

Marina BAIDINA

Master 2 TAL, Université Paris Nanterre

Rapport d'alternance

[Alternance effectuée du 17/09/2020 au 17/09/2021]

CARREFOUR HYPERMARCHES

93 Avenue de Paris, 91300 Massy

**Améliorer l'UX d'un moteur de recherche sur un site e-commerce :
limites et enjeux d'une mise en place d'une solution « clé en main »**

Sous la direction de :

Mme Marion HEMERY

18/07/2021

Université Paris Nanterre

200 Avenue de la République 92001 Nanterre cedex

Année Universitaire 2020 - 2021

Remerciements

Tout d'abord, je tiens à remercier ma tutrice, Marion Hemery, Responsable de Projet Search, de m'avoir confié des missions à la fois intéressantes, stimulantes et responsabilisantes au sein de son équipe. Son aide précieuse et accompagnement attentionné tout au long de mon alternance m'ont permis de m'épanouir professionnellement et personnellement.

J'adresse également mes remerciements à Madame Iris Taravella, Directrice du Master TAL, pour son encouragement et disponibilité au cours de cette année universitaire atypique et particulièrement chargée.

Je souhaiterais remercier les équipes Enrichissement, Merchandising et Searchandizing, ainsi que iSearch pour leur accueil amical et soutien qui m'ont permis de me ressentir un vrai esprit d'équipe.

Remerciements spéciaux à mes amis et ma famille qui étaient toujours là pour m'encourager et m'apporter leur soutien à chaque étape de mon alternance.

Table des matières

Introduction	4
1. Entreprise et son secteur d'activité	6
1.1 Groupe Carrefour	6
1.2 Carrefour France	7
1.3 Carrefour et son site e-commerce	8
2. Le cadre administratif de l'alternance	11
3. Le cadre technique de l'alternance	13
3.1 Les composants du site carrefour.fr	13
3.2 Du fournisseur à l'index	20
3.3 Business Console	23
3.4 Requête et page de résultat	24
3.5 Indicateurs de performance du moteur de recherche	26
4. Contribution à la mise en place d'une solution : Les missions effectuées	28
4.1 Amélioration de l'UX via le suivi de performance des recherches	28
4.2 Nouvelles applications pour un corpus de recherches	29
4.2.1 Extraction de recherches saisonnières	31
4.2.2 Enrichissement automatisé de configurations	34
4.3 Proposition d'évolutions à l'équipe technique	36
4.4 Résolution des bugs : cas de la lemmatisation	39
Discussion : Limites et enjeux	43
Conclusion	45
Références bibliographiques	46

Introduction

L'expérience d'un client faisant ses courses sur un site marchand dépend de beaucoup de facteurs qui sont parfois indépendants : on commence par l'architecture du site, sa rapidité, la facilité de navigation, la taille d'offre, les prix, les promotions, la procédure du paiement, et on termine par la rapidité de la livraison, ou bien, par la disponibilité des créneaux horaires en Drive, par la complétude de la commande, son bon état, la fraîcheur, la pertinence des substitutions...

Un moteur de recherche n'est qu'un seul élément dans cette grande chaîne qui commence au moment où le client ouvre le site avec ou sans intention de faire ses courses et qui se termine quand la commande a été récupérée. Néanmoins, le moteur de recherche joue un rôle primordial dans la facilitation du parcours client, permettant au dernier de non seulement trouver son produit souhaité, mais aussi de découvrir les offres disponibles, former une idée claire et précise de l'assortiment proposé par le magasin.

Idéalement, le moteur de recherche représente une solution très peu dépendante de l'humain, ayant la capacité de s'adapter aux besoins changeants des clients d'une manière automatique. Ce rapport met en lumière les limites et les enjeux que l'on peut rencontrer lors d'une mise en place d'un moteur de recherche interne sur un site de e-commerce en tant qu'une solution « clé en main », donc, une solution customisée, prête à être utilisée directement après la fin de son développement.

Cette problématique sera développée à travers la description de certaines tâches que j'ai pu effectuer en tant que Chargée de projet Search sous la direction de Marion Hemery, Responsable de Projet Search au sein de la direction E-Commerce et Animation Commerciale chez Carrefour, du 17 septembre 2020 à ce jour.

La présentation de mon alternance se déroulera selon le plan suivant :

1. La présentation de l'entreprise et son secteur d'activité : un aperçu sur la grande distribution, le Groupe Carrefour, Carrefour France et le site e-commerce carrefour.fr ;
2. La description du cadre administratif : la présentation des équipes auxquelles j'ai eu l'opportunité d'adhérer lors de mon alternance ;
3. L'explication du cadre technique : l'architecture du site, les outils utilisés, le fonctionnement général du moteur de recherche et les indicateurs de sa performance ;
4. Le développement des missions effectuées, contribuant à la mise en place d'une solution « clé en main » ;
5. La discussion mettant en lumière les défis relevés ;

6. La conclusion.

D'abord, l'entreprise et son secteur d'activité seront présentés.

1. Entreprise et son secteur d'activité

Cette partie vise à présenter l'entreprise qui m'a accueillie pour mon alternance. Mon alternance s'est déroulée dans le secteur de la grande distribution, réunissant les magasins de commerce en détail, dotés d'une grande surface, proposant une vente en format libre-service. Chacun de ces concepts composants qui servent aujourd'hui à définir ce que c'est, un magasin de grande distribution, sont apparus au cours de la première moitié du XXème siècle en France et aux Etats-Unis et ont reçu leur popularité pendant les Trentes Glorieuses, une période marquée par une forte croissance économique (Daumas, 2006).

Actuellement, en France, on compte plusieurs groupes d'enseignes de la grande distribution (Groupe Auchan, Système U, Groupe Casino, Mousquetaires, etc.), dont le Groupe Carrefour.

Dans les parties suivantes, je présenterai d'abord le Groupe Carrefour, puis, Carrefour France et, finalement, cette section sera clôturée par une description de la place que le e-commerce occupe dans le Groupe Carrefour.

1.1 Groupe Carrefour

Le groupe Carrefour est un groupe français, l'un des pionniers de la grande distribution, avec son premier hypermarché ouvert en 1963 à Sainte-Geneviève-des-Bois, en Essonne, proposant à ses clients un vaste choix de produits alimentaires et non-alimentaires « sous un même toit » (Paturle, 2005). Selon Dominique Bessire, professeur en sciences de gestion à l'Université d'Orléans, l'objectif principal de Carrefour à l'époque était de vendre moins cher que les commerces traditionnels, ce qui a été atteint en réduisant le nombre de personnel, massification des commandes, etc.

Aujourd'hui, en dehors de la France, le groupe Carrefour est directement présent dans 8 pays : Brésil, Belgique, Pologne, Italie, Taïwan, Espagne, Roumanie, Argentine, ainsi que dans 20 autres pays en tant qu'une franchise ou un partenariat. Malgré cette diversité géographique d'implantations, le groupe tâche de favoriser le développement économique local, en priorisant les producteurs locaux : selon carrefour.com, 73% de produits vendus dans les magasins du groupe sont produits localement.

Cela s'inscrit dans le programme *Act for Food*, un programme mondial dont l'objectif est d'améliorer la qualité des produits sans augmenter les prix (Figures 1 et 2).



Figure 1 (à gauche) : Actions Act for Food



Figure 2 : Logo Act For Food

En tout, le groupe compte plus de 321 000 collaborateurs dans le monde (données officielles du site carrefour.com). Son chiffre d'affaires pour l'année 2020, marquée par la meilleure croissance depuis 20 ans, s'approche de 79 milliards d'euros (selon les données du magazine LSA), dont presque la moitié a été obtenue dans les pays autres que la France. Depuis le 18 juillet 2017, le poste de Président-Directeur Général est occupé par Alexandre Bompard.

1.2 Carrefour France

En France, le groupe Carrefour est présent sous formats physiques suivants :

- hypermarchés (Carrefour) : la superficie entre 3 000 m² et 27 000 m² ;
- supermarchés (Carrefour Market) : la superficie entre 650 m² et 4 900 m² ;
- proximité (Carrefour Express, Carrefour Montagne, Carrefour Contact, Carrefour Bio, Proxi, etc.) : la superficie entre 50 m² et 900 m².

Tous formats confondus, Carrefour possède à ce jour 5 278 magasins en France.

Depuis le 1er juillet 2020, le poste de Directeur Exécutif France est occupé par Rami Baitieh. Sa politique consiste à prioriser davantage le client en appliquant la méthode 555 : 15 actions pour améliorer l'expérience client (Figure 3).

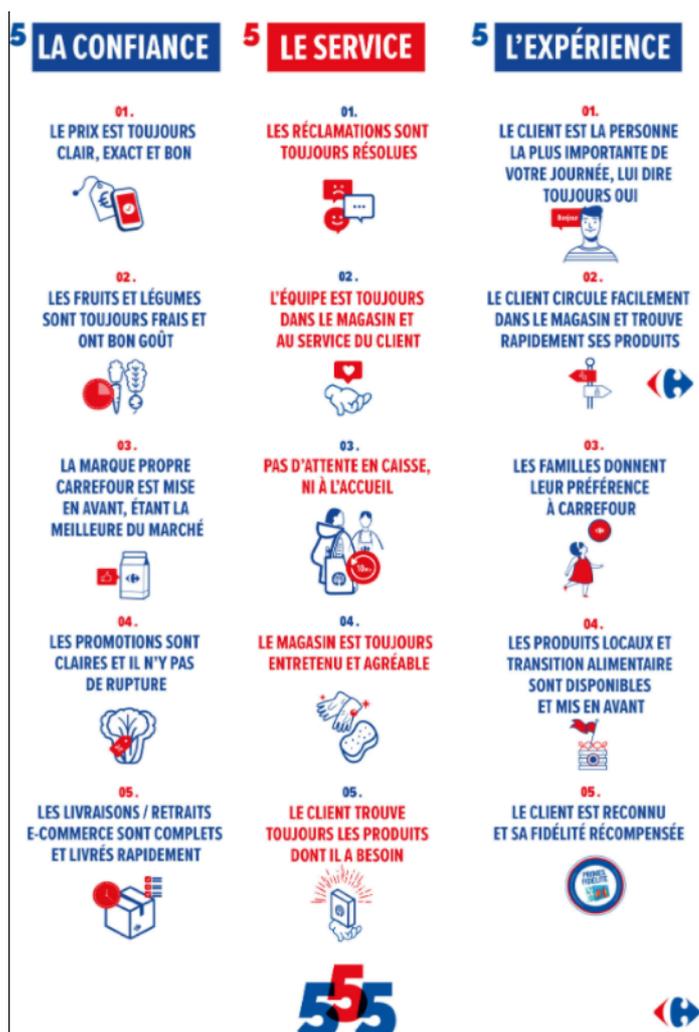


Figure 3 : “555”

Il est important de remarquer que ces règles s’appliquent non seulement aux magasins physiques, mais aussi aux magasins virtuels.

A partir de 2018, Carrefour s’est engagé à investir 2,8 milliards d’euros, afin de créer un écosystème complet réunissant les magasins physiques et le site e-commerce, permettant aux clients de passer leur commande via le site Internet dédié et puis, soit retirer leurs courses au Drive, en Click&Collect, ou bien, être livré au domicile.

La prochaine section décrit en détail le rôle que l’e-commerce joue dans le groupe Carrefour.

1.3 Carrefour et son site e-commerce

Comme mentionné précédemment, le groupe Carrefour investit dans le développement du digital, notamment, du e-commerce. Actuellement, l’e-commerce est géré par Elodie Perthuisot, Directrice Exécutive E-commerce, Data et Transformation Digitale.

Carrefour a ouvert son premier cybermarché en 2000, nommé Ooshop (Figure 4), proposant une livraison des courses à domicile.

