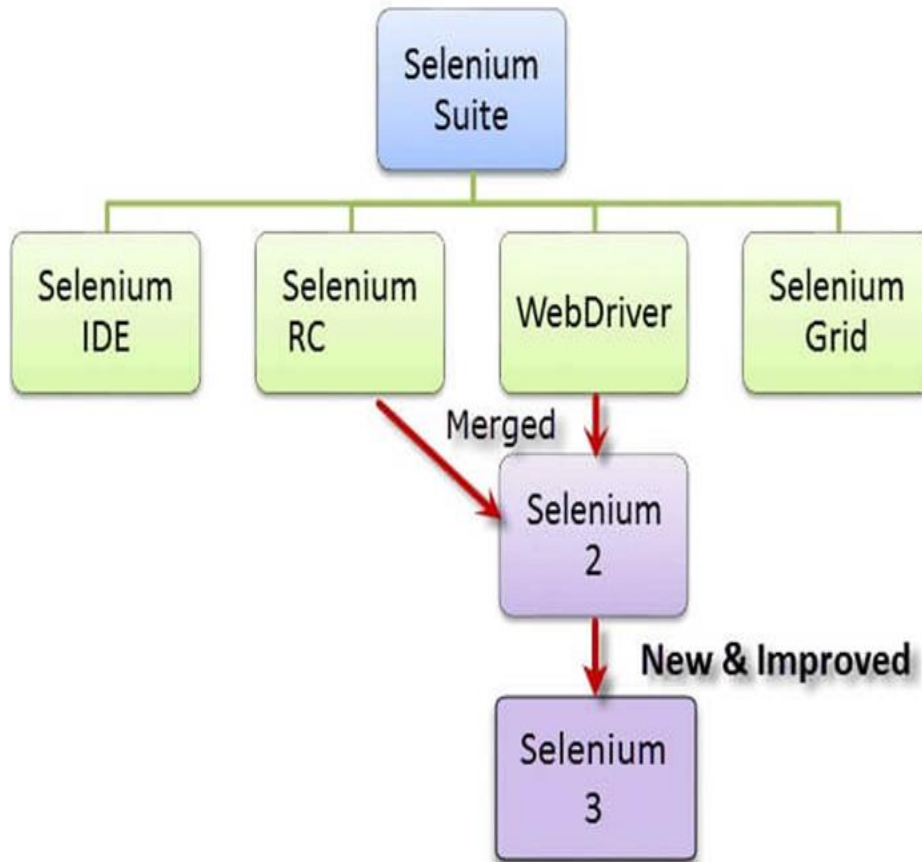
3. Mục tiêu của đề tài

Mục tiêu cơ bản của đề tài là dự đoán nhanh chóng số sao của khách sạn dựa trên những đặc trưng cho trước. Hiểu được cách thức crawl dữ liệu, xử lý dữ liệu và cách hoạt động của mô hình học máy.

4. Giới thiệu về Selenium

- Selenium là bộ kiểm thử tự động miễn phí (mã nguồn mở) dành cho các ứng dụng web trên các trình duyệt và nền tảng khác nhau.
- Selenium không chỉ là 1 công cụ độc lập mà là 1 bộ công cụ của phần mềm, mỗi bộ đều đáp ứng được nhu cầu kiểm thử khác nhau của 1 tổ chức. Nó có 4 thành phần:
 - + Selenium Integrated Development Environment (IDE).
 - + Selenium Remote Control (RC).
 - + WebDriver.
 - + Selenium Grid.

Hiện tại, Selenium RC và WebDriver được hợp nhất thành một framework duy nhất để tạo ra Selenium 2. Còn Selenium 1 thì tham chiếu đến Selenium RC.



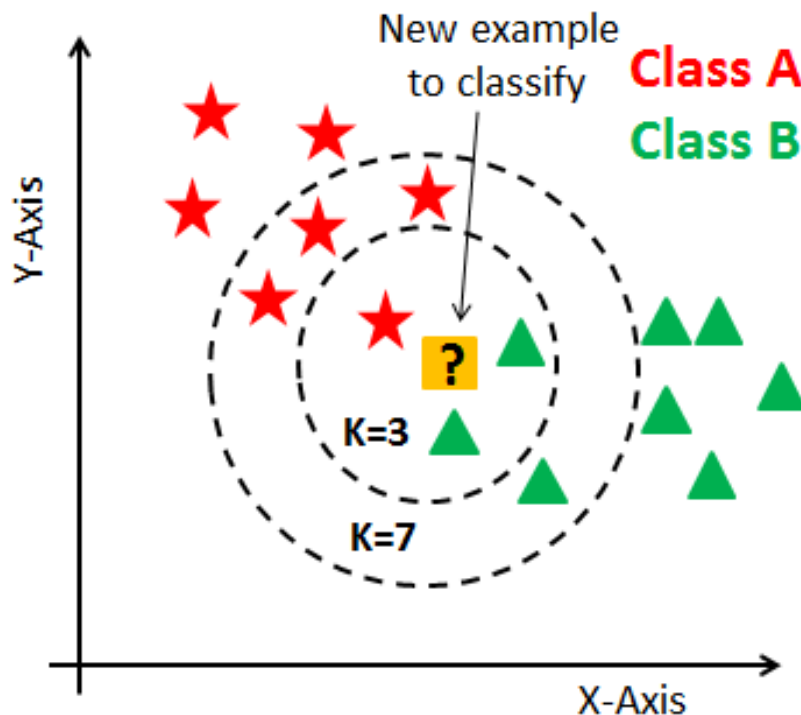
- Sự ra đời của WebDriver: Simon Stewart đã tạo ra WebDriver khoảng năm 2006 khi các trình duyệt và các ứng dụng web đang trở nên mạnh hơn và hạn chế hơn với các chương trình JavaScript như Selenium Core. Đây là khung thử nghiệm nền tảng đầu tiên có thể điều khiển trình duyệt từ cấp hệ điều hành.
- Giới thiệu qua về WebDriver: WebDriver thực hiện tiếp cận hiện đại và ổn định hơn trong tự động hoá các hành động của trình duyệt. WebDriver, không giống như Selenium RC, không phụ thuộc vào JavaScript cho Tự động hóa. Nó điều khiển trình duyệt bằng cách liên lạc trực tiếp với nó. Hỗ trợ các ngôn ngữ như: Java, Python, C#, PHP,...
- Tổng kết lại, khi crawl dữ liệu trên trang web agoda.com, nếu sử dụng Scrapy thì rất khó hoặc không thể lấy được những link khách sạn vì trang web đã được tối ưu bằng javascript, Selenium được sử dụng để tự động hóa thao tác với trình duyệt như 1 người dùng thực sự nên ta có thể sử dụng nó trong Project này, tuy nhiên thời gian chạy của Selenium là khá chậm.

