

# KHOA HỌC DỮ LIỆU

## BÁO CÁO ĐỒ ÁN

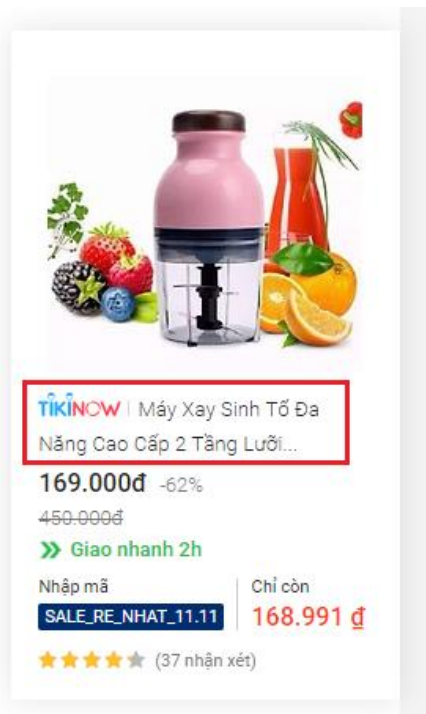
1612645 NGUYỄN ĐĂNG ANH THI

# Những phần đã update sau buổi báo cáo

- ▶ Trước đây trong file jupyter notebook chỉ tách 2 tập là tập train
- ▶ Sau chỉnh sửa đã tách thành 3 tập train, test, validation để tăng độ chính xác

# Bài toán: **Product Title Categorization** phân loại danh mục cho 1 sản phẩm

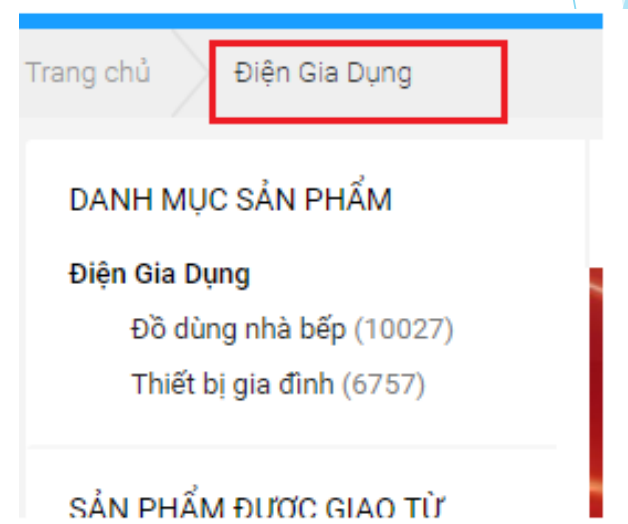
- ▶ Từ tên của sản phẩm, cần phân loại sản phẩm vào đúng danh mục



? Category

# Bài toán: phân loại danh mục cho 1 sản phẩm

- ▶ Từ tên của sản phẩm, cần phân loại sản phẩm vào đúng danh mục



# Ý nghĩa của bài toán:

- ▶ Giúp cho các nhà phân phối sản phẩm giảm thiểu công việc nhập category khi nhập các sản phẩm mới