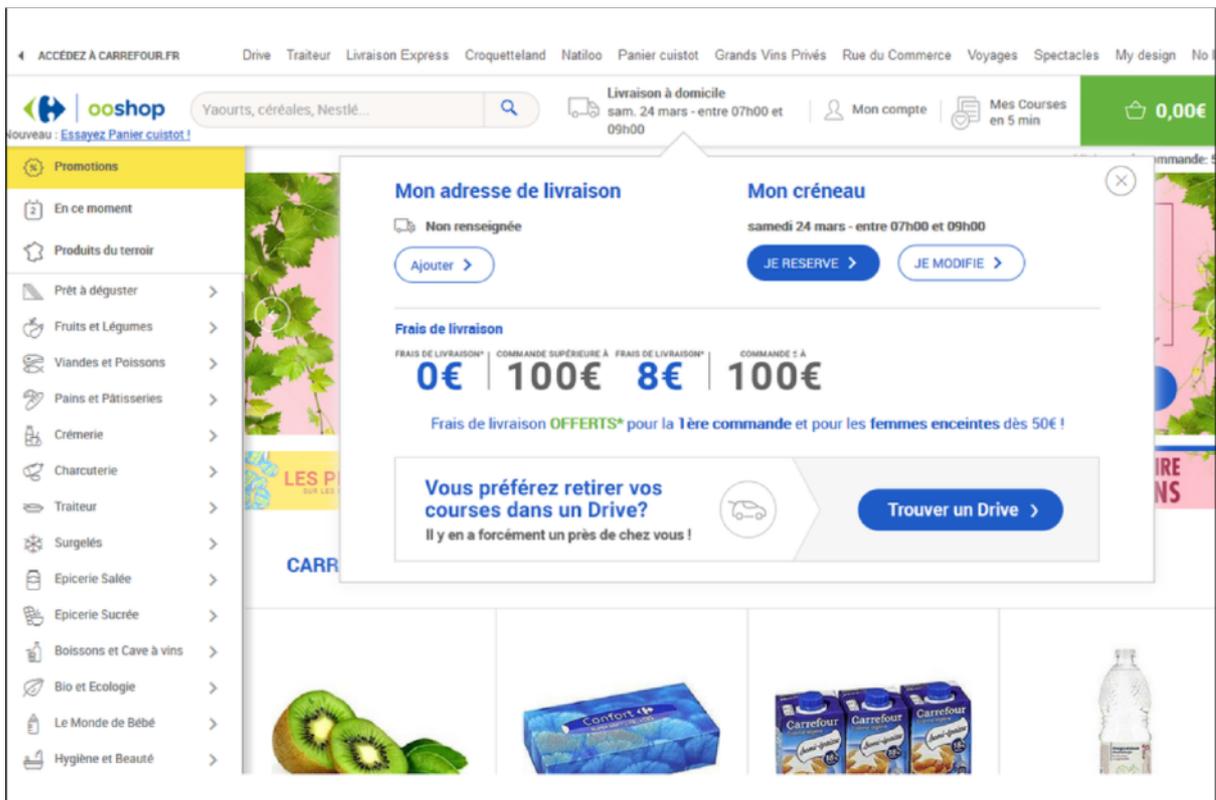


Figure 4 : page home de ooshop.fr

En 2018, Carrefour ouvre son site e-commerce unique : carrefour.fr (Figure 5), réunissant plusieurs bannières : formats de magasins qui diffèrent selon la manière dont les produits sont stockés (par exemple, en entrepôt ou en magasin), selon la manière dont le client peut récupérer sa commande (par exemple, un point Drive ou une livraison).

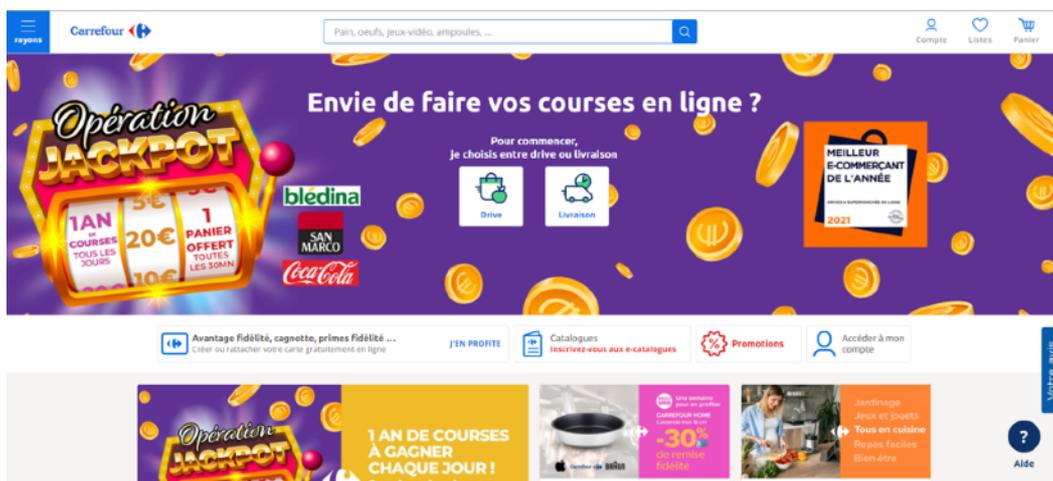


Figure 5 : page home de carrefour.fr

Les bannières actuellement disponibles sont :

- CLCV : Carrefour Livré Chez Vous, les commandes sont préparées en entrepôt et elles sont livrées directement chez le client ;
- Hypermarchés adossés, supermarchés adossés, magasins de proximité adossés : préparation de la commande dans un magasin, récupération de la commande dans un point Drive, via Ckick&Collect ou une livraison à domicile ;
- LEX : livraison express (30 minutes) ;
- PPC : Plateforme de Préparation de Commande : entrepôts ou magasins à partir desquelles une commande est préparée.

Aujourd'hui, en France, du point de vue de la fréquentation, le site carrefour.fr est le site de distribution alimentaire le plus visité, ainsi que le cinquième site e-commerce venant après les sites de Vinted, FNAC, CDiscount et Amazon, avec 1 607 000 visiteurs uniques par jour. En ce qui concerne le chiffre d'affaires, Carrefour Drive occupe actuellement la deuxième place en France, derrière E.Leclerc Drive.

Nous pouvons, donc, remarquer, que le site e-commerce carrefour.fr occupe un rôle stratégique dans le développement digital du groupe, d'où est la nécessité de faire évoluer plusieurs de ses composants.

Dans la prochaine partie, je présente le cadre administratif de mon alternance.

2. Le cadre administratif de l'alternance

Du point de vue organisationnel, mon alternance se déroule au sein de la direction E-Commerce et Animation Commerciale, au sein du pôle Enrichissement, Merchandising et Searchandising, dirigé par David Devalois. Comme son nom l'indique, le pôle est constitué de plusieurs équipes :

- E-merchandising (responsable Mirian Caballero) : analyse des performances des PLP (Product Listing Pages), par exemple, des pages de rayons ;
- Search (responsable Marion Hemery) : optimisation des performances des SRP (Search Result Pages), pages générées par le moteur de recherche ;
- Enrichissement (responsable Ophélie Debert) : contrôle de la qualité et d'authenticité des informations des fiches produits sur les PDP (Product Description Pages) ;
- CarrefourTraiteur.fr (responsable Virginie Berthelin) : proposition de l'activité du Carrefour Traiteur à travers le site dédié ;
- Sites Externes (Foire aux Vins, Rentrée des Classes, etc.) (responsable Sabine Hafid) : vente des produits à travers des services logistiques dédiés.

Pour mon alternance, j'ai adhéré à l'équipe Search, qui est constituée de deux personnes : Marion Hemery, responsable Search et Marianne Ropus, assistante E-Commerce. Plus précisément, les tâches effectuées par les membres de l'équipe Search ont pour objectif d'optimiser le fonctionnement du moteur de recherche via des configurations effectuées avec un outil dédié et indiquer des raisons d'une mauvaise performance, afin d'améliorer la satisfaction des clients. En outre, l'équipe propose des évolutions techniques du moteur de recherche qui sont ensuite mises en place par l'équipe technique.

De l'autre côté, le développement technique du moteur de recherche est pris en charge par l'équipe iSearch, avec son PO Mathieu Manzano. Le Product Owner est un chef de projet dans le cadre de la méthode Agile, dont les responsabilités consistent à identifier et formuler les besoins du métier et à définir et concevoir le produit (en définissant le Backlog, liste des fonctionnalités à développer par l'équipe technique). En plus d'un PO, l'équipe technique iSearch est constituée d'un ingénieur QA (Quality Assurance) et 6 développeurs.

Le fonctionnement de l'équipe technique est segmenté en sprints, des périodes d'environ une semaine. Chaque sprint commence et termine par une réunion.

Chaque journée de travail commence par une courte réunion d'équipe de 15 minutes, appelée le "stand-up meeting" (car, en conditions normales, les participants se tiennent debout tout au long de la réunion), lors de laquelle les membres de l'équipe s'échangent autour du travail effectué le jour précédent, ainsi que des projets pour la journée à venir.

Les actions effectuées par l'équipe technique ont pour l'objectif de faire évoluer les performances techniques du moteur de recherche et portent sur les composants suivants :

- A2
- Batch Admin
- Business Console
- Router

Le contenu de chaque composante sera développé dans la partie « Le cadre technique de l'alternance ».

3. Le cadre technique de l'alternance

3.1 Les composants du site carrefour.fr

Avant d'entamer la description du fonctionnement technique du moteur de recherche, il est nécessaire de décrire l'architecture et les composants du front-office¹ (FO) carrefour.fr.

Le FO du site e-commerce de Carrefour dispose de plusieurs types de pages, permettant de non seulement afficher le produit aux clients, mais aussi de proposer le contenu textuel comme des articles. Voici quelques types de pages du site carrefour.fr.

1. Product Listing Page (PLP)

Les PLPs sont les pages qui proposent une liste de produits, où pour chaque produit sont disponibles les informations suivantes : libellé indiquant la nature du produit (pain au maïs) et sa marque (CARREFOUR) ; le format (le pain de 300 g) ; le prix (à l'unité et au kilogramme) ; le visuel. Le cas échéant, le produit peut être doté d'une vignette, indiquant s'il s'agit d'un produit bio, frais, surgelé, sans gluten, etc.

Carrefour.fr propose les PLP en formats *pages de rayons*, *SRP* (page du résultat du moteur de recherche, Search Result Page), *pages promo* et *pages de boutiques partenaires*. Ensuite, les features de chacune de ces pages seront décrits.

a) SRP

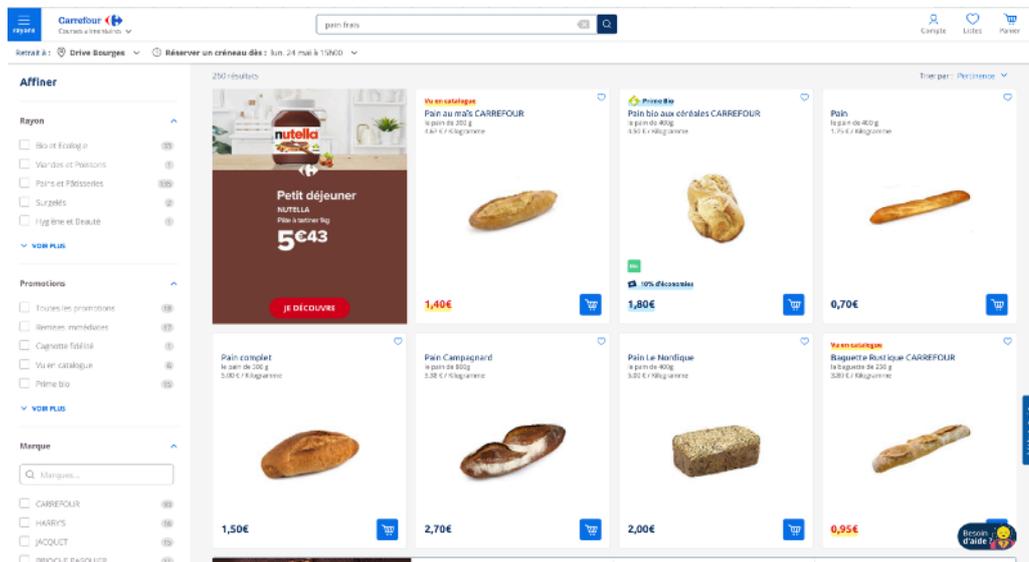


Figure 6 : Page Search de la recherche "pain frais"

¹ partie du site visible par les internautes (source)

Les SRP (Figure 6) sont générées par le moteur de recherche à partir d'une requête qu'un client peut faire entrer dans la fenêtre dédiée (Figure 7).

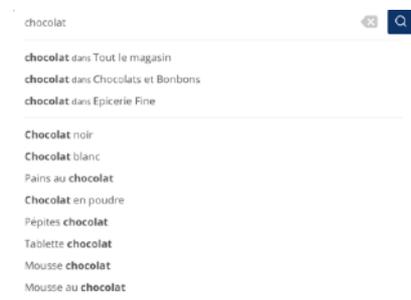


Figure 7 : Fenêtre Search avec un menu d'auto-complétion

La figure 7 illustre la manière dont marche l'auto-complétion de la fenêtre Search du site : les clients sont d'abord proposés de limiter leur recherche à un seul rayon (par exemple, ne rechercher le chocolat que dans le rayon Chocolats et Bonbons, ou bien dans l'Épicerie Fine), puis, ils peuvent choisir entre plusieurs options de complétions (chocolat noir, chocolat blanc, pains au chocolat). Les complétions sont extraites d'un corpus de recherches, le choix et l'ordre de recherches utilisées pour une complétion prend en compte plusieurs facteurs, par exemple, le nombre de recherches et d'achats effectués, etc.