5. Thuật toán Random Forests

- Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. Random forests có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng.
- **Ưu điểm:** Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting.
- **Nhược điểm:** Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian.
- **3 Thành phần cơ bản của Random Forests:**
 - + Randomization and no pruning.
 - + Combination.
 - + Bagging.
- **Thuật toán:**
 - + Input: training data D.
 - + Learning: grow K trees as follows:
 - Generate a training set D_i by sampling with replacement from D.
 - Learn the i^{th} tree from D_i :
 - + At each node:
 - Select randomly a subset S of attributes.
 - Split the node into subtrees according to S.
 - + Grow this tree up to its largest size without pruning.
 - + Prediction: taking the average of all predictions from the individual trees.

6. Thuật toán KNN

- K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất trong Machine Learning. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based, Lazy learning hay Memory-based learning.
- **Ý tưởng của thuật toán:**
 - + Không có giả định cụ thể nào về hàm được học.
 - + Giai đoạn học chỉ lưu trữ tất cả dữ liệu đào tạo.
 - + Dự đoán 1 trường hợp mới dựa trên các “hàng xóm” gần nhất của nó trong dữ liệu đào tạo.
- ⇒ Như vậy KNN được gọi là phương pháp phi tham số (Không có giả định cụ thể về classifier/regressor).
- **2 thành phần chính của thuật toán:**
 - + The similarity measure (distance) between instances/objects.
 - + The neighbors to be taken in prediction.



Distance functions

Euclidean $\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$

Manhattan $\sum_{i=1}^k |x_i - y_i|$

Minkowski $\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$

➤ **Ưu điểm:**

- + Thuật toán đơn giản, dễ triển khai.
- + Độ phức tạp tính toán nhỏ.
- + Xử lý tốt với tập dữ liệu nhỏ.

➤ **Nhược điểm:**

- + Với K nhỏ dễ gặp nhiễu dẫn tới kết quả đưa ra không chính xác.
 - + Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.
 - + Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.
- Chuẩn hóa thuộc tính, trọng số của các thuộc tính đôi khi rất quan trọng để KNN hoạt động tốt hơn.
- Đặc biệt KNN có thể xử lý missing value rất tốt, trong nhiều trường hợp nó sẽ tốt hơn sử dụng giá trị trung bình của các giá trị lân cận hoặc sử dụng giá trị có tần suất nhiều nhất để điền. Trong project của nhóm em, nhóm em sử dụng thuật toán KNN để điền 1 số trường giá trị trống.

