Cdiscount

Catégorisation de produits

Du sémantique (k-NN) à l'image (CNN)

Qui suis-je ?

Chef d'activité data chez Cdiscount
Je travaille avec deux équipes pour exploiter au maximum la donnée associée aux produits

Référentiel : structuration et complétion des données décrivant les produits chez Cdiscount

Parcours
Data scientist chez **Cdiscount** depuis 2017
Data scientist chez MFG Labs (Paris)
Thèse Thales – LIP6 (Paris)
Telecom Bretagne (Brest)

Data science
L'exploitation de sources de données à disposition pour assister l'humain dans ses activités

Modèles
Factorisation matricielle
Séparateur à Vaste Marge

Technos
SQL (PSQL/HiveQL)
Python

Sommaire

Une rapide histoire de Cdiscount

Cdiscount
Plateforme française de vente de produits et de services pour répondre aux besoins de nos clients

1998
Création à **Bordeaux** de Cdiscount
Vente de CD à prix discount

2012
Lancement de la Marketplace pour permettre à d'autres vendeurs de proposer leurs produits

2019
30 millions de produits et services disponibles sur cdiscount.com

Marketplace

Marketplace
Cdiscount héberge les propositions commerciales d'autres vendeurs et met en relation avec ses clients des vendeurs du monde entier

Offre et produit
Les vendeurs placent des offres (prix, stock, mode de livraison) sur des produits

Intégration en continu
La création manuelle est possible mais la majorité est automatisée

Volumes importants
~ 1 M demandes de créations de produits par semaine

La catégorisation

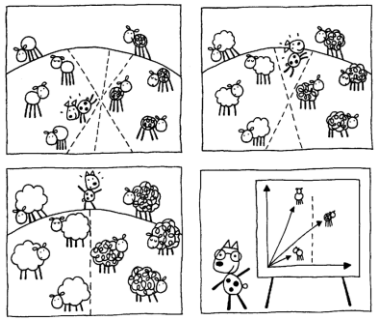
Catégorie
Une catégorie est une typologie de produit et est unique pour chaque produit.
Elle caractérise les propriétés du produit et structure sa fiche sur le site

Arborescence
Les catégories sont regroupées dans une arborescence à quatre niveaux

Evolutive
L'arbre des catégories est amélioré en continu : créations, déplacements et suppressions de catégories

8000 catégories
Au niveau le plus fin. Chaque produit doit avoir une catégorie

La classification



Apprentissage supervisé
Etant donné un ensemble d'exemples pour lesquels la réponse est connue, créer un modèle capable de prédire la meilleure étiquette possible pour de nouveaux exemples

Minimisation
Une fonction f(x) qui prédit y dont les paramètres sont estimés pour minimiser un coût

X un vecteur
Les données sont représentées par un vecteur x dont chaque dimension est une caractéristique

Feature engineering
Le choix des caractéristiques peut être manuel (ex. : TF-IDF) ou laissé au modèle (ex. : deep learning)

Illustration de Bernhard Schölkopf et Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002