La page résultante contient certaines fonctionnalités qui ont pour objectif d'optimiser l'UX des clients (Figure 8) :

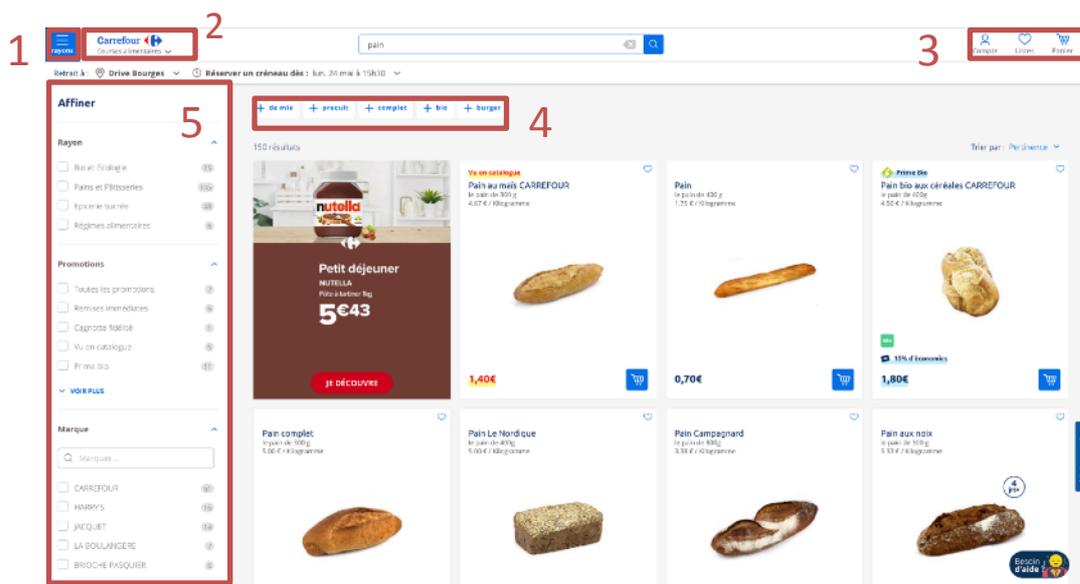


Figure 8 : Page Search “pain” et ses features

Les fonctionnalités indiquées sur la Figure 11 sont :

1. Menu de navigation : permet de visualiser le menu de rayons ;
2. Retour vers la page Home ;
3. Visualisation du compte, liste de courses, panier ;
4. Recherches affinées : une fonctionnalité disponible pour certaines recherches qui permet d'ajouter des mots à la recherche initiale, donc, l'affiner ; par exemple, le client peut choisir d'ajouter « de mie » à la recherche « pain » et ainsi être renvoyé vers la page de la recherche « pain de mie ». Le contenu des recherches affinées vient d'une donnée issue d'utilisation du moteur de recherche : elle représente les séquences de mots-clés les plus recherchés par les clients, en occurrence, il est évident qu'un nombre important de clients cherchent « pain de mie » directement après la recherche « pain ». Ensuite, l'équipe Data extrait ces séquences et les transforme en recherches affinées. En ce qui concerne l'UX, l'ajout des recherches affinées permet d'optimiser le parcours client, en s'adaptant au mieux à leurs besoins et habitudes.
5. Filtres : les filtres peuvent être utilisés pour affiner le choix, et donc, de filtrer l'offre proposée sur la page. Par exemple, un client peut choisir d'afficher des produits d'une certaine marque, ayant certaines propriétés (sans allergènes, de couleur rouge, ayant un format particulier, etc.). Selon les données issues des observations de l'utilisation des filtres, ils trouvent notamment leur application sur les pages des recherches non-alimentaires, permettant de personnaliser le choix, en précisant les critères comme la taille d'écran, la couleur, la marque, etc.

Le ranking de produits de la page search est un facteur très important, contribuant à la satisfaction des clients. Il est effectué en deux étapes : d'abord, par un algorithme prenant en compte le calcul TF IDF, puis, par le Métier via des configurations manuelles. Le calcul du score sera décrit en détails prochainement.

En ce qui concerne l'adresse URL des pages search, elle contient le texte de la recherche après un signe d'interrogation, indiquant une requête :

<https://www.carrefour.fr/s?q=pain+frais>

b) Rayon

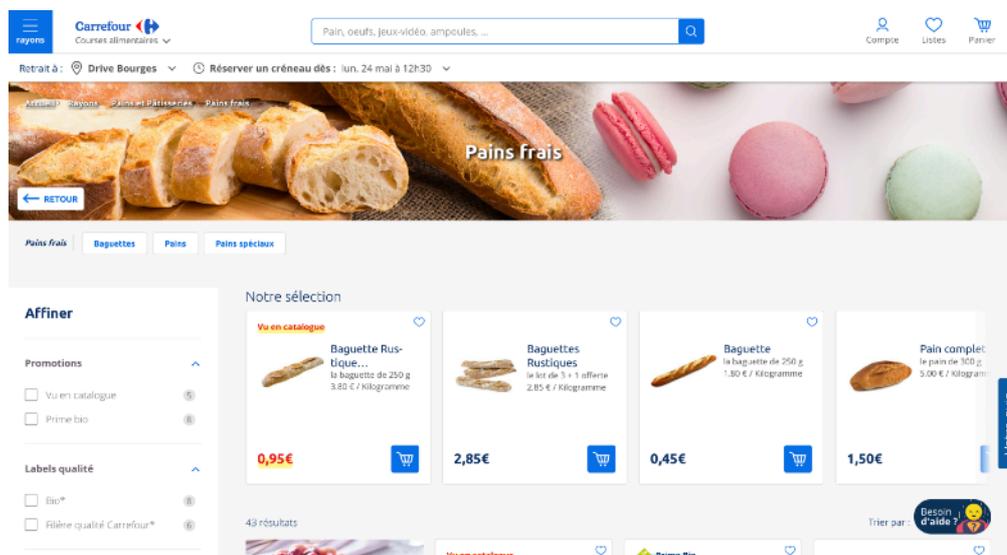


Figure 9 : Page du rayon Pains Frais

Les pages de rayons (Figure 9) permettent de recréer une expérience proche du magasin physique : le client a la possibilité de faire ces courses à partir de pages de rayons uniquement, sans avoir recours à la fonction de recherche. Selon les chiffres issus de l'utilisation du site, environ 50% de clients ont recours aux pages de rayons, afin de faire leurs courses.

En ce qui concerne les fonctionnalités disponibles depuis une page de rayon, elles reprennent, en grande partie, les fonctionnalités disponibles sur une page search : retour vers la page Home ; menu déroulant "Rayons", proposant d'autres rayons ; visualisation du contenu du panier, compte client, liste de courses ; accès au moteur de recherche.

Néanmoins, il existe des fonctionnalités qui sont propres aux rayons. Elles sont illustrées dans la Figure 10.

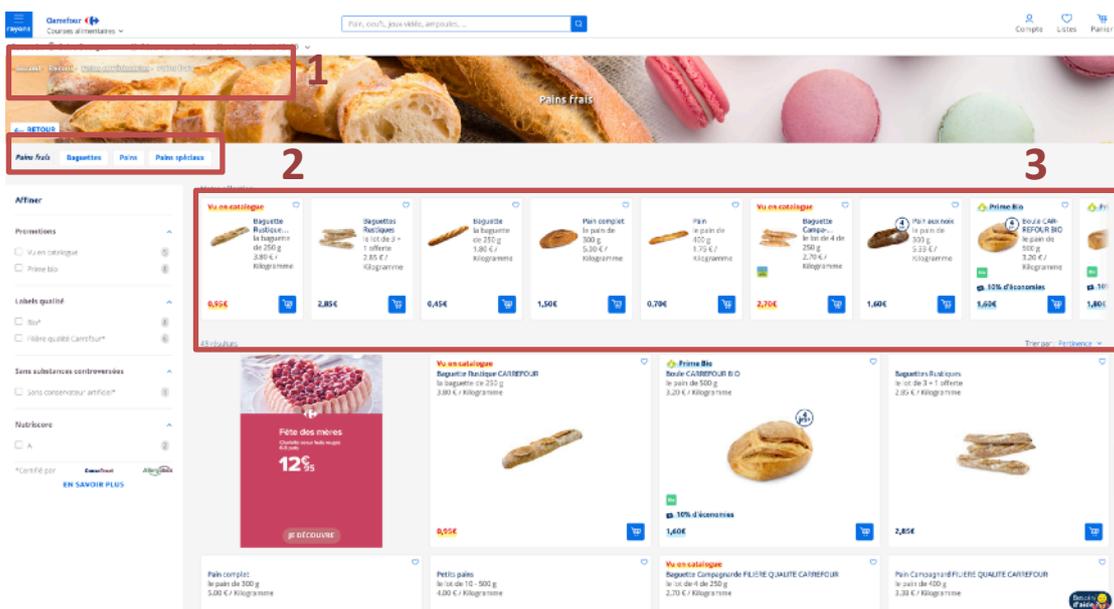


Figure 10 : Page du rayon Pains Frais et ses features

Les composants encadrés en rouge sont :

1. File d'ariane : permet de revenir en arrière (par exemple, le sous-rayon Pains Frais fait partie du rayon Pains et Pâtisserie), donc, pour le client, il suffit de cliquer dessus pour visualiser le rayon parent ;
2. La TagLine : permet d'accéder aux rayons-sœurs (en occurrence, Baguettes, Pains, Pains Spéciaux) ;
3. Carrousel "Notre Sélection" : une sélection de produits mise en avant par l'équipe merch, ou par un calcul, basé sur les promotions, les top références, etc.

La disposition de produits sur les pages rayons est effectuée par défaut par un calcul qui prend en compte plusieurs critères, par exemple, la masse de marge, les ventes en volume, la saisonnalité, la marque, la nature de produit, top UB². Autrement, l'équipe Merch met en place des stratégies spécifiques appliquées sur un ou plusieurs rayons prenant en compte d'autres facteurs, par exemple, les promos, la taille (sur le rayons Couches), etc.

L'URL d'une page rayon reprend l'arborescence, en contenant non-seulement le rayon sur lequel on se situe, mais aussi les rayons-parents :

<https://www.carrefour.fr/r/pains-et-patisseries/pains-frais>

² Unité de Besoin — unité de produits unique correspondant au besoin (REF)

c) Pages de promotions

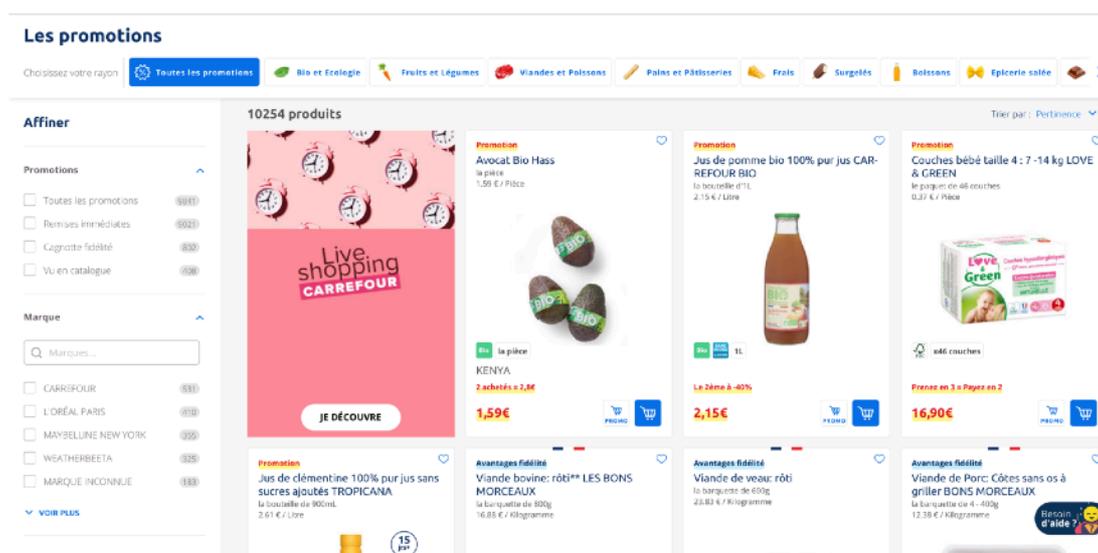


Figure 11 : Page de promotions

Les pages de promotions (Figure 11) sont créées pour mettre en lumière les promotions disponibles sur le site.