7. Các bước thực hiện đề tài

1) Crawl dữ liệu

- Đầu tiên nhóm em vào trang <https://www.agoda.com/vi-vn/country/vietnam.html>

Khám phá các khu vực của Việt Nam

Thành phố Hồ Chí Minh	8957 khách sạn	Tỉnh Vĩnh Phúc	185 khách sạn	Tỉnh Tây Ninh	30 khách sạn
Thành phố Hà Nội	5787 khách sạn	Tỉnh Đắk Lắk	154 khách sạn	Tỉnh Thái Nguyên	30 khách sạn
Thành phố Đà Nẵng	3560 khách sạn	Tỉnh Hòa Bình	145 khách sạn	Tỉnh Vĩnh Long	29 khách sạn
Tỉnh Bà Rịa-Vũng Tàu	3091 khách sạn	Tỉnh Hà Tĩnh	133 khách sạn	Tỉnh Điện Biên	28 khách sạn
Tỉnh Lâm Đồng	2913 khách sạn	Tỉnh Sơn La	109 khách sạn	Tỉnh Cà Mau	22 khách sạn
Tỉnh Khánh Hòa	2485 khách sạn	Tỉnh Nghệ An	106 khách sạn	Tỉnh Lạng Sơn	21 khách sạn
Tỉnh Quảng Ninh	1492 khách sạn	Tỉnh Đồng Nai	103 khách sạn	Tỉnh Bạc Liêu	17 khách sạn
Tỉnh Quảng Nam	1315 khách sạn	Tỉnh Bình Dương	98 khách sạn	Tỉnh Nam Định	16 khách sạn
Tỉnh Kiên Giang	969 khách sạn	Tỉnh An Giang	97 khách sạn	Tỉnh Thái Bình	15 khách sạn
Tỉnh Bình Thuận	761 khách sạn	Tỉnh Quảng Ngãi	84 khách sạn	Tỉnh Tuyên Quang	14 khách sạn
Tỉnh Thừa Thiên Huế	651 khách sạn	Tỉnh Gia Lai	83 khách sạn	Tỉnh Đắk Nông	12 khách sạn
Tỉnh Lào Cai	604 khách sạn	Tỉnh Bắc Ninh	75 khách sạn	Tỉnh Sóc Trăng	12 khách sạn
Thành phố Hải Phòng	595 khách sạn	Tỉnh Bến Tre	74 khách sạn	Tỉnh Hà Nam	10 khách sạn
Tỉnh Bình Định	575 khách sạn	Tỉnh Cao Bằng	64 khách sạn	Tỉnh Long An	10 khách sạn
Tỉnh Ninh Bình	496 khách sạn	Tỉnh Tiền Giang	52 khách sạn	Tỉnh Bắc Giang	9 khách sạn
Tỉnh Cần Thơ	434 khách sạn	Tỉnh Bắc Kạn	48 khách sạn	Tỉnh Bình Phước	6 khách sạn
Tỉnh Phú Yên	275 khách sạn	Tỉnh Đồng Tháp	45 khách sạn	Tỉnh Hưng Yên	6 khách sạn
Tỉnh Thanh Hóa	257 khách sạn	Tỉnh Kon Tum	42 khách sạn	Tỉnh Lai Châu	6 khách sạn
Tỉnh Hà Giang	251 khách sạn	Tỉnh Yên Bái	34 khách sạn	Tỉnh Hậu Giang	5 khách sạn
Tỉnh Quảng Bình	234 khách sạn	Tỉnh Hải Dương	31 khách sạn	Tỉnh Phú Thọ	4 khách sạn
Tỉnh Ninh Thuận	187 khách sạn	Tỉnh Quảng Trị	30 khách sạn	Tỉnh Trà Vinh	4 khách sạn

- Trên đây là danh sách những thành phố có chứa khách sạn, khi kiểm tra thì em thấy tất cả phần link của các thành phố trên đều nằm trong 1 thẻ section và 1 class nên sử dụng selenium có thể dễ dàng lấy được các link thành phố và em sẽ cho vào 1 mảng (do số lượng link khá ít nên không cần lưu vào file csv).
- Sau đó em thử vào từng thành phố, thì lại xuất hiện những khu vực trong thành phố đó, cũng làm tương tự như trên để thu được tất cả các khu vực trong từng thành phố rồi cho vào 1 mảng (do số lượng link khá ít nên không cần lưu vào file csv).

Khám phá khu vực ở Hồ Chí Minh

Hồ Chí Minh

8934 khách sạn


Đi An

15 khách sạn


Cần Giờ

8 khách sạn

- Tiếp theo em vào từng khu vực, cuộn chuột xuống 1 chút thấy có Button xem tất cả số khách sạn ở khu vực đó, nên để lấy được tất cả khách sạn thì phải thực hiện sự kiện click vào button đó đối với từng khu vực, web sẽ redirect người dùng sang trang khác chứa tất cả link của khách sạn trong mỗi khu vực, tuy nhiên trang web không hề để link trang web trong phần button, trang web mới có phần đầu dạng chung là "https://www.agoda.com/vi-vn/search?" được để trong 1 đoạn script, ta có thể lấy được link trong đoạn script đó bằng các phương pháp xử lý chuỗi thông thường → lưu tất cả link này vào 1 mảng (do số lượng link khá ít nên không cần lưu vào file csv).



Hong Vina Hotel
★★ Quận 1, Hồ Chí Minh - Xem trên bản đồ
"Great staff, centrally located and nice rooms."



Rất tốt
511 nhận xét

7.7


Giá trung bình mỗi đêm
481.267 đ

Kiểm tra lượng phòng trống

Xem tất cả 7884 khách sạn ở Hồ Chí Minh

- Sau đó ta có những link chứa tất cả khách sạn của từng khu vực. Tuy nhiên nếu ta thực hiện lấy page_source của link đó luôn thì sẽ chỉ thu được 1 vài khách sạn vì trang web sử dụng javascript để load nên khi ta cuộn chuột đến đâu thì khách sạn sẽ được hiện thêm đến đó để tối ưu giao diện trang web, khi kéo được khoảng 100 khách sạn thì sẽ được chia thành trang mới. Nhóm em sẽ viết 1 hàm cuộn chuột xuống cuối cùng để tất cả khách sạn được hiện lên khi vào mỗi link (để sleep

1 vài giây để cho trang web load kịp). Để sang trang khác thì khi làm nhóm em có phát hiện được là chỉ cần sửa page=2 trên url thì có thể chuyển sang trang 2, tương tự với trang 3, 4, 5.



~~447.695~~
189.104 đ

Trang 1 trên 41

Trang kế

[Nam \(37.989\)](#) >
 [Thành phố Hồ Chí Minh \(8.957\)](#) >
 [Hồ Chí Minh \(8.934\)](#) >

Chúng tôi	Điểm du lịch	Đối tác của chúng tôi	Tải ứng dụng
ientsMAX iễn dụng io chí ặt ký mạng	Quốc gia Thành phố	Cổng thông tin đối tác YCS Giải pháp đối tác Đối tác liên kết Đối tác kết nối	Ứng dụng iOS Ứng dụng Android




























