# B1: Thu thập dữ liệu, lấy ở đâu?

## Tiki.vn

The screenshot shows the Tiki.vn homepage during a Black Friday sale. The top banner features the text "BLACK FRIDAY GIẢM ĐỒNG LOẠT ĐẾN 50% TOÀN BỘ HÀNG CHÍNH HÃNG" and "TỪ 26.11- 02.12.2019 Miễn phí vận chuyển từ 150k\* XEM NGAY". Below the banner is a navigation bar with icons for Ticketbox, Trợ lý Tiki, Ưu đãi đối tác, Đặt khách sạn, Đặt vé máy bay, Sản hàng tồn, Khuyến Mãi HOT, Hàng quốc tế, and Bán hàng cùng Tiki. The main header includes the Tiki logo, a search bar with the placeholder "Tìm sản phẩm, danh mục hay thương hiệu mong muốn...", and links for "Theo dõi đơn hàng", "Thông báo của tôi", "Đăng nhập Tài khoản", and "Giỏ hàng 0". A secondary navigation bar lists "DANH MỤC SẢN PHẨM" and "Bạn muốn giao hàng tới đâu?". The left sidebar contains a category menu with items like "Điện Thoại - Máy Tính Bảng", "Điện Tử - Điện Lạnh", "Phụ Kiện - Thiết Bị Số", "Laptop - Thiết bị IT", "Máy Ảnh - Quay Phim", "Điện Gia Dụng", "Nhà Cửa Đời Sống", "Hàng Tiêu Dùng - Thực Phẩm", "Đồ chơi, Mẹ & Bé", "Làm Đẹp - Sức Khỏe", "Thời trang - Phụ kiện", and "Thể Thao - Dã Ngoại". The main content area features a large "12.12 BẮM LÀ CÓ" flash sale banner with a man and a woman, and a "FLASH SALE CHỈ TỪ 12K" offer. To the right, there are several promotional tiles: "Ưu đãi đối tác Nhận ngay nhiều ưu đãi TỪ NGÂN HÀNG" with a 0% interest rate, "Giày thể thao đình Sale linh đình ƯU ĐÃI ĐẾN 50%", "Top sản phẩm Phụ kiện cho xe ƯU ĐÃI ĐẾN 49%", "Xả kho giá sốc Sách - VPP GIẢM ÍT NHẤT 40%", "Tã bỉm chính hãng giá tốt Phát hiện hàng giả", "Thứ 2 vui vẻ Hốt cú chốt - Giá luôn sốc", "Đầu tháng Lương về Sắm ngay Để yêu", and "Flash sale Deal bất ngờ".

# Có thể thực hiện crawl data sau khi kiểm tra file robot.txt

---

```
# Disallow all crawlers access to certain pages.
```

```
User-agent: *  
Disallow: /customer/account/  
Disallow: /customer/account/edit/  
Disallow: /customer/account/login/  
Disallow: /customer/account/create/  
Disallow: /customer/wishlist/  
Disallow: /customer/review/  
Disallow: /customer/reward/  
Disallow: /customer/bookcare/  
Disallow: /customer/notification/  
Disallow: /sales/order/history/  
Disallow: /order/tracking/  
Disallow: /checkout/cart/  
Disallow: /checkout/shipping/  
Disallow: /checkout/payment/  
Disallow: /catalogsearch/result/  
Disallow: /*q=
```

```
User-agent: adsbot-google  
Crawl-delay: 1
```

Sau khi tiến hành thu thập dữ liệu ta được bộ dữ liệu như sau

	product_title	category
3	Áo Khoác Ca Rô Cực Đẹp AK002	Thời trang - Phụ kiện
4	Bộ Thú Cưng Hoang Dã Của Bé Lego Duplo	Đồ chơi, Mẹ & Bé
5	Bộ Phát Wifi Di Động 3G Alcatel Y580 (21	Laptop - Thiết bị IT
6	3.5mm 1 to 2 Double Earphone Headphone Y	Hàng quốc tế
7	Folding Transparent 240 Holes Stud Earring	Hàng quốc tế
8	Ấm Đun Siêu Tốc AUX AK-15N01 (5L)	Điện Gia Dụng
9	Phấn phủ bột khoáng kèm dầu LUA	Làm Đẹp - Sức Khỏe
10	Chai xịt sát trùng nhanh lành vết thương cho	Hàng Tiêu Dùng - Thực Phẩm
11	Bàn Phím Microsoft Surface Go	Hàng quốc tế
12	Combo 4 Tã Dán Pampers Sơ Sinh Nội Địa Nhật	Đồ chơi, Mẹ & Bé
13	Dép cao su 2 quai chéo	Thời trang - Phụ kiện
14	Mẫu Dán Decal Điện Thoại Minion-15 (12 x 21	Phụ Kiện - Thiết Bị Số
15	Thảm Lót Sàn Xe 7 Chỗ Michelin 903-489 (4	Xe Máy, Ô tô, Xe Đạp
16	Siro Đào Vinasyrup 2200ml	Hàng Tiêu Dùng - Thực Phẩm



## B2: Tiền xử lý dữ liệu

- ▶ Xóa bỏ các stopword cho cột title như là các mạo từ: của ,các , cho, các...
- ▶ Xóa bỏ dấu chấm câu, các kí tự đặc biệt
- ▶ Xóa các kí tự chứa số
- ▶ Chuyển các kí tự hoa thành kí tự bình thường
- ▶ Ví dụ: 'Combo 4 hộp Gà Hàm Vissan(150)'  
Qua bước tiền xử lý => 'hộp gà hàm vissan'