Dans les URLs des pages promotions, le nom du domaine principal est suivi par le répertoire : “promotions” :

<https://www.carrefour.fr/promotions>

d) Boutiques de partenaires

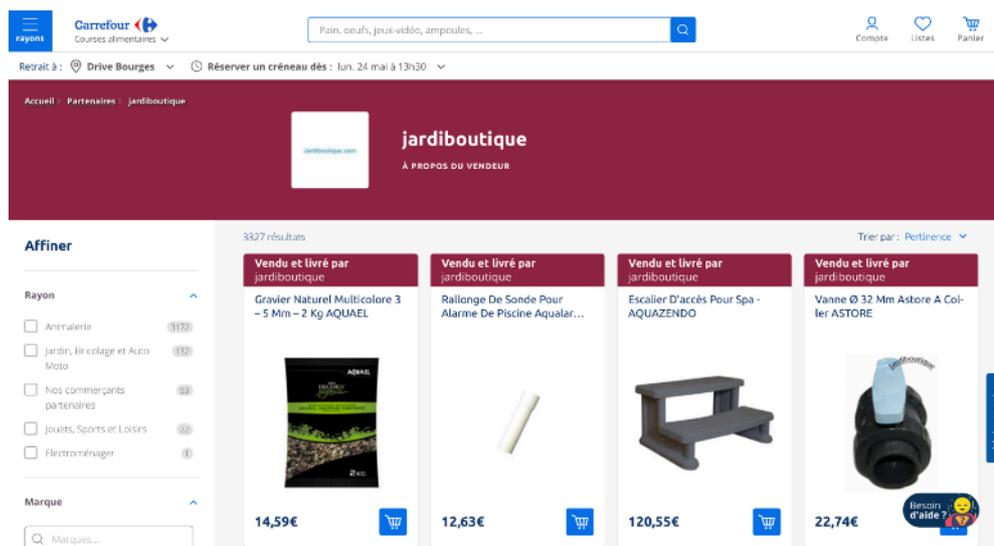


Figure 12 : Page de partenaire Jardiboutique

Les pages de partenaires (Figure 12) contiennent l'offre complémentaire proposée par les vendeurs partenaires dans le cadre de l'action Marketplace. Une page de boutique permet au client d'ajouter au panier des produits venus et livrés par un vendeur indépendant.

Les URLs de ce genre de pages indiquent, entre autres, l'ID et le nom du vendeur :

<https://www.carrefour.fr/v/2318?sq=jardiboutique&noRedirect>

En somme, les PLP proposent une vision sur plusieurs produits et leur listing est généré par le moteur de recherche.

2. Product Detail Page (PDP)

L'objectif d'une PDP est de fournir aux clients les informations nécessaires sur le produit en question, par exemple, son poids, son contenu, sa description, son nutriscore, sa saisonnalité, les allergènes, les recommandations, ainsi que d'autres informations qui peuvent être utiles. Le client peut accéder à la PDP d'un produit en lui cliquant dessus dans la PLP, ou bien, à partir du moteur de recherche externe.

La figure 13 présente un exemple d'une PDP.

← RETOUR Accueil > Rayons > Bio et Ecologie > Epicerie sucrée > Galettes de riz et Pains > Pain complet 3 céréales bio BJORG

Pain complet 3 céréales bio BJORG
le paquet de 500 g
1,89€
3,78 € / Kilogramme

ACHETER

A **B** **C** **D** Nutriscore **400** Score Innit **10**

VOIR TOUS LES PRODUITS DU RAYON

Pain complet à base de seigle, avoine et orge

Bio **V**

Description

Découvrez le pain complet 3 céréales Bjorg et faites le plein d'énergie. Bio, riche en fibres et source de magnésium et de phosphore, ce pain préparé à base de seigle, d'avoine et d'orge est un véritable allié pour concilier bienfaits nutritionnels et plaisir.

Nos clients ont également acheté

Fromage blanc nature 0% MG Bio VRAI le pot de 500 g 3,90 € / Kilogramme 1,95€	Terrine forestière bio BJORG la boîte de 125 g 18,48 € / Kilogramme 2,31€	Tartines crousties sarrasin CEREAL BIO la boîte de 145 g 12,41 € / Kilogramme 1,00€
--	--	--

Figure 13 : PDP Pain complet 3 céréales bio BJORG

L'adresse URL d'une PDP contient le répertoire /p (produit), suivi d'un nom de produit et de son code EAN³, par exemple :

<https://www.carrefour.fr/p/pain-complet-3-cereales-bio-bjorg-3229820019192>

3. Pages Editoriales

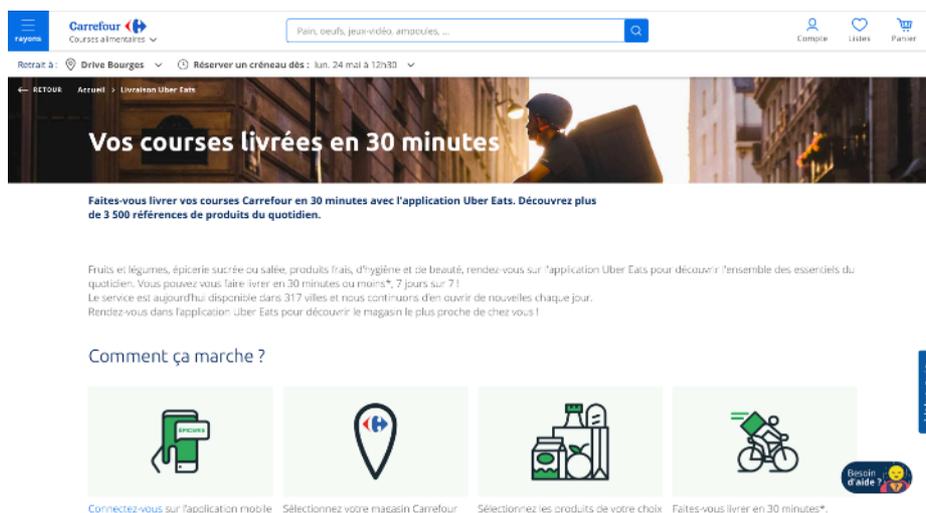


Figure 14 : Page éditoriale « Vos sources livrées en 30 minutes »

Les pages éditoriales (Figure 14) ont pour objectif d'informer le client de services proposés par Carrefour, par exemple, la livraison express, retrait 2h, etc. Normalement, ces pages ne contiennent que le contenu textuel.

L'URL d'une page éditoriale indique également son type :

<https://www.carrefour.fr/edito/livraison-uber-eats>

Toutes les types de pages représentent un système uni, permettant aux clients de faire leurs courses, être au courant des actualités et opérations commerciales, tout en respectant leurs habitudes.

Dans la prochaine partie, le fonctionnement technique du moteur de recherche, ainsi que des composants associés, sera mis en lumière.

3.2 Du fournisseur à l'index

Les données sur les produits sont récupérées par Carrefour via l'outil **Alkemics**, un logiciel collaboratif, permettant aux fournisseurs de remplir les données sur leurs produits. Ensuite, l'équipe Enrichissement effectue la mise en forme de ces informations et les

³ European Article Numbering

dépose dans une base de données via l'outil **DDM** (IBM) qui contient les informations structurées sur les produits proposés par Carrefour qui alimente une base de données nommée **Product**.

Product contient également les informations sur la disponibilité des produits. Vu la nature de données y contenues, cette base de données est constamment modifiée : non seulement les nouveaux produits y sont ajoutés tous les jours, mais aussi les informations sur les produits existants ne sont pas à l'abri du changement, par exemple, les produits ne sont pas toujours disponibles dans tous les magasins, les prix sont régulièrement modifiés et mises à jour, etc.

Ce fonctionnement est primordial, car du point de vue du client du magasin, il est nécessaire d'être au courant de la disponibilité d'un produit. Autrement dit, le client doit être sûr de pouvoir retrouver le produit sélectionné et ajouté au panier au moment de la livraison de sa commande. Afin de permettre le flux continu, le logiciel **Apache Kafka** a été mis en place, permettant de collecter et transmettre en temps réel les changements, ou, dans la terminologie de Kafka, les événements, qui ont lieu au sein de Product au moteur de recherche.

Ces changements sont renvoyés en temps réel au **Batch Admin** du moteur de recherche, une technologie qui permet de transformer de l'événementiel en batch pour effectuer des indexation et des échanges entre deux systèmes de larges volumes de données d'une manière automatique et autonome (sans participation d'un humain). En occurrence, il s'agit du processus **Extract, Transform, Load** (ETL) : les données sont extraites d'une base de données de départ (data source), elles sont ensuite transformées pour pouvoir être exploitées par un outil précis, avant d'être injectées dans la base de données d'arrivée (data target) (Cogoluegnes et al, 2011).

Dans le cas de données sur les produits, la source de données est le Product que Kafka utilise pour communiquer les données en temps réel au Batch Admin qui, ensuite, transforme ces données en format JSON pour les passer au moteur de recherche, l'ElasticSearch.

L'ElasticSearch (ES) est un moteur de recherche open source, permettant de stocker, interroger et analyser en temps réel des données de plusieurs types (données textuelles structurées et non structurées, données numériques, etc). En outre, l'ES dispose d'une architecture distribuée : son index peut être subdivisé en plusieurs **shards**, unités de segmentation d'un index en parties indépendantes. Un ensemble de shards constitue un **nœud (node)** ayant un ou plusieurs rôles, par exemple, un "data node" est responsable de garder les données et participer à l'indexation, alors qu'un "master node" permet de gérer les activités du **cluster** en sauvegardant son état et ses métadonnées. Un cluster, à

son tour, est une collection de nœuds connectés. Finalement, un cluster contient des données sur les catalogues de plusieurs magasins, les configurations mises en place, redirections, rankings, suggestions, etc. Tous les catalogues de Carrefour France sont situés sur cinq clusters.

En ce qui concerne la recherche, elle est effectuée avec un moteur de recherche **Lucene**, un logiciel open source Apache qui effectue une recherche plein texte, c'est-à-dire, qu'il cherche un ou plusieurs termes dans une série de **documents** texte. Le document, à son tour, est un document structuré en plusieurs **champs**, par exemple, le nom du produit, son packaging, le nom de la marque, etc. Chaque champ contient un ou plusieurs **termes**, constituant son contenu.

D'une manière générale, Lucene permet non seulement d'effectuer une recherche, mais aussi d'organiser la disposition de résultats sur la page search selon le scoring. La partie suivante développe et explique le mécanisme de fonctionnement de ces deux fonctionnalités.

D'abord, pour effectuer une recherche, Lucene stocke les informations sur les statistiques de chaque terme employé dans les champs indiqués comme « searchable », c'est-à-dire, utilisés par la recherche, dans un **index inversé** : une architecture de données faisant une correspondance entre le contenu et un ensemble de documents dans lesquels il peut être trouvé. Par exemple, l'index des documents "fromage blanc nature" et "fromage blanc vanille" donnerait l'index suivant :

fromage : 1,2

blanc : 1,2

nature : 1

vanille : 2

Avant d'effectuer l'indexation, les termes subissent quelques étapes de traitement. Par exemple, le terme "Faisselle coulis mûre" sera modifié selon de la manière suivante :

1. tokenization => "Faisselle", "coulis", "mûre" ;
2. conversion en minuscule => "faisselle", "coulis", "mûre" ;
3. re-encodage en ASCII (ASCII folding) => "faisselle", "coulis", "mure".

Selon le terme et le besoin du Métier, une autre couche de traitement peut être ajoutée, plus précisément, nous pouvons définir des synonymes pour un mot-clé. Par exemple,

pour traiter les cas comme “télévision” et “tv”, en rendant la page de résultat de la première recherche identique à la dernière, les deux sont déclarés comme des synonymes.

Les configurations de ce genre sont effectuées manuellement par le Métier avec l’outil Business Console qui présente une interface graphique utilisée pour la génération d’un document JSON de configurations. La partie suivante décrit en détaille les fonctionnalités disponibles avec la Business Console.

3.3 Business Console

Comme mentionné précédemment, la **Business Console** (BC) est un outil utilisé par le Métier, permettant de faire des configurations du moteur de recherche avec une interface graphique. Les configurations ont pour but d’améliorer la performance commerciale du moteur de recherche et consistent souvent à optimiser le ranking, élargir ou réduire l’assortiment proposé, renvoyer vers une page du rayon, créer des facettes, corriger le comportement non-souhaité. En outre, la BC permet d’obtenir les informations sur les produits et de visualiser leur ranking sur la PLP.

Plus précisément, la BC permet une mise en place des configurations suivantes.

- Création de synonymes de deux types, simples et unidirectionnels. Les synonymes simples permettent de lier deux ou plusieurs mot-clés, afin de rendre les pages de résultat uniformes, par exemple, avoir le même résultat pour la recherche “tasse” et “mug”. L’autre type de synonymes, unidirectionnels, donne la possibilité de remplacer la page de résultat d’un mot-clé par une page de résultat d’un ou plusieurs mot clés, par exemple, avoir les pommes de terre des variétés matte, charlotte et amandine comme résultat de la recherche “pomme de terre raclette”.