Une contrainte opérationnelle forte

Précision supérieure à 90 %
Seuil exigé après estimation de la précision humaine

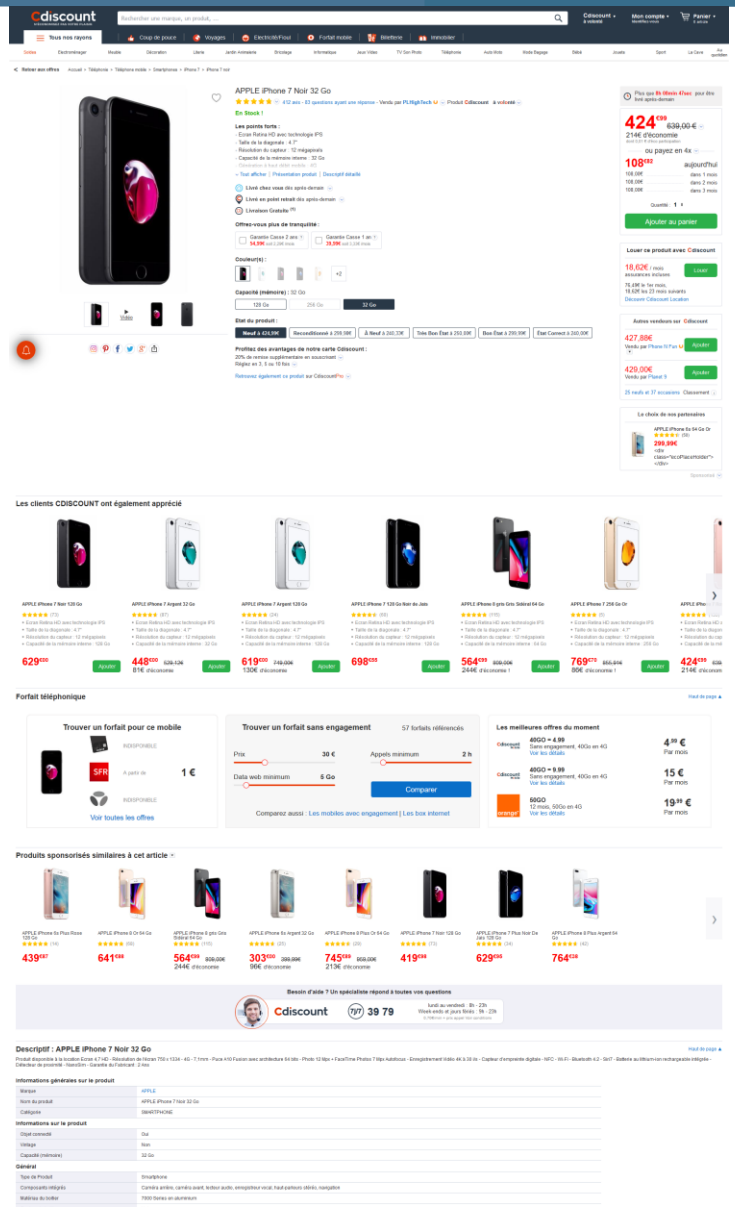
Validation manuelle
La qualité des prédictions est évaluée, sur un échantillon, chaque semaine

Taux de classification
Les produits pour lesquels le classifieur n'est pas assez confiant demandent un traitement supplémentaire

Qualité
Maintenir le catalogue de plus de 30 millions de produits bien catégorisé

Sommaire

Les éléments d’une fiche produit



Titre

Texte court caractérisant le produit visible également sur les listes de produit

Image(s)

Visuel représentant le produit, éventuellement mis en situation, visible également sur les listes de produit

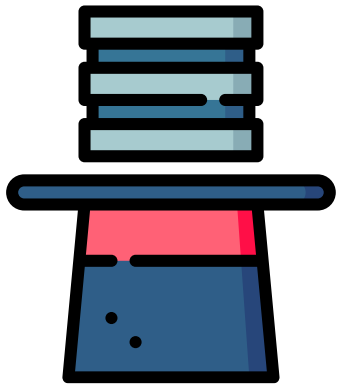
Description

Texte naturel long permettant de décrire le produit et ses caractéristiques

Caractéristiques

Texte structuré listant les propriétés techniques du produit (dimensions, poids, taille d’écran, nombre de places…)

Utiliser le texte des produits



Vectorisation
Transformation du texte naturel sous forme de vecteur pour pouvoir l'utiliser dans un algorithme d'apprentissage

Sac de mot
Représentation vectorielle où chaque dimension représente un mot observé

TF-IDF
Calcule une valeur pour chaque mot de chaque produit
TF : occurrences dans le document
IDF : inverse des occurrences dans l'ensemble des produits

Embeddings / word2vec
Apprentissage d'un espace vectoriel latent ou chaque mot est un vecteur

K plus proches voisins

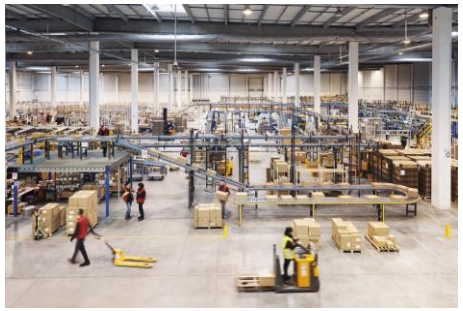
Vote des produits les plus ressemblants
1ere étape : sélection des k produits de la base d'apprentissage les plus ressemblants
2eme étape : vote entre ces k produits pour prédire la catégorie

Base d'apprentissage
Ensemble de m produits dont la catégorie est connue

Distance
Fonction attribuant une valeur réelle à une paire de vecteurs d'autant plus grande que les vecteurs sont distincts

