Organizing Data
Sắp xếp dữ liệu

Hiểu về chuẩn hóa dữ liệu

Trước khi bắt tay vào phân tích dữ liệu để trả lời các giả thuyết và câu hỏi, chúng ta cần phải hiểu được thông tin mà ta có. Dữ liệu được sắp xếp theo một bộ những quy luật được chuẩn hóa cụ thể để giúp ta nhìn dữ liệu một cách dễ dàng hơn. Trong công việc, chúng ta sẽ thường làm việc với các bảng dữ liệu trong bảng tính (spreadsheets), không phải bộ dữ liệu, nhưng nhiều nguyên tắc về cách tổ chức dữ liệu tương tự được áp dụng.

Bảng dữ liệu là một tập dữ liệu được sắp xếp thành các cột, mỗi bản ghi dữ liệu riêng biệt trong một hàng. Hệ thống sắp xếp này cho phép máy tính có thể phân tích dữ liệu và nhận ra sự tương đồng cho phép chúng ta đưa ra những kết luận chung về dữ liệu

Chuẩn hóa dữ liệu

X

Species	Habitats
AH	Shrubland, Grassland
AS	Grassland
SH	Grassland, Woodland, Shrubland

Species	Habitat 1	Habitat2	Habitat3
AH	Shrubland	Grassland	
AS	Grassland		
SH	Grassland	Woodland	Shrubland

Species	Habitat
AH	Shrubland
AH	Grassland
AS	Grassland
SH	Grassland
SH	Woodland
SH	Shrubland

Chuẩn hóa dữ liệu

Khi làm việc với một bộ dữ liệu, các thông tin có thể đến từ nhiều nguồn khác nhau, có các trường không đầy đủ, có các cấu trúc khác nhau và chứa những lỗi như nhập hai lần hoặc lỗi chính tả. Những điều này làm phức tạp quá trình phân tích và mặc dù chúng ta nhận ra những lỗi này nhưng máy tính thì không có khả năng đó

Chuẩn hóa dữ liệu hay làm sạch là quá trình làm sạch dữ liệu và là một bước quan trọng trong dữ liệu báo chí.

Một trong những bước đầu tiên trong làm sạch dữ liệu là đảm bảo tất cả các tiêu đề của các cột chính xác và đầy đủ; đồng thời các dữ liệu trong mỗi hàng khớp với tiêu đề cột.

Tiêu đề

- Dòng đầu tiên chứa tên của các biến.
- Không dùng nhiều hơn 1 dòng cho các tên biến
- Điền vào các tiêu đề còn thiếu/rỗng

Ô - cell

- Mỗi ô trong bản tính chứa một mẩu dữ liệu. Không điền nhiều hơn một thứ trong một ô
- Tránh merge các ô

Các ô trống: các lựa chọn

- Xóa những dòng trống nhưng không xóa những ô riêng lẻ trừ phi bạn chắc rằng về lý do tại sao nó trống
- Dùng một số mã (code) phổ biến cho dữ liệu bị thiếu (missing data).
- Điền vào các chỗ trống trong dữ liệu phân tách (disaggregate data) bằng (data average).

Chuẩn hóa dữ liệu

Nếu bộ dữ liệu có các trường về địa chỉ, ngày tháng, tuổi, đơn vị đo, bước đầu tiên là quyết định fomat chuẩn để điền thông tin những trường này và bộ dữ liệu

Date
12 February 2012
12/2/2012
2/12/2012
12/2/12
12/feb/2012

Chuẩn hóa dữ liệu

Không có định dạng đúng bắt buộc nào, miễn là các ngày tháng được định dạng theo cùng một cách và máy tính hiểu được định dạng ngày này. Quan trọng là cần chọn định dạng nào thuận tiện nhất cho cả bộ dữ liệu.

Trong trường hợp này, chọn kiểu DD/MM/YYYY. Dữ liệu sau khi làm sạch sẽ như thế này:

Date
12 February 2012
12/2/2012
2/12/2012
12/2/12
12/feb/2012

DATE
12/02/2012
12/02/2012
12/02/2012
12/02/2012
12/02/2012

Chuẩn hóa dữ liệu

Nguyên tắc cơ bản là đảm bảo tất cả các dữ liệu được nhập theo cùng một định dạng, thường là tất cả bằng các chữ in hoa, không có bất kỳ khoảng trắng thừa nào.

Mỗi loại dữ liệu nên có cột riêng.

Chuẩn hóa dữ liệu: Dữ liệu lộn xộn

Tên	Ngày sinh	Địa chỉ	Mức lương
Phạm Nhật Tân	16 April 1992	Bảo tàng thiên nhiên Việt Nam	$1250
Trần Thị Thanh Hải	31/5/1990	24H2 Khu đô thị mới Yên Hòa, Phường Yên Hòa, Quận Cầu Giấy, Yên Hoà, Cầu Giấy, Hà Nội	US$1000

Chuẩn hóa dữ liệu: Đã làm sạch

TÊN	NGÀY SINH	ĐỊA CHỈ	XÃ/PHƯỜNG	QUẬN/HUYỆN	TỈNH/THÀNH PHỐ	MỨC LƯƠNG (USD)
Phạm Nhật Tân	16/04/1992	18 Hoàng Quốc Việt	Nghĩa Đô	Cầu Giấy	Hà Nội	1250
Trần Thị Thanh Hải	31/05/1990	24H2 Khu đô thị mới Yên Hòa	Yên Hòa	Cầu Giấy	Hà Nội	1000

Lời khuyên khi bắt đầu đánh giá dữ liệu

- Đảm bảo số lượng bản ghi mà bạn cần phải có và bạn có đủ số lượng đó
- Kiểm tra tính nhất quán ở tất cả các trường thông tin
- Đảm bảo các tiêu đề cột rõ ràng và các đơn vị đồng nhất
- Tìm hiểu những chỗ trống hoặc N/A có nghĩa là gì
- Nếu bạn đã scrape một bản dữ liệu tóm tắt, nghĩ ra cách tốt nhất để cấu trúc tệp dữ liệu dễ dàng cho việc phân tích.