Les synonymes sont souvent utilisés pour rendre compte des abréviations, oubliés ou ajouts des espaces inter-mots, erreurs d’orthographe, synonymie, variation dans les libellés de produits, etc.

- Modification du comportement du Query Relaxing, un algorithme qui permet une recherche approximative en cas où aucun résultat n’a été trouvé pour une recherche en question. Parfois, l’algorithme de QR n’est pas suffisant pour proposer des options optimales, dans ce cas-là, les synonymes sont utilisés pour proposer d’autres produits (par exemple, en cas d’absence du fromage à tartiner de la marque St Môret, le moteur affichera du fromage à tartiner nature).

- Création de dictionnaires, permettant d’enrichir des données. Notamment, les dictionnaires sont utilisés par l’équipe Merch pour la tâche de ranking, création de facettes. Ce qui est plus important encore, c’est que les dictionnaires permettent

d'uniformiser le contenu des champs de produits, les rendant plus facilement exploitables.

- Création de facettes (affichés comme des filtres dans la Figure 10), ressourcées par le contenu des champs.
- Mise en place de redirections vers les pages rayons ou vers les pages search filtrées.
- Modification des scores avec des boosts de produits, ce qui permet de remonter ou redescendre certains produits, en se basant sur le contenu des champs choisis (par exemple, il est possible de faire remonter les produits du rayon "Jambons blancs" sur la page résultat de la recherche jambon).
- L'indexation qui permet de sauvegarder les modification et de les visualiser sur un magasin sélectionné.

Les configurations sont mises en place sur le site du Drive de la ville de Bourges, ainsi que sur les catalogues de la MKP et du retrait 2h, qui sont uniques pour tous les magasins. Deux fois par semaine, les configurations appliquées aux trois catalogues sont dupliquées vers tous les autres.

L'intégration de la BC est possible grâce au **plugin A2** qui permet de récupérer les fichiers JSON contenant des configurations et les intégrer aux requêtes de l'ES, ainsi renvoyant plus d'informations au FO.

Les détails sur le traitement d'une requête sont décrits dans la prochaine partie.

3.4 Requête et page de résultat

La requête de client est récupérée depuis le FO par le **router** et est traitée ensuite par l'ES.

Plus précisément, elle subit d'abord un nombre de modifications, par exemple, la requête "Lait à la fraise" sera modifiée de la manière suivante :

- tokenization => "Lait", "à", "la", "fraise" ;
- suppression de stop words => "Lait", "fraise";
- lemmatisation ;
- conversation en minuscule => "lait", "fraise" ;
- ascii-folding.

Ensuite, dans le cas d'une requête multi-mots, comme la requête 'lait fraise', l'ES fait une requête par token, donc, il cherche les deux tokens séparément dans l'index, et puis, renvoie le résultat que se trouve au croisement des deux requêtes. Grâce à ce fonctionnement, le moteur de recherche ne renvoie pas des fraises et du lait frais sur la recherche "lait fraise", ce qui améliore son rappel.

En fonction de la requête et des configurations mises en place, les modifications suivantes peuvent également être appliquées :

- redirection vers une page rayon ou page boutique ;
- suggestions : ajout ou suppression d'une ou deux lettres, par exemple, Perier > Perrier ;
- ajout des synonymes : configurations manuelles permettant de mettre en équivalence deux ou plusieurs chaînes de caractères ;
- redirection vers une page filtrée : renvoie vers une page search avec un filtre appliqué, par exemple, la recherche "eau 5L" a été programmé pour renvoyer vers la page search "eau" avec un filtre "format" ayant une valeur de 5L.
- Query Relaxing : recherche approximative en cas de 0 résultat, qui consiste à supprimer de caractères, ne prendre en compte qu'une partie de la requête.

Une fois que les produits sont sélectionnés, le moteur procède au calcul du score qui est définitif dans la tâche de ranking. Le calcul du score prend en compte deux facteurs : la pertinence sémantique et la pertinence commerciale d'un produit/document vis-à-vis la requête. Prochainement, les deux facteurs seront développés.

D'abord, le moteur de recherche calcule un TF-IDF pour obtenir la pertinence que les termes de la requête ont par rapport à chaque document de la sélection. Comme déjà mentionné, les documents sont constitués de plusieurs champs, par exemple, le nom de produit, la marque, le packaging, la description, l'EAN, le rayon, la nature, etc. Afin de mieux correspondre aux objectifs commerciaux, certains champs sont pondérés, c'est-à-dire qu'en cas où un terme se trouve dans le champ "nature", il a plus de poids comparé à s'il se trouvait dans le champ "description".

Ensuite, afin d'améliorer la pertinence commerciale, les scores de certaines catégories de produits sont davantage augmentés ou diminués par des boosts configurés manuellement par l'équipe Search, qui multiplie le score initial du produit par la valeur entrée. Ces boosts permettent de modifier le ranking en remontant ou redescendant certains produits en fonction de leur marque, apparentante à un certain rayon, packaging, opérations commerciales etc. Un exemple d'un boost serait de remonter des glaces de la

marque “Magnum” sur la recherche “glace”, car les données sur les achats effectués sur le site montrent que les glaces de cette marque sont les plus ajoutées au panier par les clients, donc, il est intéressant de les proposer en haut de la page. Ainsi donc, les décisions de la modification du ranking sont prises en résultat d’analyse de performances des pages, d’un suivi hebdomadaire des KPIs⁴.

Les technologies présentées dans cette partie servent à améliorer l’expérience utilisateur de clients du site e-commerce de Carrefour. La partie suivante décrit les indicateurs qui sont utilisés pour le suivi de performance de celui-ci.

3.5 Indicateurs de performance du moteur de recherche

Contrairement aux moteurs de recherche généralistes comme Google, Bing ou Qwant, le moteur de recherche de carrefour.fr a un objectif commercial, qui consiste à augmenter le volume de ventes effectuées sur le site. En conséquence, quand il s’agit d’une évolution de la performance du moteur de recherche, les indicateurs orientés marketing ont plus de poids que d’autres indicateurs comme la précision et la f-mesure. En effet, le comportement des clients peut être imprévisible et il est parfois nécessaire de chercher des raisons autres que celles liées à une mauvaise précision.

Un exemple qui prouve l’importance de la priorisation des indicateurs orientés marketing est le fait que le produit le plus ajouté au panier sur la page de la recherche “lait écrémé” et le lait demi-écrémé. Autrement dit, la page produite par le moteur de recherche aurait beau être bonne en précision, les ventes continueraient à baisser. Donc, l’ajout du lait demi-écrémé avec des configurations manuelles avait permis d’améliorer la performance de la recherche en question, en dépit de la précision du résultat.

Néanmoins, il est important que les pages générées par le moteur montrent un bon rappel, c’est-à-dire, contiennent tous les produits correspondants à la recherche entrée par le client.

En retournant aux indicateurs orientés comportement des clients, dans le cadre d’évaluation de la performance du moteur, nous suivons les données sur :

- Le nombre de recherches → peut être augmenté dû aux opérations commerciales, saisonnalité, mauvais fonctionnement du moteur de recherche, absence du résultat.
- Le taux d’ajouts au panier (nombre d’ajouts au panier uniques/nombre de sessions de recherche⁵) → baisse quand le client ne trouve pas le produit désiré, ou bien, quand il

⁴ Indicateurs clés de performance

⁵ Une session de recherche qui peut contenir plusieurs recherches

utilise le site pour obtenir les informations sur les prix, disponibilité de produit, ou promotions.

- Le taux de sorties (nombre de sorties⁶/nombre de sessions de recherche) → contrairement au taux d'ajout au panier, monte quand le client ne trouve pas le produit désiré.
- NPS (Net Promoter Score) → un indice permettant de mesurer la satisfaction client, basé sur les réponses aux questions du type "Recommanderiez-vous ce service à vos amis/votre entourage ?". Selon la réponse choisie, les clients sont classés comme des détracteurs (note basse), passifs (note moyenne), ou des promoteurs (note élevée). Le calcul du NPS consiste à soustraire le pourcentage des détracteurs du pourcentage des promoteurs. Le score du NPS peut varier entre -100 (100% détracteurs) et 100% (100% promoteurs).
- Verbatims → les commentaires et les avis de clients issus des différentes sources, comme par exemple, les questionnaires laissés sur carrefour.fr ou sur les sites tiers. Normalement, les verbatims permettent d'avoir une image plus complète des irritants clients.
- Le taux de clic (le nombre de fois qu'un élément de la page a été cliqué/le nombre de fois qu'un élément a été affiché) → est utilisé pour voir l'efficacité d'utilisation d'un élément.
- La position du produit le plus ajouté au panier → permet d'évaluer la pertinence du ranking mis en place.

Ensemble avec le rappel, l'ensemble de ces indicateurs sont utilisés pour proposer des configurations et les évolutions pertinents du moteur de recherche.

L'objectif de cette partie était de décrire le cadre technique de mon alternance, notamment, les fonctionnalités du site carrefour.fr, l'architecture du moteur de recherche et composants liés, ainsi que les indicateurs suivis pour évaluer sa performance.

La prochaine partie décrit certaines missions que j'ai eu l'opportunité d'effectuer, afin de contribuer à la mise en place d'une solution pérenne pour le moteur de recherche.

⁶ Le nombre de session ayant terminé depuis la page search

4. Contribution à la mise en place d'une solution : Les missions effectuées

4.1 Amélioration de l'UX via le suivi de performance des recherches

Une partie importante des tâches réalisées au sein de l'équipe Search est liée à l'amélioration de l'UX avec une mise en place des configurations, effectuée avec la BC. Habituellement, le processus peut être divisé en trois parties : découverte du problème, sa résolution et, finalement, le suivi de performance. Dans cette partie, j'illustrerai ce cycle en utilisant un problème concret issu de mon expérience.

Premièrement, le problème est découvert : à la fin de l'année 2020, une baisse consécutive du Chiffre d'Affaires (CA) des Produits Frais Traditionnels (PFT) (par exemple, le fromage à la coupe, des fruits et légumes, etc.) sur le Drive a été signalée. Il a été remarqué que certains produits PFT sont placés trop bas dans la page, donc, moins visibles pour les clients. En conséquence, une repasse a été programmée pour améliorer la visibilité de ce type de produit sur les SRP.

Dans le cas de cette demande de repasse, le fonctionnement non-désiré a été signalé par une personne de l'autre service. Néanmoins, les demandes de ce type peuvent également venir d'autres sources, par exemple, une baisse des KPIs, des bugs inaperçus par l'équipe technique, ou bien, des recommandations données par des fournisseurs.

Ensuite, les configurations mentionnées dans la partie 3.3, visant à résoudre le problème signalé, ont été mises en place. Dans le cas des PFT, un script Python a été élaboré, afin de permettre une écriture automatisée d'un code JSON, afin de remonter les catégories de produits désirés.

Finalement, une fois les configurations mises en place, nous avons effectué un suivi de six semaines, pour voir l'évolution hebdomadaire des KPIs : nombre de recherches, taux d'ajout au panier et taux de sorties. Au cas où une dégradation des performances a été observée, des configurations supplémentaires étaient mises en place.

A la fin du suivi, selon la performance observée, la configuration peut être considérée comme réussite et, donc, ne nécessitant pas d'autres modifications, autrement, nous contions à chercher des améliorations possibles, afin d'atteindre la performance optimale.

Cette approche à une résolution de problèmes est très difficilement automatisable, car il s'agit des modifications ponctuelles, nécessitant une expertise en matière des ventes.

Les parties suivantes décriront d'autres tâches effectuées dans le cadre de mon alternance. Dans le cas des tâches liées à la mise en place des configurations du moteur de recherche, chaque mission a été divisée en trois sous-tâches décrites ci-dessus. Pour des raisons de concision, seulement la deuxième partie sera détaillée. J'adapterai une

approche critique aux tâches réalisées, afin de mettre en lumière les améliorations possibles.

4.2 Nouvelles applications pour un corpus de recherches

Chaque mois, l'équipe Analytics publie un tableur extrait du BigQuery⁷ (BQ) contenant le top 10 000 mot-clés les plus avec recherchés, ainsi que les données sur leur volumétrie. Afin d'améliorer l'exploitabilité de ces données, ainsi que de les utiliser pour certaines actions, il a été décidé de créer un fichier rassemblant toutes les recherches de l'année 2020, ainsi que leurs KPIs mensuelles, jugées pertinentes pour le suivi. Ce fichier devrait permettre d'évoluer rapidement la normalité de comportement d'une recherche, lors de son suivi.