- Đối với mỗi trang web, khi cuộn chuột xuống cuối cùng, để lấy link mỗi khách sạn thì chỉ cần sử dụng thư viện BeautifulSoup tìm kiếm những thẻ a có class là "PropertyCard__Link" rồi lưu vào file .csv.

```

<li class="hotel-item container">
  <div class="DealProperty">...</div>
  <div data-selenium="selectedHotelContainer" class="JacketContent JacketContent--DealProperty">
    <a href="/vi-vn/the-myst-dong-khoi/hotel/ho-chi-minh-city-vn.html?finalPriceView...spTypes=15,18&los=1&searchrequestid=fa105ff8-3515-463d-8dff-dd43913a7259" target="_blank" property-id="16237" data-element-index="0" class="PropertyCard__Link" data-is-singleroom="false" data-element-name="property-card-content" data-event-key="1607402922665|985">...</a>
  </div>
</li>
    
```


- Ta đã thu được danh sách khoảng gần 30000 link tại Việt Nam, tuy nhiên đây là bài toán dự đoán số sao khách sạn mà trang Agoda lại cho chung khách sạn và 1 số dạng khác như resort, nhà nghỉ, ... ở chung 1 nhóm các link nên cần xóa bỏ những link chứa “resort” “hostel” “villa”....
- ➔ Thu được 1 file csv chứa khoảng gần 8000 link khách sạn.

➤ Sau đó em sẽ trích xuất thông tin trong từng link khách sạn:


	Ngôn ngữ sử dụng <div>  tiếng Việt  Tiếng Anh  Tiếng Nhật </div>
	Khả năng tiếp cận cho người khuyết tật <div> <input checked="" type="checkbox"/> Thang máy </div>
	Truy cập Internet <div>  Wi-Fi miễn phí trong tất cả các phòng!  Wi-Fi ở nơi công cộng </div>
	Thư giãn & Vui chơi giải trí <div>  Bể bơi [ngoài trời]  Mát-xa  Spa </div> <div>  Chuyến du lịch  Phòng tập  Xông khô </div> <div>  Dịch vụ vé  Phòng xông ướt </div>
	Độ sạch sẽ và an toàn <div> <input checked="" type="checkbox"/> bộ dụng cụ sơ cứu <input checked="" type="checkbox"/> nước rửa tay </div>
	Ăn uống <div>  Dịch vụ phòng  Quán bar cạnh bể bơi  Nhà hàng </div> <div>  Quán cà phê </div>
	Dịch vụ và tiện nghi <div>  Bàn tiếp tân [24 giờ]  Được phép đưa thú nuôi vào  Nhận/trả phòng [nhanh] </div> <div>  Bảo vệ [24 giờ]  Giặt khô  Nhận/trả phòng [riêng] </div> <div>  Dịch vụ bưu chính  Giữ hành lý  Nhân viên chăm sóc khách hàng </div> <div>  Dịch vụ giặt là  hoàn toàn không hút thuốc  Nhân viên trực cửa </div> <div>  Dọn phòng hằng ngày  Két sắt  Đổi ngoại tệ </div> <div>  Khu vực hút thuốc </div>
	Đi lại <div>  Dịch vụ taxi  Đưa đón sân bay  Ô tô cho thuê </div>
	Dành cho trẻ em <div>  Phòng gia đình </div>
	Trang bị trong mọi phòng <div>  Bàn làm việc  Dép đi trong nhà  Máy sấy tóc </div> <div>  Bồn tắm  Điều hòa  Nước đóng chai miễn phí </div> <div>  Cách âm  Giá treo quần áo  TV [màn hình phẳng] </div> <div>  Các loại khăn  Két sắt trong phòng  Vòi sen </div>


Nhận phòng/ Trả phòng


 Nhận phòng từ: 14:00

 Trả phòng đến: 12:00


Di chuyển

 Phí đưa đón sân bay: 600000 VND


 Khoảng cách từ trung tâm thành phố: 0.1 km


 Thời gian đến sân bay (phút): 40


Thông tin khác


 Phí bữa sáng (áp dụng khi không bao gồm trong giá phòng): 410000 VND


Về khách sạn


 Phòng / tầng không hút thuốc: Yes

 Số tầng khách sạn: 12

 Số lượng nhà hàng: 1

 Số lượng phòng: 108

 Điện áp trong phòng: 220


 Khách sạn được xây vào năm: 2018

NHẬN XÉT TRÊN AGODA (1984)

NHẬN XÉT TRÊN BOOKING.COM (1089)

8,9/10

Tuyệt vời 

 Dựa trên 1984 bài
đánh giá

Độ sạch sẽ 9,1



Vị trí 9,3

Đáng giá tiền 8,4

— Điểm cao đối với Hồ Chí Minh

Tiện nghi 8,7

Dịch vụ 9,1

 2018  The Myst Dong Khoi

★★★★★ 6-8 Đường Hồ Huân Nghiệp, Phường Bến Nghé, Quận 1, Hồ Chí Minh, Việt Nam, 700000 - [TRÊN BẢN ĐỒ](#)