Décision locale
Seuls les k voisins interviennent dans une décision

Une chaine complète



Garantir la précision
L'objectif premier du catégoriseur est sa précision

Utilisation de seuils de confiance pour rejeter les prédictions moins certaines

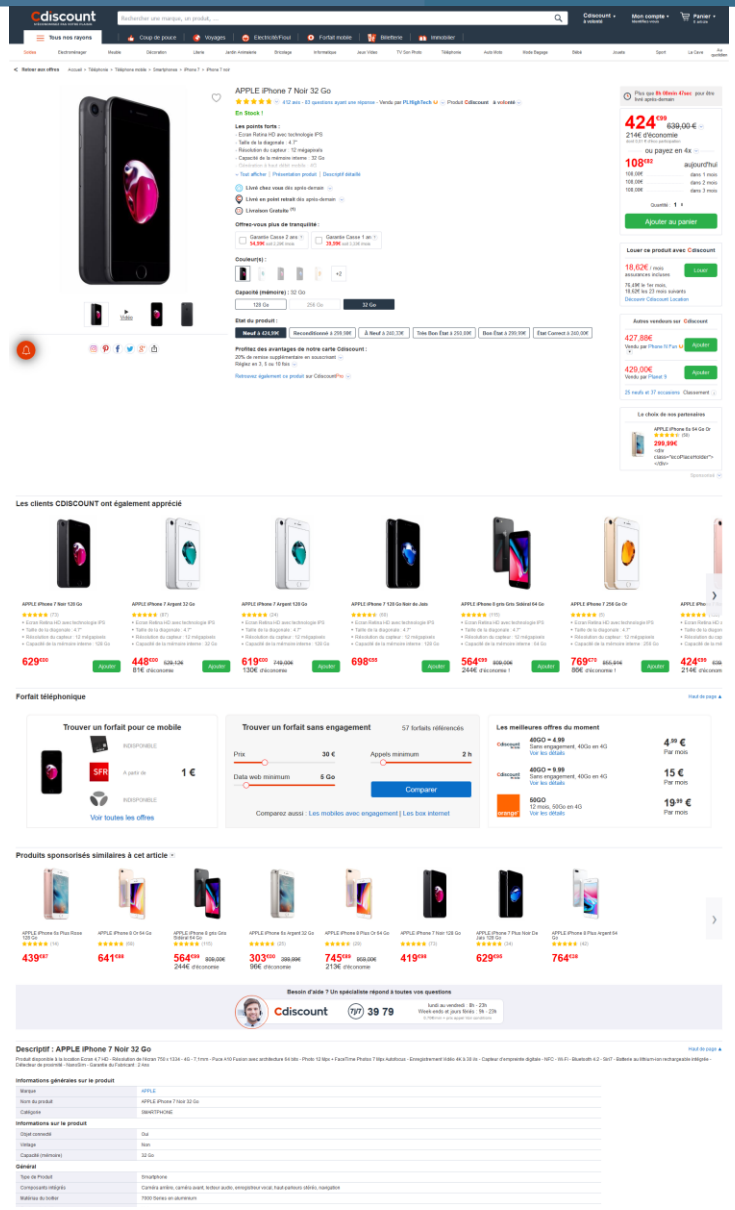
Base d'apprentissage
Sélectionner un échantillon de produits correctement catégorisés représentatif de la diversité du catalogue au sein de chaque catégorie

TF-IDF + k-NN
Transformation du texte brut en vecteur puis attribution d'une catégorie avec un score

Seuils de confiance
Si le score prédit est inférieur à un seuil, dépendant le catégorie, la prédiction est rejetée

Sommaire

Les éléments d’une fiche produit



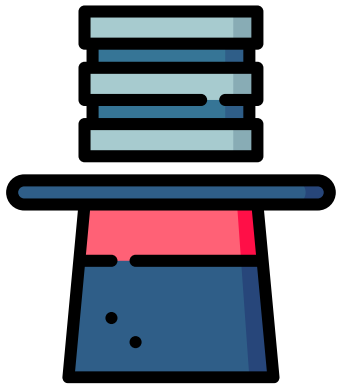
Titre
Texte court caractérisant le produit visible également sur les listes de produit

Image(s)
Visuel représentant le produit, éventuellement mis en situation, visible également sur les listes de produit

Description
Texte naturel long permettant de décrire le produit et ses caractéristiques

Caractéristiques
Texte structuré listant les propriétés techniques du produit (dimensions, poids, taille d’écran, nombre de places…)

Utiliser les images des produits



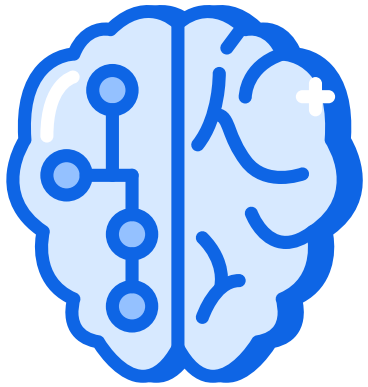
Vectorisation
Transformation de l'image sous forme de vecteur pour pouvoir l'utiliser dans un algorithme d'apprentissage