Ainsi, donc, il a été choisi d'organiser le corpus en trois fichiers : chaque fichier contenant une liste de mots-clés recherchés en 2020 et une concaténation de valeurs d'un des indicateurs suivants : le nombre de recherches, le taux d'ajouts au panier par session, le taux de sorties.

N'ayant pas à ma disposition de libraires nécessaires, ainsi que n'ayant pas une possibilité de les installer sur ma machine, j'ai codé une fonctionnalité permettant une concaténation de plusieurs fichiers, son algorithme était le suivant :

⁷ entrepôt de données d'entreprise géré par Google

```

REP ← le contenu du répertoire des BQ
I ← 0
Tant que I < longueur(REP)
|
|   Tester si la variable DICO du type dictionnaire est déclarée en tant qu'une variable locale
|   ou globale
|   Si DICO n'existe pas, alors
|   |   DICO ← dictionnaire
|   |   Extraire le couple TEXT18 et VOL19 du REP[I]
|   |   Extraire le couple TEXT2 et VOL2 du REP[I+1]
|   |   T ← 0
|   |   Pour T allant de 1 au longueur (TEXT1)
|   |   |   Si TEXT1[T] est dans le TEXT2
|   |   |   |   IND ← index du TEXT1[T] dans TEXT2
|   |   |   |   Ajouter TEXT1[T] en tant que clé du DICO
|   |   |   |   et la liste VOL1[T], VOL2[IND] en tant que ses valeurs
|   |   |   Sinon
|   |   |   |   Ajouter TEXT1[T] en tant que clé du DICO
|   |   |   |   et la liste VOL1[T], "NA" en tant que ses valeurs
|   |   |   Fin Si
|   |   Fin Pour
|   |   I ← I+1
|   Sinon
|   |   Extraire le couple TEXT,VOL du REP[I]
|   |   Pour T allant du 1 à la longueur(TEXT)
|   |   |   Si TEXT[T] existe dans la liste de clés du DICO, alors
|   |   |   |   VOL est ajouté à la liste de valeurs du DICO[TEXT[I]]
|   |   |   Sinon
|   |   |   |   DICO[TEXT[T]] ← liste contenant "NA"
|   |   |   |   Si I > 1, alors
|   |   |   |   |   J ← 0
|   |   |   |   |   Tant que J < I-1
|   |   |   |   |   |   Ajouter "NA" au DICO[TEXT[T]]
|   |   |   |   |   |   J ← J+1
|   |   |   |   |   Fin Tant que
|   |   |   |   Ajouter VOL[T] au DICO[TEXT[T]]
|   |   |   Fin Si
|   |   Fin Pour
|   |   Pour T allant du 1 à la longueur(clés du DICO)
|   |   |   Si clé[T] n'est pas dans TEXT
|   |   |   |   Ajouter "NA" aux valeurs du DICO
|   |   |   Fin Si
|   |   Fin pour
|   |   I ← I+1
|   Fin Si
Fin Tant que

```

Pseudo-code de la fonctionnalité "concaténation"

⁸ la variable contient une liste des mots-clés du fichier

⁹ la variable contient une liste des volumétries en question (nombre de recherches, taux d'ajouts panier ou taux de sorties) du fichier

Comme résultat, j'ai pu obtenir des fichiers au format csv, contenant des volumétries concaténées pour chaque recherche du corpus. La figure 15 montre un extrait du tableur obtenu qui associe les mots-clés au nombre de recherches effectuées.

recherche	Dec_2019	Jan_2020	Fev_2020	Mar_2020	Apr_2020	Mai_2020	Juin_2020
lait	99141	126302	105887	539696	261609	156824	109339
jambon	89332	117861	113776	550566	354040	212734	136218
beurre	81715	86290	74620	413744	268529	159669	97827
foie gras	79660	6623	2282	2715	4463	2403	2035
fromage	70461	69352	67602	439829	386344	180174	101098
oeufs	70200	99760	78908	319565	266256	126726	100495
pain de mie	67238	81453	66541	446353	380651	172170	89475
chocolat	65234	39337	36112	226359	250496	126779	64247
lessive	56783	61587	58830	139213	92117	72521	56591
saumon	55292	33143	27211	64883	77242	55208	36275
creme fraiche	54626	64843	60485	193260	161199	104687	32556
papier toilette	54100	67838	76804	571766	243913	156233	102229
eau	53537	78385	68166	375710	228771	151737	115347
yaourt	52333	61334	62385	357539	213131	118632	91407
salade	52172	69891	61887	273576	219263	130380	85927

Figure 15 : extrait du fichier de concaténation

En tout, le fichier contient 19 667 mot-clés et leurs volumétries.

Les informations contenues dans les fichiers sont utilisés par les membres de l'équipe Search, afin de vérifier si les recherches analysées suivent les tendances globales, ou bien, si leurs performances sortent du comportement habituel et nécessitent une inspection méticuleuse.

Le corpus obtenu a été ensuite exploité dans la réalisation de trois tâches suivantes : 1) extraction de recherches saisonnières ; 2) extraction de recherches ayant une distance lexicale minimale ; 3) enrichissement de la classe de synonymes "Espace inter-mots".

Les prochaines parties décrivent les tâches effectuées.

4.2.1 Extraction de recherches saisonnières

Une d'applications possibles d'un fichier qui concatène les KPIs manuelles de chaque recherche effectuée au cours de l'année 2020 est d'obtenir une liste de recherches marquées par une forte variation saisonnière.

Souvent, les recherches peu performantes sont repérées quand le niveau d'insatisfaction de la page générée par le moteur de recherche (produits manquants, aucun résultat, mauvais ranking) est assez élevé. Le but de cette action est d'analyser les SRP de recherches en question et de mettre en place les configurations nécessaires en amont de la saison.

Afin de réaliser cet objectif, j'ai entrepris les étapes du travail suivantes :

1. calcul des écarts par mois pour choisir un indice de performance le plus impacté par les variations saisonnières ;
2. extraction de recherches dont la variation de valeurs est la plus importante ;
3. recherches de mot-clés les plus affectées au cours de la saison printemps-été ;
4. mise en place des configurations et suivi.

Lors de la première étape, un script Python a été employé pour calculer un écart moyen¹⁰ entre chaque valeur mensuelle et la moyenne annuelle. Il est important de noter que les données du mois de mai n'ont pas été prises en compte, car elles étaient fortement impactées par le début du premier confinement 2020, marqué par une pénurie de plusieurs types de produits, donc, les chiffres obtenus ce mois-ci pourraient biaiser les résultats.

Les figures 16 - 18 illustrent une vue réduite sur les écarts obtenus pour la recherche "foie gras" pour les trois types de KPIs.

Dec 2019	Jan 2020	Fev 2020	Apr 2020	Mai 2020	Juin 2020	Juillet 2020	Aout 2020	Sept2 020	Oct 2020	Nov 2020	Dec 2020	Jan 2021
262.72	-69.84	-89.61	-79.68	-89.06	-90.73	-90.52	-93.26	-93.26	-75.66	10.77	548.76	-50.63

Figure 16 : Ecart des nombres de recherches "foie gras"

Dec 2019	Jan 2020	Fev 2020	Apr 2020	Mai 2020	Juin 2020	Juillet 2020	Aout 2020	Sept2 020	Oct 2020	Nov 2020	Dec 2020	Jan 2021
-4.77	-25.42	-19.9	38.43	26.65	12.96	11.92	10.3	5.45	-35.73	-39.05	2.98	-25.81

Figure 17 : Ecart des taux d'ajouts au panier "foie gras"

Dec 2019	Jan 2020	Fev 2020	Apr 2020	Mai 2020	Juin 2020	Juillet 2020	Aout 2020	Sept2 020	Oct 2020	Nov 2020	Dec 2020	Jan 2021
26.12	23.7	-6.02	-35.83	-24.55	-0.88	0.43	-0.22	6.22	11.76	21.49	7.63	2.35

Figure 18 : Ecart des taux de sortie "foie gras"

Vu que les taux d'ajouts au panier et de sortie sont relatifs au nombre de recherches, nous pouvons constater que c'est le dernier indice qui reflète le plus les variations saisonnières : l'écart de la recherche "foie gras" devient négatif pour les mois du janvier au novembre et s'envolent au moins de décembre, car c'est un élément essentiel du repas de Noël.

¹⁰ formule

Conséquemment, ce sont les données sur le nombre de recherches effectuées qui seront utilisées pour la détection de recherches saisonnières.

Ensuite, afin de faciliter le travail et l'interprétation des chiffres, je transforme les nombres de recherches mensuels en indices de saisonnalité, calculé de la manière suivante : la valeur mensuelle séparée par la moyenne annuelle. Le plus l'index s'approche de 1, le plus la valeur mensuelle s'approche de la moyenne annuelle. Ensuite, pour rendre compte de la variabilité, l'écart-type de tous les indices a été calculé.

En se basant sur une analyse brève du tableau, j'ai considéré que l'écart-type supérieur à 0.70 signifie une présence de variations dues au changement de saison. L'extrait du tableau récapitulatif est présenté dans le tableau 19.

recherche	ecart type
foie gras	1.915
chocolat	0.782
champagne	0.966
saumon fume	0.705
switch	0.994
raclette	0.92
toast	1.333
buche	2.184

Tableau 19 : Extrait du tableau récapitulatif

Comme on peut voir dans la figure 19, cette approche n'a pas permis de distinguer les variations importantes de volumétrie liées à la saisonnalité (raclette), des événements (foie gras, champagne, bûche, etc.) et aux réalisations de produits (switch).

La troisième étape consistait à classifier les recherches extraites selon leurs rayons d'appartenance et les saisons marquées par un intérêt important de la part des clients, ainsi qu'à analyser les pages des recherches sélectionnées comme des recherches de la saison printemps-été. La figure 20 présente l'extrait du document Google Sheet résultant.

Recherche	Ecart-type	Rayon	HIVER	PRINT	ETE	AUT
cerise	0.977	Fruits et Légumes			X	
poussette canne	0.72	Bébé			X	
panache	0.763	Boissons		X	X	
gazeuse minerale naturelle san pelle	0.797	Boissons			X	
biere aromatisee tequila desperados	0.906	Boissons			X	
cerises	0.962	Fruits et Légumes			X	
velo femme	0.878	Sports et Loisirs			X	
plancha	0.817	Electroménager		X	X	
insecticide	0.881	Entretien et nettoyage		X	X	

Figure 20 : Fichier saisonnalité

En tout, 90 mot-clés fortement recherchés pendant le printemps et l'été ont été repérés.

Grâce aux filtres sur les saisons, ainsi que sur les événements, les membres de l'équipe peuvent filtrer la vue sur le tableur et facilement obtenir les informations souhaitées. En plus de cela, cette tâche a permis d'anticiper la popularité de plusieurs recherches et mettre en place des configurations nécessaires (synonymes, boosts, redirections vers les rayons) avant le début de la saison.

4.2.2 Enrichissement automatisé de configurations

Une autre application du corpus obtenu est d'enrichir les configurations du moteur de recherche existantes d'une manière automatique, car le corpus de recherches 2020 représente une donnée réelle, issue de l'utilisation du moteur de recherche.

Extraction de recherches ayant une distance lexicale minimale

Comme mentionné précédemment, les recherches donnant aucun résultat sont traitées d'une manière approximative, QueryRelaxing, autrement, le client est renvoyé vers une page de résultat vide (figure 21).

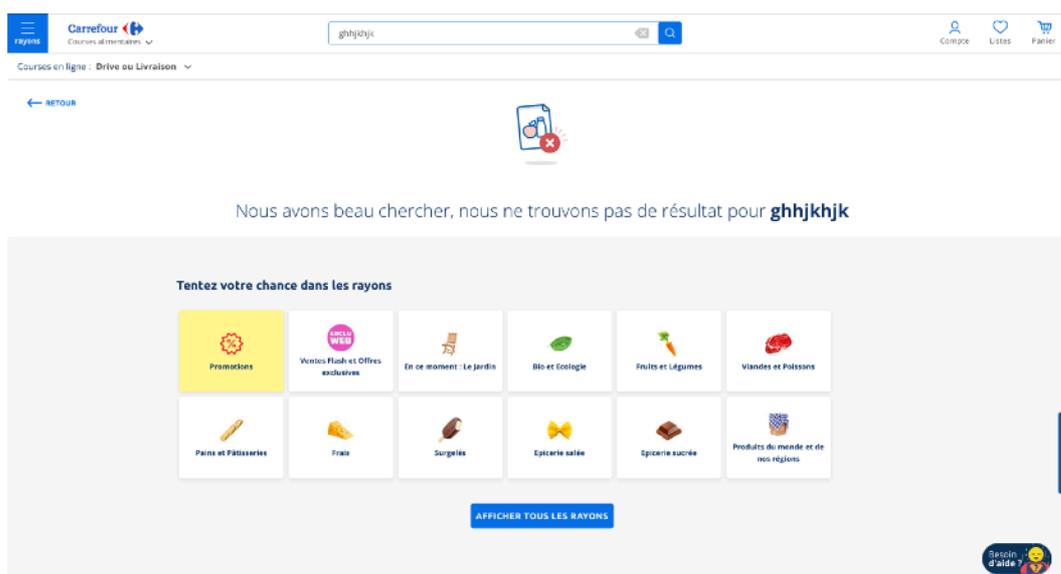


Figure 21 : page 0 résultat

Néanmoins, le QR ne permet pas de rendre compte de toutes les variations orthographiques. Ce comportement est corrigé en créant des synonymes dans le groupe “o résultat”.