Khách sạn 5-star

Bán chạy nhất

Mới được xây dựng vào năm 2018

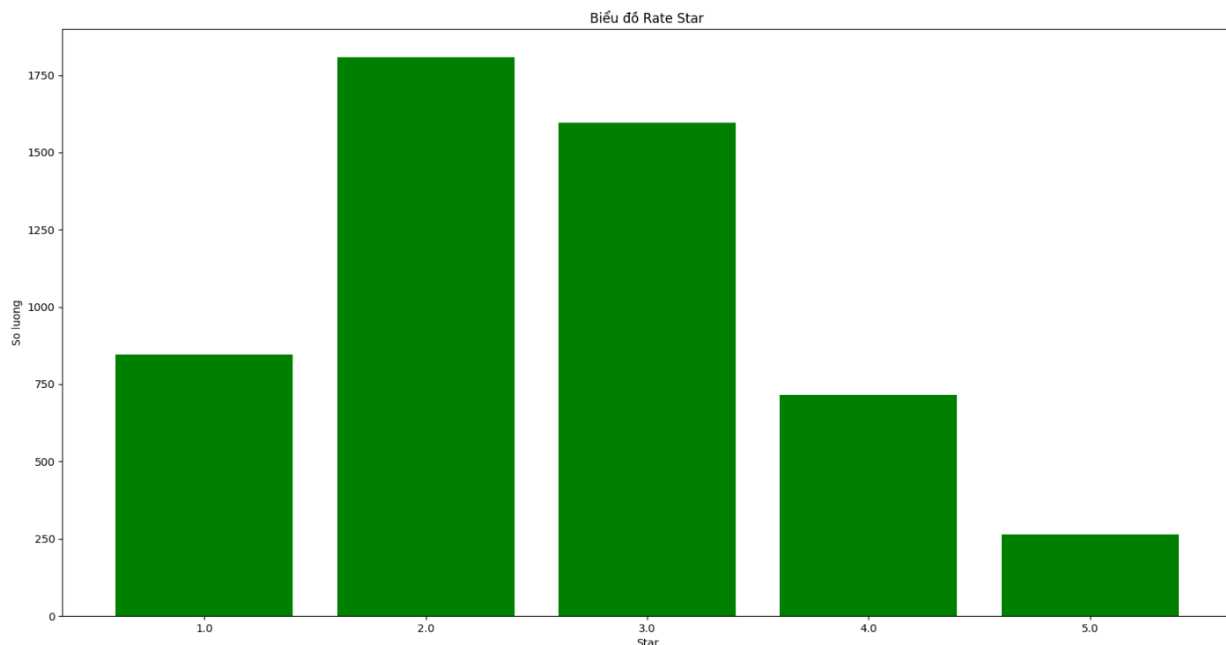
Wifi miễn phí

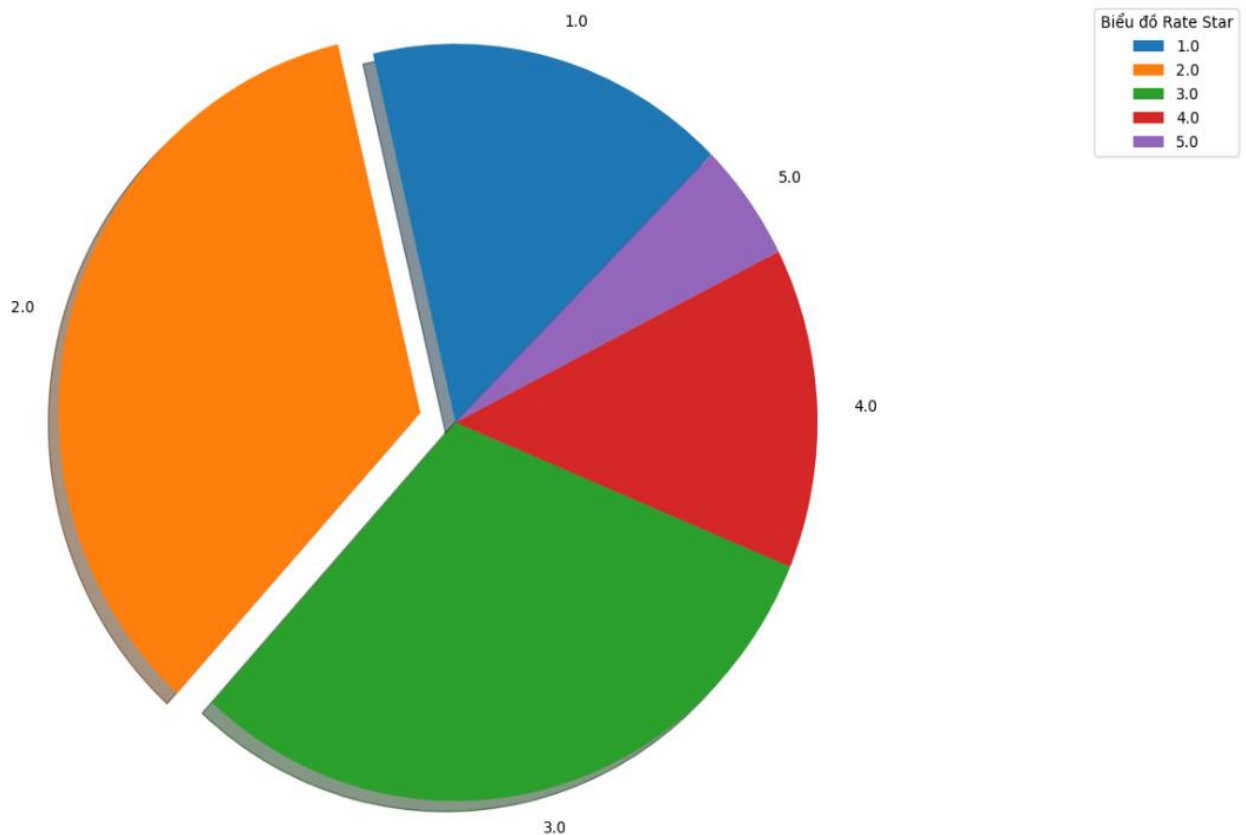
 Agoda
PREFERRED

- Thông thường mỗi khách sạn sẽ có những thông tin như trên, nhóm em có đề xuất những đặc trưng sau để lấy về: Độ sạch sẽ, Sự thoải mái và chất lượng phòng, Dịch vụ, Vị trí, Tiện nghi, Số lượng phòng, Số lượng nhà hàng, Số quán bar, Thang máy, Tiêu chuẩn về an toàn, Bể bơi, Bồn tắm, Ghế Sofa, Phòng xông hơi, Spa, Mát-xa, Phòng tập, Sân golf, Sân quần vợt. Em lấy mỗi đặc trưng dựa trên class và icon tương ứng của mỗi đặc trưng (vì icon thường là duy nhất nên lấy sẽ nhanh hơn), tuy nhiên thời gian lấy là rất lâu do phải để sleep, mỗi link lấy trong khoảng 30s-50s, nhóm em chia ra 3 máy treo khoảng 2 ngày. Một số khách sạn bị thiếu 1 số đặc tính thì tùy thuộc vào đặc tính em sẽ lưu vào file csv là NaN hoặc là 0 (Ví dụ như thông tin về sân golf nếu không có thì tức là 0 chứ không phải bị sót).