Aplatissement
Une image peut être vue comme trois matrices de pixel, une par dimension dans l'espace de couleur (RVB)

Points d'intérêt
Résume l'information autour de points d'intérêts (*SIFT*)

Deep learning
Laisser en charge au modèle l'apprentissage du « bon vecteur »

Une histoire rapide du deep learning



Réseau de neurones
Formalisation mathématique d'un ensemble de modèles connectés pour produire une décision finale

1958
Rosenblatt propose le perceptron

Somme pondérée des entrées

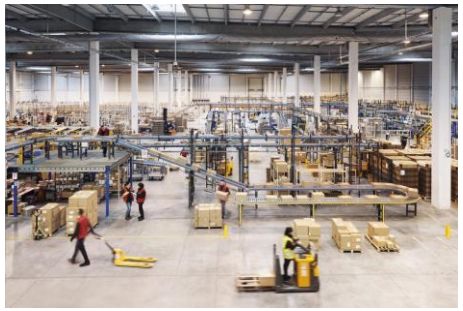
1988
Rumelhart, Le Cun : rétropropagation

Plusieurs sommes pondérées, ensuite repondérées

2012
Krizhevsky utilise un CNN dans la compétition ImageNet

L'architecture du réseau permet de trouver le « bon vecteur »

Une chaine complète



Garantir la précision
L'objectif premier du catégoriseur est sa précision

Utilisation de seuils de confiance pour rejeter les prédictions moins certaines

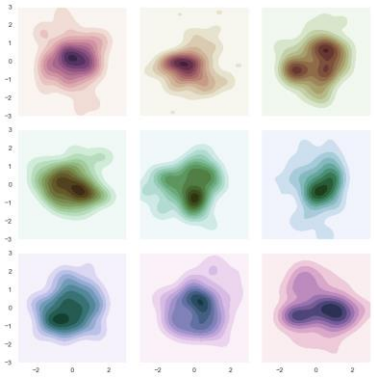
Base d'apprentissage
Sélectionner un échantillon de produits correctement catégorisés représentatif de la diversité du catalogue au sein de chaque catégorie

Architecture DL
Transformation du texte brut en vecteur puis attribution d'une catégorie avec un score

Seuils de confiance
Si le score prédit est inférieur à un seuil, dépendant le catégorie, la prédiction est rejetée

Sommaire

Problèmes de distribution



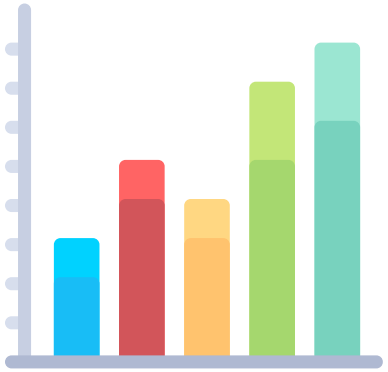
Distribution sous-jacente aux données
Classiquement, les modèles d'apprentissage supposent une distribution fixe dans le temps qui génère les données d'apprentissage et d'entrainement

Apprentissage vs monde réel
Biais dans la sélection des produits formant la base d'apprentissage

Evolution des produits
Les produits vendus sur Cdiscount évoluent et leur description visuelle et textuelle avec eux

Nouveaux produits
Régulièrement, de nouveaux types de produits apparaissent

Evolution des catégories



Equilibre des catégories
Un modèle d'apprentissage ne peut prédire que des catégories qu'il a déjà observées avant et suppose souvent une répartition équitable des exemples connus parmi les catégories possibles

Déséquilibre naturel
Toutes les catégories n'ont pas le même nombre de produits : fendeur de bûches vs décorations de Noël

Modèle textuel
Possible d'entrainer le modèle sur un ordinateur classique en quelques heures

Modèle visuel
Infrastructure spécifique nécessaire et apprentissage (et validation) long

La vraie chaîne complète

L'humain est nécessaire
La gestion du bon fonctionnement au jour le jour d'un tel système de classification nécessite une supervision humaine

Base d'apprentissage
La bonne sélection des produits à utiliser comme référence est la clé des performances

Evaluation du modèle
Une personne experte sur l'arbre des catégories peut évaluer la précision des prédictions et intervenir si besoin

Communication
Un score entre 0 et 1 n'est pas une explication en soit lors des échanges en interne ou avec nos vendeurs

Merci !

Aux organisateurs
Pour cette journée d'échange autour de la data science

A vous
Pour votre écoute lors de cette présentation