Ainsi, donc, ma tâche consistait à trouver un moyen d’extraire du corpus les recherches qui n’ont pas la performance souhaitée à cause d’erreurs d’orthographe. Il a été choisi de se baser sur la distance de Levenshtein, une distance qui mesure le nombre de modifications qu’il faut accomplir, afin de rendre une chaîne de caractères équivalente à l’autre, considérant comme “proches” deux mot-clés ayant une distance inférieure à 3, mais supérieure à 1.

J’ai organisé mon travail en étapes suivantes :

1. groupement de recherches en fonction de leur implantation rayon du niveau 1, ce qui permet de limiter le nombre d’erreurs produites par le calcul de distance (ex. la distance entre “lard” et “card” est égale à 1) ;
2. calcul de Levenshtein par paire de mots-clés au sein de chaque rayon ;
3. enrichissement des données obtenues par les informations sur le taux d’ajout au panier et taux de sorties de chaque recherche ;
4. extraction des recherches ayant le taux d’ajout de panier plus bas et le taux de sortie plus haut que la valeur moyenne du groupe, basé sur le calcul d’écart ;
5. vérification manuelle du résultat.

La dernière étape a été nécessaire, car l’approche basée sur la distance de Levenshtein a néanmoins généré des erreurs, par exemple, les recherches “jambon” et “jambon bio” ont été considérées comme étant proches, alors que du point de vue de l’assortiment et des performances attendues, elles sont assez éloignées. Conséquemment, je peux noter que même si la distance de Levenshtein m’a permis d’automatiser le processus du regroupement, lui seul n’était pas suffisant et une repasse manuelle a été nécessaire.

Ensuite, une repasse a été faite sur 152 recherches, regroupées en fonction d’action à entreprendre : synonymes simples (whiskey == wiskhy == whisky), synonymes unidirectionnels (galette bretonne > galette sarrasin) et d’autres configurations.

Enrichissement de la classe de synonymes Espaces inter-mots

Une autre application du grand corpus a été l’extraction automatique des groupes des recherches ayant comme différence des espaces insérées, ou oubliée au sein du mot-clé, par exemple, “jambon bio” et “jambonbio”.

Les cas de ce genre ne sont pas toujours pris en compte par le QR, donc, il est nécessaire de passer des configurations à la main pour éviter que ces recherches tombent sur un résultat vide.

Mon action a permis d'enrichir la classe de synonymes Espaces Inter-mots, en y injectant plus de 200 synonymes, augmentant la capacité du moteur de donner un résultat pertinent.

Comme il a été remarqué, grâce au fichier du corpus de recherches de l'année 2020, nous avons pu faciliter l'analyse effectuée par l'équipe, en donnant un aperçu facile et compréhensible sur le comportement montré par des recherches au cours de l'année passée, ainsi que d'enrichir les configurations existantes, d'une manière semi-automatique, permettant une injection de plusieurs configurations à la fois.

En ce qui concerne les difficultés rencontrées, d'abord, mon travail a été compliqué par un manque d'outils nécessaires. En outre, j'ai dû confronter l'enjeu consistant à mettre en place une méthode d'automatisation de traitement du corpus qui serait à la fois peu chronophage dans son développement, mais aussi ne pas nécessitant une repasse manuelle.

En ce qui concerne ce dernier aspect, une repasse manuelle s'est avérée nécessaire dans le cas de toutes les tâches effectuées, car les données exploitées, ainsi que les méthodes choisies ne permettaient pas d'avoir la finesse nécessaire.

4.3 Proposition d'évolutions à l'équipe technique

Une de mes missions consistait à communiquer les propositions d'évolution auprès de l'équipe technique. Ces propositions, nommées des expressions de besoin (EB), concernaient plusieurs types de fonctionnalités, ayant pour objectif d'améliorer la performance du moteur de recherche, ainsi que de faciliter le travail du Métier. L'objectif d'une EB est de présenter le problème d'une manière claire et concise, sans pour autant donner des solutions.

Dans cette partie, un exemple d'une EB, ainsi que son apport seront présentés.

La proposition d'évolution décrite dans cette partie a pour l'objectif d'améliorer la réactivité du moteur de recherche, ainsi que de faciliter la création de boosts pour le Métier.

La description du problème, les zones impactées et les KPIs qui devront être suivies pour s'assurer d'un bon fonctionnement de la fonctionnalité sont présentés dans la figure 22.

Overview		
Business owner : Marion Hemery Writer : Marina Baidina	Date : 06/04/21	Priority : ★★☆☆
Business Requirement : Correspondance partielle de contenu du champ à booster en BC		
Summary : - Problème : il est difficile de booster des produits en utilisant les champs dont le contenu est peu homogène. => Actuellement, le boost ne permet que le match exact de la valeur du champ. Cela est pratique lorsqu'il s'agit de boosts de champs bien homogénéisés, par exemple, le noeud product.brand_name . => Néanmoins cette fonctionnalité a des limites, notamment, elle ne permet pas de booster des champs contenant une valeur particulière, ou de ne pas être sensible à la casse (ex. booster le mot "Parme" pour remonter les produits ayant comme qualifier "de Parme AOP", "cru de parme AOP", etc). => Cela impacte la création de boosts des champs product.qualifier , product.packaging et product.nature , dont les configurations doivent être multipliées,, ce qui entraîne des problèmes de performance (ralentissement du chargement de la page) et rendent le processus de configuration chronophage (un boost == une variation de valeur).		
Impact Pages : - Search Area		
KPI : - Ranking - NPS - Taux de Sortie (amélioré si meilleures performances > réduction du volume de configurations)		

Figure 22 : Sommaire EB

Le problème décrit dans cette EB est le suivant : en tant qu'utilisateurs de la BC, nous sommes souvent amenés à créer plusieurs boosts pour compenser pour le manque d'uniformité dans l'enrichissement des produits (les variations sont souvent dues au facteur humain). La solution souhaitée consiste à remonter les produits en se basant sur le contenu partiel du champ et non pas complet, comme c'est le cas aujourd'hui.

Les KPIs choisis sont censés évoluer l'efficacité de l'évolution demandée, notamment, l'analyse du ranking pour analyser la disposition des produits sur la page ; les NPS pour voir si le nombre de détracteurs augmente ou diminue après la mise en place de la nouvelle fonctionnalité et puis, finalement, le taux de sortie, qui peut augmenter ou diminuer en fonction de l'impact du changement.

Ensuite, les spécifications fonctionnelles attendues sont mises en lumière (Figure 23).

Functional Requirement - Correspondance Partielle		
Business owner : Marion Hemery Writer : Marina Baidina		
#	Requirement	Priority
FR01	En créant des boosts pour les champs de type chaîne de caractères (à date : product.qualifier , product.packaging et product.nature), je peux choisir entre un match exact ou un match partiel de la valeur du champ.	MUST
FR02	Le boost créé permet de trier les produits comprenant la valeur en question dans leur champ si le mode "match partiel" est demandé, même si il comprend aussi d'autres caractères.	MUST
FR03	Le boost créé permet également de matcher les valeurs sans respecter la casse.	MUST
FR04	Le boost créé prend en compte les limites des mots. (ex : "eau-x" vs. "bateau-x")	MUST
FR05	Le boost créé permet également de matcher les valeurs correspondant aux lemmes.	SHOULD

3

Figure 23 : Spécifications Fonctionnelles

Chaque spécification est attribuée une priorité MUST ou SHOULD, en fonction de la nécessité de sa mise en place. Par exemple, il est nécessaire que le boost créé prend en compte les limite des mots ; en revanche, il n'est pas nécessaire qu'il corresponde aux valeurs correspondant aux lemmes.

Finalement, l'EB contient des exemples du besoin (Figure 24), ainsi que du fonctionnement souhaité (Figure 25).

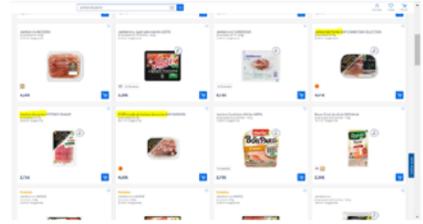
Functional Requirement - Correspondance Partielle

Business owner : Marion Hemery
Writer : Marina Baidina

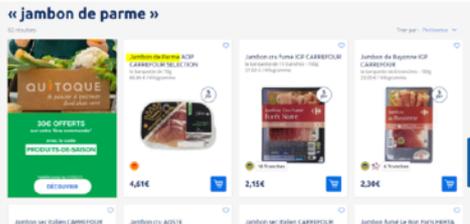
Exemple du problème à résoudre :

Exposé du besoin :
La page de la recherche "jambon de parme" fait apparaître trop bas les références pertinentes : le première se situe à la fin de la troisième ligne (Image à gauche), ce qui les rend difficilement trouvable par les utilisateurs.

Ranking actuel



Ranking souhaité



4

Figure 24 : Problème à résoudre

Functional Requirement - Correspondance Partielle

Business owner : Marion Hemery
Writer : Marina Baidina

Exemple du problème à résoudre :

Solution souhaitée :
Il n'existe pas de rayon "Jambon de Parme", conséquemment, la seule manière de faire remonter les produits pertinents est de booster le champ **product.qualifier**. Comme illustré, le contenu du champ n'est pas homogénéisé, ("de Parme AOP" vs. "cru de parme AOP") nécessitant soit une repasse de la part de l'équipe Enrichissement, soit, une création de boosts individuels pour chaque produit que l'on veut remonter ce qui multiplie considérablement le nombre de configurations et affecte la performance.



	PRODUIT	OFFRES	TECHNIQUES	SCORE
Filter				
product name	Jambon de Parme AOP CARREFOUR SELECTION			Jambon cru de parme AOP
product.qualifier	de Parme AOP			cru de parme AOP
product.sub_brand_name (Marque)	CARREFOUR SELECTION			
product.brand_name	CARREFOUR SELECTION			PARMACOTTO

5

Figure 25 : Solution souhaitée

La mise en production de la solution des boosts permettant une correspondance partielle est prévu pour la fin du mois de juin. Sa mise en place permettra à la fois, de

rendre compte de l'enrichissement non-homogène, améliorer la rapidité du moteur de recherche, en diminuant le nombre de configurations et d'optimiser le travail du métier.

Le plus grand enjeu de cette tâche est d'illustrer et expliquer le besoin Métier, sans pour autant de proposer des solutions concrètes, car cette mission est réservée à l'équipe SI. Donc, il est nécessaire que le besoin soit clair, bien fondé et compréhensible, mais il n'est pas censé donner des moyens de résolution du problème.

4.4 Résolution des bugs : cas de la lemmatisation

Afin de permettre une prise en compte des variations de mots recherchés en genre et en nombre, les textes de recherches subissent une couche de traitement supplémentaire : lemmatisation.

L'utilisation de la lemmatisation par un moteur de recherche e-commerce permet d'assurer l'apparition complète de l'offre, malgré les différences dans l'écriture des libellés, ainsi que dans les requêtes écrites par les utilisateurs du site.

Néanmoins, l'utilisation de la lemmatisation par le moteur de recherche de Carrefour est relativement récente : avant l'introduction de la prise en compte des lemmes, le moteur utilisait le stemming, faisant apparaître des **croissants** sur la page de la recherche "la **croix**".

L'introduction du dictionnaire de lemmatisation réduit¹¹, dont l'extrait est présenté dans la figure 22 a résolu les problèmes dont un a été décrit précédemment, en créant d'autres.

asperge	asperge	Ncfs
aspergeant	asperger	Vmpp
asperger	asperger	Vmn
asperges	asperge	Ncfs
asperge	asperger	Vmps-sm
aspergee	asperger	Vmps-sf
aspergees	asperger	Vmps-pf
asperges	asperger	Vmps-pm

Figure 26 : les formes "asperge" dans le dictionnaire de lemmatisation

En ce qui concerne le format du dictionnaire, il représente un fichier au format TXT, composé de trois colonnes de valeurs séparées par des tabulations. La première colonne correspond à la forme fléchiée ; deuxième, au lemme correspondant ; et puis, le troisième, à l'étiquetage morpho-syntaxique selon le système MULTEXT (Erjavec, 2020). Vu la chaîne

¹¹ plusieurs couples forme - lemme verbales ont été exclus, car en plus d'être inexploitable dans le contexte e-commerce, elles génèrent des erreurs.

de traitement subie par le texte de requête décrite dans les parties antérieures, le dictionnaire est transformé en ASCII, donc, ne contient pas de caractères diacrités.