2) Tiền xử lý dữ liệu

- Sau khi thu được thông tin của từng khách sạn, trong những thuộc tính thì thuộc tính số phòng và nhãn là rất quan trọng nên những khách sạn nào không có thông tin về số phòng hoặc số sao thì em sẽ loại bỏ chúng → Thu được 5528 khách sạn cuối cùng.
- Em nhận thấy rằng có tất cả 9 nhãn: 1 sao, 1.5 sao, 2 sao, 2.5 sao, 3 sao, 3.5 sao, 4 sao, 4.5 sao, 5 sao, điều đó không hợp lý cho lắm vì thông thường khách sạn chỉ từ 1, 2, 3, 4, 5 sao nên em sẽ chuyển những sao lẻ bằng với phần nguyên của nó, ví dụ 1.5 sao thì thành 1 sao.

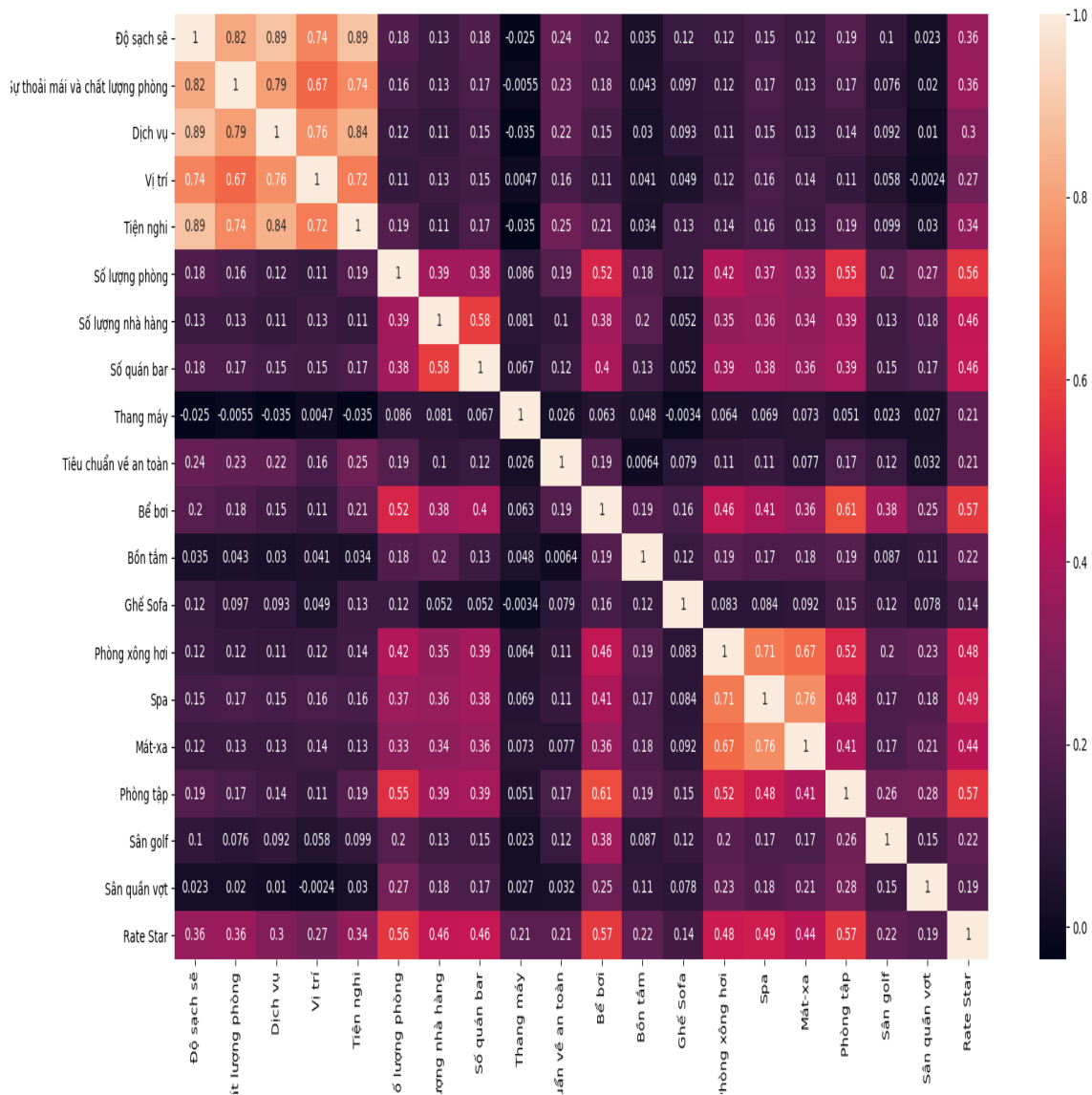




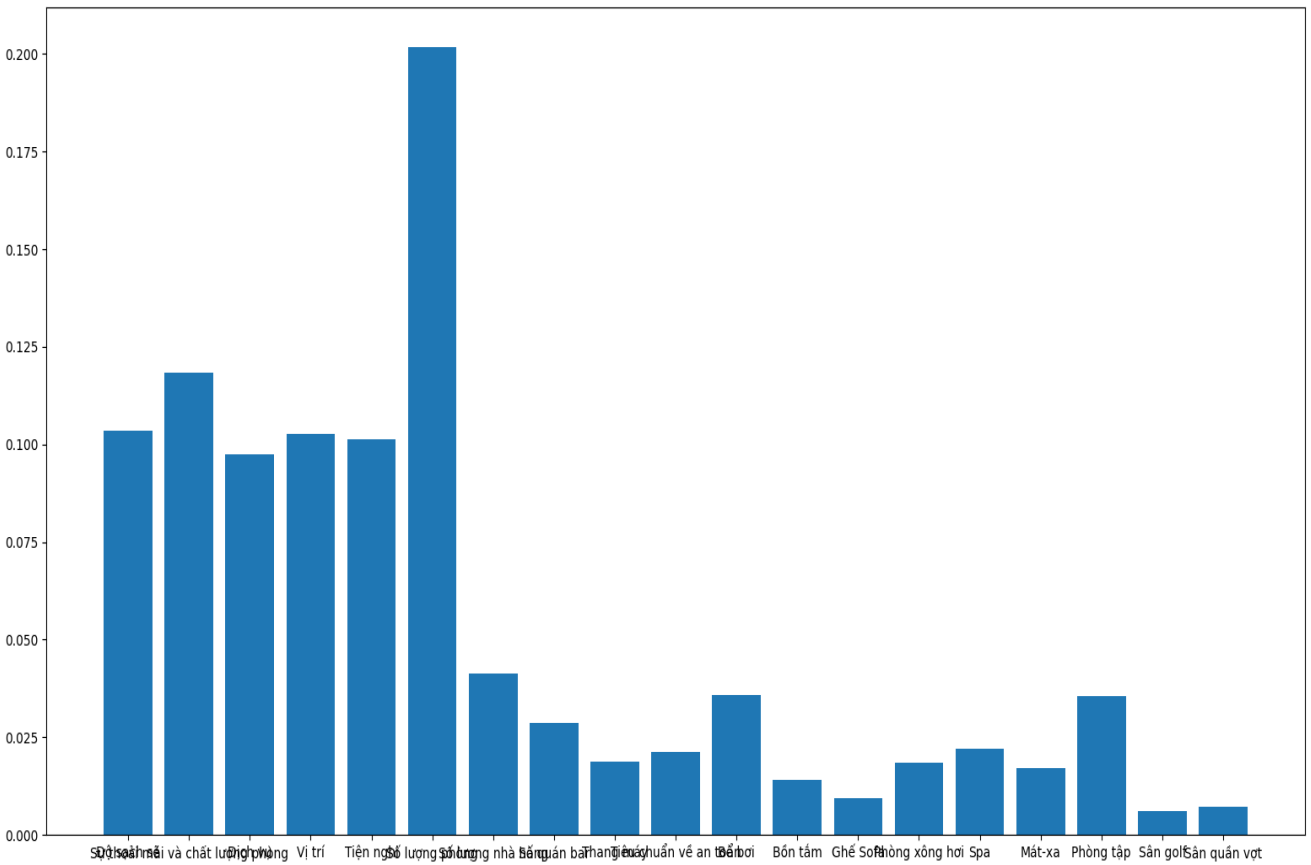
- Đây là biểu đồ phân bố nhãn (số sao) trong tập dữ liệu, nhận thấy rằng ở Việt Nam số khách sạn 5 sao là ít nhất so với còn lại, khoảng 264 khách sạn, khách sạn 4 sao khoảng 715 khách sạn, chủ yếu là khách sạn 2 sao và 3 sao với 1809 và 1595 khách sạn, trong đó khách sạn 1 sao là 845
- Em kiểm tra những feature có giá trị là NaN thì thấy có: Độ sạch sẽ, Sự thoải mái và chất lượng phòng, Dịch vụ, Vị trí, Tiện nghi là chứa giá trị NaN, em sử dụng thuật toán KNN để điền giá trị bị khuyết ở những trường này (có thể sử dụng cách điền trung bình, điền dựa trên tần suất hoặc giá trị lớn nhất, thậm chí là điền 0) tuy nhiên điền như vậy thì so với sử dụng thuật toán KNN sẽ không tối ưu bằng.