## B2: Tiền xử lý dữ liệu

### ► Thêm cột category\_id

Dữ liệu ban đầu

	product_title	category
0	Giày Nam Thể Thao Tăng Chiều Cao 8cm Ohazo!	Thời trang - Phụ kiện
1	Bộ quần áo thun cotton 4 chiều cao cấp	Đồ chơi, Mẹ & Bé
2	Áo Khoác Ca Rô Cực Đẹp AK002	Thời trang - Phụ kiện
3	Bộ Thú Cưng Hoang Dã Của Bé Lego Duplo	Đồ chơi, Mẹ & Bé
4	Bộ Phát Wifi Di Động 3G Alcatel Y580 (21	Laptop - Thiết bị IT
5	3.5mm 1 to 2 Double Earphone Headphone Y	Hàng quốc tế
6	Folding Transparent 240 Holes Stud Earring	Hàng quốc tế
7	Ấm Đun Siêu Tốc AUX AK-15N01 (5L)	Điện Gia Dụng
8	Phấn phủ bột khoáng kèm dầu LUA	Làm Đẹp - Sức Khỏe
9	Chai xịt sát trùng nhanh lành vết thương cho	Hàng Tiêu Dùng - Thực Phẩm

Sau bước tiền xử lý

	product_title	category	category_id
0	giày nam thể thao tăng chiều cao ohazo	Thời trang - Phụ kiện	0
1	bộ quần áo thun cotton chiều cao cấp	Đồ chơi, Mẹ & Bé	1
2	áo khoác ca rô cực đẹp	Thời trang - Phụ kiện	0
3	bộ thú cưng hoang dã bé lego duplo	Đồ chơi, Mẹ & Bé	1
4	bộ phát wifi di động alcatel	Laptop - Thiết bị IT	2
5	to double earphone headphone y	Hàng quốc tế	3
6	folding transparent holes stud earring	Hàng quốc tế	3
7	ấm đun siêu tốc aux	Điện Gia Dụng	4
8	phấn phủ bột khoáng kèm dầu lua	Làm Đẹp - Sức Khỏe	5
9	chai xịt sát trùng nhanh lành vết thương	Hàng Tiêu Dùng - Thực Phẩm	6

## B3: Chuyển data từ dạng text xang vector

- ▶ Giày nam thể thao tăng chiều cao ohazo



```
[[0.236, -0.141, 0.000, 0.045],  
[0.006, 0.652, 0.270, -0.556],  
[0.305, 0.569, -0.028, 0.496],  
[0.421, 0.195, -0.058, 0.477],  
[0.236, -0.141, 0.000, 0.045],  
[0.844, -0.001, 0.763, 0.201]]
```

## B4: Dùng các mô hình máy học khác nhau để huấn luyện dữ liệu và đưa ra dự đoán

- ▶ Tách tập dữ liệu thành 3 tập: train, validation, test và thực hiện huấn luyện
- ▶ Dùng các mô hình khác nhau: naïve bayes, SGD, logistic regression.
- ▶ Dùng tập huấn luyện để train các model này, sau đó tính score với tập validate. Tìm mô hình huấn luyện có điểm số cao nhất
- ▶ Dùng mô hình tốt nhất huấn luyện lại tập train + validation sẽ cho ra mô hình tối ưu và chính xác nhất
- ▶ So sánh, đối chiếu, tính score với tập test

# Thống kê kết quả score trung bình của các mô hình

Mô hình	Score
Naïve bayes	80.60
Linear support vector machine	74.83
Logistic Regression	83.64



Ta chọn mô hình Logistic Regression để dự đoán dữ liệu tiếp theo

# Dự đoán

Đầu vào	Kết quả dự đoán
Bàn phím acer	Laptop - Thiết bị IT
lược sử loài người	Sách, VPP & Quà Tặng
Trái đất hình thành như thế nào?	Sách, VPP & Quà Tặng
Như là giấc mơ	Sách, VPP & Quà Tặng
Găng tay xe đạp	Thời trang - phụ kiện
Bình nước nóng sanyo 220ml	Điện Gia Dụng

Thank you