Cependant, le dictionnaire seul ne permet pas de faire une distinction entre plusieurs formes, ce qui a entraîné des problèmes liés à la visualisation de l'offre complète sur plusieurs SRP. Plus précisément, dans le cas où une forme était ambiguë (correspondante aux plusieurs lemmes dans le dictionnaire de lemmatisation), le choix de lemme correspondant se passait d'une manière aléatoire : il n'y avait pas de moyen de déterminer quel lemme serait choisi.

Les figures 27 et 28 illustrent des exemples du comportement obtenu pour les recherches simples et multi-mots.

Impact vision Front

La recherche "couche" (~2000 recherches par semaine) ne donne que deux résultats non MKP :

Extrait du dictionnaire de lemmatisation :

```

couchers coucher Ncmp
couches couche Ncfp
couches-culottes couche-culotte Ncfp
couchette couchette Ncfs
couchettes couchette Ncfp
couche coucher Vmps-sm
couchees coucher Vmps-sf
couchees coucher Vmps-pf
couches coucher Vmps-pm
couci-couca couci-couca Rgp
    
```

Page de la recherche "couche" : 42 résultats

Page de la recherche "couches" : 148 résultats



Figure 27 : recherches "couche" et "couches"

Impact vision Front recherche multi-mots

La recherche "petits beurres" (~500 recherches par mois) ne donne aucun résultat non MKP.

Extrait du dictionnaire de lemmatisation :

```

beurre beurre Ncms
beurrer beurrer Vmn
beurres beurre Ncmp
beurre beurrer Vmps-sm
beurree beurrer Vmps-sf
beurrees beurrer Vmps-pf
beurres beurrer Vmps-pm
beurs beur Afpfp
    
```

Page de la recherche "petit beurre" : 37 résultats

Page de la recherche "petits beurres" : 2 résultats



Figure 24 : recherches "petit beurre" et "petits beurres"

Dans le but de modifier le comportement du moteur rapidement, une classe de synonymes “Lemmatisation” a été créée, permettant au Métier de mettre en place des configurations nécessaires pour une résolution temporaire des problèmes signalés.

Nous avons pu également constater l’impact causé en regardant les KPIs des recherches mal-lemmatisées avant la création des synonymes et après (Figure 25).

Mots clés	Nombre de recherche	Taux d'évolution volumétrie vs S-1	Taux d'ajout panier (Session)	Taux d'ajout panier (Recherche sur Ajout panier global)	Taux de sortie	Semaine d'analyse
pate brise	897	-16,48%	86,57%	108,47%	4,39%	S11
pate brise	1045	25,90%	2,23%	3,44%	0,37%	S13
eponge	3430	-18,59%	82,45%	94,69%	5,74%	S11
eponge	3761	17,83%	43,97%	44,56%	4,27%	S13
petits beurres	58	5,45%	112,50%	115,52%	28,13%	S11
petits beurres	67	52,27%	33,93%	44,78%	1,79%	S13

Figure 25 : KPIs de recherches affectées

Les taux d’ajout panier extrêmement bas et les taux de sortie trop hauts pour les recherches alimentaires montent l’urgence de la mise en place d’une solution pérenne.

Une solution envisagée consistait à effectuer une annotation morpho-syntaxique du corpus de recherches de l’année 2020, corrigées avec un outil de correction orthographique Language Tool, pour remettre les accents ; pour ensuite supprimer les entrées ambiguës du dictionnaire : par exemple, enlever *beurres* comme verbe de la deuxième personne singulier, mais garder *beurres* le nom au pluriel, car il est peu probable que le premier soit présent dans le corpus analysé.

L’annotation morpho-syntaxique a été effectuée avec le modèle `fr_dep_news_trf` de SpaCy, parce que c’est le modèle le plus performant pour ce type d’annotation (son accuracy est supérieur à 0.96).

Néanmoins, après la correction orthographique et l’annotation du corpus, dans le corpus annoté, constitué de 20154 mots, 453 tokens sont restés ambigus, ce qui n’a pas permis d’exploiter le corpus. En revanche, l’annotation du corpus a permis d’avoir une idée sur la répartition des catégories grammaticales (Figure 26).

NOUN	:	5265
PROPN	:	1523
ADJ	:	1400
VERB	:	557
X	:	216
NUM	:	203
ADV	:	70
PRON	:	44
ADP	:	39
DET	:	29
PUNCT	:	16
AUX	:	15
CCONJ	:	9
SCONJ	:	4
SPACE	:	3
SYM	:	3

Figure 26 : répartition des catégories grammaticales dans le corpus

Il a été proposé d'automatiser la prise en compte de noms de manière prioritaire sur les autres parties du discours, ou bien, de rechercher sur plusieurs lemmes sur une recherche donnée.

Les propositions ont été jugées techniquement irréalisables par l'équipe technique, donc, une solution moins sophistiquée a été mise en place : pour toutes formes ambiguës, les lemmes verbaux du dictionnaire ont été remplacées par des lemmes nominaux.

La figure 27 illustre l'extrait du dictionnaire obtenu.

asperge	asperge	Ncfs
aspergeant	asperge	Vmpp
asperger	asperge	Vmn
asperges	asperge	Ncfp
asperge	asperge	Vmps-sm
aspergee	asperge	Vmps-sf
aspergees	asperge	Vmps-pf
asperges	asperge	Vmps-pm

Figure 27 : extrait de la nouvelle version du dictionnaire de lemmatisation

Finalement, j'ai défini une liste de mot-clés à utiliser pour les tests de l'évolution.

La solution mise en place manque sans doute d'affinement, mais malgré, cela elle est suffisamment efficace et a bien permis de résoudre les problèmes générés par le dictionnaire de lemmatisation initial et a évité un développement long et coûteux d'une approche plus sophistiquée.

Discussion : Limites et enjeux

Les actions décrites dans la partie précédente ont été mises en place dans le but d'élaborer une solution « clé en main », donc, une solution qui peut être utilisée par le demandeur directement après sa finalisation. En effet, ces actions ont permis de contribuer à une optimisation progressive des performances du moteur de recherche et, conséquemment, d'améliorer l'expérience des clients, faisant leurs courses sur carrefour.fr.

Néanmoins, malgré l'efficacité évidente des solutions mises en place, il y existe des limites et des enjeux auxquels il peut être particulièrement difficile de faire face. Cette partie vise à mettre en exergue les facteurs dont j'ai observé une influence directe sur le déroulement de ma mission.

Enjeux :

1. Facteur humain. Comme mentionné précédemment, les données utilisées par le moteur de recherche, notamment, l'enrichissement des informations sur les produits est effectué d'une manière manuelle par une équipe dédiée. Cela provoque plusieurs problèmes liés à l'homogénéité du contenu des champs, utilisés pour la recherche, ce qui rend l'automatisation complète de la recherche difficile et nécessite une action supplémentaire de la part de l'équipe Search, qui risque de ralentir la réactivité du moteur de recherche.

2. Imprévisibilité. La manière dont les clients expriment leurs besoins n'est pas toujours évidente. Cela se traduit par des requêtes peu compréhensibles (« lait de vache » au lieu de « les deux vaches »), ou bien trop larges (par exemple, « boisson », « condiments »). Cependant, il peut être difficile de décoder le vrai besoin du client et pour le faire, il faut avoir du recul, un esprit analytique et une bonne connaissance de l'offre. Néanmoins, il peut prendre des semaines pour faire décoller la performance de ces genres de recherches, ce qui se traduit pour le temps et l'argent perdu.

3. Temps. Souvent, une prise de décision est influencée par le fait que nous ne pouvons pas nous permettre de passer trop de temps à résoudre des problèmes, car chaque jour, ou parfois même chaque heure compte. Comme c'était vu dans le cas de la mise en place du dictionnaire de lemmatisation, la solution a été choisie non à cause de sa finesse, mais surtout pour des raisons économiques. Donc, souvent, une résolution d'un problème, c'est une course contre la montre qui impacte la complexité et l'affinement des solutions techniques mises en place.

Limites :

1. Nécessité d'un contrôle humain permanent. Lors de mon alternance, j'ai remarqué que la perspective d'une automatisation complète du moteur de recherche était difficilement atteignable. En effet, une mise en place des modifications nécessite un contrôle humain, analyse, regard de la page du point de vue utilisateur, ce qui ne peut pas être effectué par une machine. Cela veut dire que dans un futur proche, le moteur de recherche ne pourra être rendu « entièrement autonome » du point de vue de l'indépendance complète d'un expert.

2. Complexité. Le moteur de recherche n'est qu'un composant du site, donc, pour que son fonctionnement soit optimisé, il est nécessaire que les autres composants soient également opérationnels, notamment, les autres services du site, par exemple, le catalogue, le FO, etc., mais aussi les données utilisées par le moteur : l'enrichissement de produits. J'ai remarqué que souvent les limites du développement sont imposées non par le moteur de recherche lui-même, mais par les autres composants, par exemple, les configurations sont compliquées par l'inhomogénéité d'enrichissement, ou bien, le fonctionnement souhaité n'est pas atteignable à cause des limites dues à la performance du FO. Le fait que la manière dont fonctionne le moteur de recherche repose sur un grand nombre de faits indépendants, constitue une limite importante, empêchant une proposition d'évolutions et la mise en place de certaines solutions.

Les limites et les enjeux observés m'ont permis de découvrir la complexité que représente une amélioration de l'UX d'un moteur de recherche sur un site e-commerce.

Conclusion

L'objectif de ce rapport était de présenter mon alternance au sein de l'équipe Search de Carrefour, qui m'a permis de découvrir une application inattendue de mes connaissances et compétences de taliste acquises au cours des deux années du Master en Traitement Automatique des Langues. Cet alternance a confirmé mon idée que le TAL peut servir d'une boîte à outils, que l'on peut appliquer à une multitude de domaines, même là, où l'utilisation des technologies du TAL ne semble pas être très intuitive. Ainsi, donc, dans mon cas, j'ai eu l'opportunité de non seulement mettre en pratique mes acquis en programmation, bases de données, recherche d'information et gestion de projet, mais aussi de découvrir un domaine dynamique d'e-commerce, qui m'est été inconnu auparavant.

Je suis ravie d'avoir eu l'opportunité de contribuer à l'amélioration de l'UX des clients de Carrefour, ainsi que de travailler sur des données réelles, issues d'utilisation du moteur de recherche. Au cours de mon alternance, j'ai été emmenée à élaborer et mettre en place certaines approches permettant une automatisation des actions, menées au sein de l'équipe, ainsi que de proposer des évolution plus générales du moteur de recherche. Egalement, j'ai pu découvrir les enjeux que représente une mise en œuvre d'un moteur de recherche sur un site marchand.

En plus de cela, grâce aux plusieurs interactions avec mes collègues, j'ai pu davantage développer plusieurs soft skills : présentation claire des idées, résilience, le sens du collectif.

Plus important encore, cette alternance m'a fait découvrir le métier de gestion de produit côté Métier, en tant qu'un possible débauché professionnel après le Master TAL. Cependant, dans l'avenir proche, je voudrais m'orienter vers un métier lié au SI (Systèmes d'Information), ce qui me permettrait de travailler directement avec la partie technique du projet.

Références bibliographiques

Apache. Apache Lucene - Index File Formats, Consulté le 23/05/2021, < https://lucene.apache.org/core/3_0_3/fileformats.html>.

Bessire, D. (1997) Grande distribution : l'efficacité de la régulation interne. Jacques Marseille (dir.), La Révolution commerciale en France. Du « Bon Marché » à l'hypermarché, Paris, Le Monde éditions, p. 165-182.

Carrefour. Mieux connaître Le Groupe Carrefour, Consulté le 30/04/2021, <<https://www.carrefour.com/fr/groupe>>.

Cogoluègues, A. Templier, T. Gregory, G. Bazoud, O. (2012) Spring Batch in Action, Manning Publications.

Delvalée, J. (2021) Résultats de Carrefour : « sa meilleure performance depuis 20 ans », LSA, c-Consulté le 30/04/2021, <<https://www.lsa-conso.fr/resultats-de-carrefour-sa-meilleure-performance-depuis-20-ans,373373>>.

Daumas, J. (2006). Consommation de masse et grande distribution: Une révolution permanente (1957-2005), Vingtième Siècle. Revue d'histoire, 3(3), 57-76.

Paturle, H. (2005). Marcel Fournier, l'hyperman : il était une fois l'homme qui inventa la grande distribution, Paris : Editions de la Martinière.