3) Xây dựng mô hình phân loại

- Nhóm em chia dữ liệu thành 2 phần là train và validate với 75% dùng làm dữ liệu huấn luyện, sử dụng thuật toán Random Forests để huấn luyện mô hình.
- Chúng em sử dụng Correlation Matrix để xem mối quan hệ giữa các biến không phụ thuộc với nhau và quan hệ của các biến phụ thuộc và các biến không phụ thuộc



- Random Forests có thể cho thấy tầm quan trọng của những đặc trưng trong mô hình:



- Ta có thể thấy rằng đặc trưng Số phòng là chiếm vai trò quan trọng nhất, đó là điều hết sức dễ hiểu trong thực tế, tiếp theo là những đặc trưng như Độ sạch sẽ, Sự thoải mái và chất lượng phòng, Dịch vụ, Vị trí, Tiện nghi. Chúng ta đưa ra 1 nhận xét đó là dù khách sạn đó có sân bóng, phòng tập, mát-xa, spa hay sân golf nhưng nếu mắc những lỗi cơ bản như không đạt độ sạch sẽ, thoải mái, chất lượng dịch vụ cũng như chất lượng cơ sở vật chất đảm bảo, ở vị trí xấu thì cũng không thể được đánh giá cao về số sao. Một khách sạn được sao cao thì số lượng phòng phải đủ lớn, những đánh giá như sạch sẽ, chất lượng, dịch vụ tiện nghi phải tốt, kèm theo đó sẽ là những dịch vụ đi kèm như golf, tennis, nhà hàng, quán bar...

4) Đánh giá mô hình

- Nhóm em sử dụng độ đo accuracy, precision, recall và F1 score để đánh giá mô hình
- Accuracy là tỉ lệ số mẫu được đoán đúng trong toàn bộ số mẫu kiểm định.
- Precision là đại lượng đo tỉ lệ nhãn được tính là True Positive trên tổng số nhãn được dự đoán là Positive

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

- Recall là đại lượng đo tỉ lệ nhãn được tính là True Positive trên tổng số nhãn thực sự là Positive.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

- F1 score là đại lượng đo trung bình điều hòa giữa Precision và Recall, chúng ta luôn mong đại lượng này càng cao càng tốt.

$$F1 = 2 \frac{Precision * Recall}{Precision + Recall}$$

- Mô hình đạt được Accuracy = 0.847 trên tập Validate, các đại lượng khác ứng với từng lớp như sau:

0.8477429227237949				
	precision	recall	f1-score	support
1.0	0.88	0.73	0.80	211
2.0	0.82	0.86	0.84	442
3.0	0.83	0.89	0.86	385
4.0	0.91	0.87	0.89	196
5.0	0.91	0.85	0.88	73
accuracy			0.85	1307
macro avg	0.87	0.84	0.85	1307
weighted avg	0.85	0.85	0.85	1307

8. Thực nghiệm

- Em sử dụng thư viện Flask trong python để deploy model lên web.
- Em viết 1 trang html, css chứa 1 form nhận đầu vào là các đặc trưng, sau khi người dùng nhập các thông tin tương ứng rồi click vào button Dự đoán thì sẽ hiển thị ra số sao của khách sạn được dự đoán.

Nhập thông tin khách sạn

Tên khách sạn:

Độ sạch sẽ (Nhập thang từ 0-10):

Sự thoải mái và chất lượng phòng (Nhập thang từ 0-10):

Dịch vụ (Nhập thang từ 0-10):

Vị trí (Nhập thang từ 0-10):

Tiện nghi (Nhập thang từ 0-10):

Số lượng phòng:

Số lượng nhà hàng:

Số quán bar:

Thang máy (1: có, 0: không có):

Dự đoán

Tiêu chuẩn về an toàn (1: có, 0: không có):

Bể bơi (1: có, 0: không có):

Bồn tắm (1: có, 0: không có):

Ghế Sofa (1: có, 0: không có):

Phòng xông hơi (1: có, 0: không có):

Spa (1: có, 0: không có):

Mát-xa (1: có, 0: không có):

Phòng tập (1: có, 0: không có):

Sân golf (1: có, 0: không có):

Sân quần vợt (1: có, 0: không có):

- Giao diện demo model em xây dựng khá đơn giản, sau khi nhập các giá trị và dự đoán thì sẽ như sau:

Nhập thông tin khách sạn

Tên khách sạn:	Tên khách sạn : Đào Minh	Tiêu chuẩn về an toàn (1: có, 0: không có):
<input type="text"/>	<input type="text"/>	<input type="text"/>
Độ sạch sẽ (Nhập thang từ 0-10):	Tuần Độ sạch sẽ : 9 Sự thoải mái : 6 Dịch vụ : 7 Vị trí : 8	Bể bơi (1: có, 0: không có):
<input type="text"/>	Tiện nghi : 8 Số lượng phòng : 40 Số nhà hàng : 1	<input type="text"/>
Sự thoải mái và chất lượng phòng (Nhập thang từ 0-10):	Số quán bar : 1 Thang máy : 1	Bồn tắm (1: có, 0: không có):
<input type="text"/>	1 Tiêu chuẩn an toàn : 1 Bể bơi : 1 Bồn tắm : 0 Ghế sofa : 1 Phòng xông hơi : 0 Spa : 1	<input type="text"/>
Dịch vụ (Nhập thang từ 0-10):	1 Mát-xa : 0 Phòng tập : 0	Ghế Sofa (1: có, 0: không có):
<input type="text"/>	Sân golf : 0 Sân quần vợt : 1	<input type="text"/>
Vị trí (Nhập thang từ 0-10):	<h2>Predicted Rate Star:</h2> <h1>3</h1>	Phòng Xông Hơi (1: có, 0: không có):
<input type="text"/>		<input type="text"/>
Tiện nghi (Nhập thang từ 0-10):		Spa (1: có, 0: không có):
<input type="text"/>		<input type="text"/>
Số lượng phòng:		Mát-xa (1: có, 0: không có):
<input type="text"/>	<input type="text"/>	<input type="text"/>
Số lượng nhà hàng:	Phòng tập (1: có, 0: không có):	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>
Số quán bar:	Sân golf (1: có, 0: không có):	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>
Thang máy (1: có, 0: không có):	Sân quần vợt (1: có, 0: không có):	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>

- Có thể nhập tiếp rồi dự đoán những khách sạn tiếp theo.
- Có thể thấy được với những thông số như trên thì 3 sao là hợp lý.

Trên đây là bản báo cáo của nhóm chúng em, chúng em mong Thầy đưa ra những góp ý và nhận xét để nhóm chúng em tăng thêm hiểu biết cũng như hoàn thiện hơn về cách làm, chúng em xin trân thành cảm ơn! Em có để code trên github:

