POLYTECH TOURS
64 avenue Jean Portalis
37200 TOURS, FRANCE
Tél +33 (0)2 47 36 14 14
www.polytech.univ-tours.fr

Rapport de stage 2021-2022

Services d'Analytics SEO TV

CONFIDENTIEL

Entreprise :
LIFAT
60 rue du plat d'Etain, 37000, Tours

Etudiant :
Segura Aurélien
DI4 2021-2022

Tuteur Entreprise :
Delalandre Mathieu
Maître de conférences

Tuteur académique :
Lenté Christophe

Table des matières

REMERCIEMENTS

Je souhaite remercier mon tuteur d'entreprise Mathieu Delalandre pour m'avoir offert l'opportunité d'effectuer ce stage, et plus particulièrement pour sa disponibilité et son investissement conséquent, qui ont contribué à rendre ce stage réellement épanouissant.
Je tiens également à remercier Angèle, autre stagiaire sur le projet Station TV, pour les échanges que nous avons eus et les données fournies pour la réalisation de mon travail.
Enfin, je souhaite remercier mon tuteur académique Christophe Lenté pour sa disponibilité et son assurance du bon déroulement du stage.

PARTIE 1

INTRODUCTION

Le Laboratoire d'informatique fondamentale et appliquée de Tours, LIFAT, est chargé de concevoir et de développer des modèles, méthodes et algorithmes, et assembler des ressources et outils logiciels pour extraire des informations, puiser des connaissances à partir de données en y intégrant les interactions homme-machine, et de résoudre des problèmes d'optimisation combinatoire avec la volonté d'obtenir de bons résultats en considérant fortement la problématique de temps d'exécution.

Le laboratoire est organisé en trois groupes de recherche :

- Base de Données et Traitement Automatique du Langage Naturel (BdTln)
- Recherche Opérationnelle, Ordonnancement et Transport (ROOT)
- Reconnaissance de Formes et Analyse d'Images (RFAI)

Le projet Station TV du LIFAT traite différentes thématiques autour de la capture et analyse de flux TV.
Ces thématiques sont constituées principalement de traitements techniques de « bas niveau » informatique, et la problématique SEO s'inscrit dans une démarche annexe qui permet d'avoir une analyse de « haut niveau » sur certaines données récupérées par ces captures. L'outil développé ici est appliqué au traitement de ces données (titres), mais peut être réutilisé dans tout secteur souhaitant porter une analyse de popularité sur ses contenus et améliorer son référencement sur le Web.
L'objectif de ce stage est de déterminer le volume de recherche associé à des mots-clés sur Google. L'accès public à ces données sur internet est très restreint. En effet, Google, dominant majoritairement le marché des moteurs de recherche, ne permet que d'accéder à ces informations en payant pour leurs services publicitaires tels qu'AdWords. En utilisant le service Google Trends fournissant des données précises et relatives pour des mots-clés spécifiés (pourcentages compris entre 0 et 100), nous allons calculer les valeurs de trafic absolues et volumes de recherche, à l'aide d'outils SEO gratuits qui nous fourniront des données de référence.

Le trafic d'un site web dépend de la visibilité du site Web sur les SERPs (Search Engine Result Page) qui dépend lui-même du SEO (Search Engine Optimization).
Il existe aujourd'hui de très nombreux outils / plateformes permettant de mettre en place une approche de SEO, et notre intérêt est porté ici sur le service Google Trends.
Ils se déclinent peu sous forme d'APIs et de services desquels il serait possible d'extraire des données. Une alternative est la caractérisation SEO de données issues de scraper Web. Grâce à cette approche, nous pouvons avoir des indicateurs sur le trafic potentiel d'un site / portail exploitant les données du scraper.

L'objectif est la mise en œuvre d'outils SEO pour la prédiction de trafic Web, et cela nécessite la mise en place d'une chaîne de traitement complète.

PRÉSENTATION DU SEO, MODÈLE LONGUE TRAINE

SEO (Search Engine Optimization) signifie en français : « Optimisation pour les moteurs de recherche ». Ce terme défini l'ensemble des techniques mises en œuvre pour améliorer la position d'un site web sur les pages de résultats des moteurs de recherche (SERP). On l'appelle aussi référencement. L'objectif d'un expert en référencement est d'améliorer la visibilité des sites web qu'il prend en charge en leur faisant gagner des places sur les moteurs de recherche (Google, mais aussi Yahoo ! Bing, etc.). En effet, il existe une compétition extrême entre les sites web pour atteindre la première page de résultats de recherche, et si possible les premiers résultats sur cette dernière. Il est également possible d'apparaitre parmi les premiers résultats en diffusant des annonces, mais cela ne correspond pas à un référencement naturel, mais sponsorisé. Le but est de faire se rencontrer les internautes intéressés par des produits / services ou du contenu informatif.

En statistique, la queue ou traîne d'une loi de probabilité correspond à la portion éloignée de la « tête » ou valeur centrale de la loi. Une loi de probabilité est dite à longue traîne si une plus grande partie de la loi est contenue dans sa traîne par rapport à celle de la loi normale.
Dans le cadre du référencement web, et de la vente en ligne, on parle de longue traine pour représenter un ensemble de produits/recherches plus ciblées, moins populaires, mais qui lorsqu'elles sont agrégées peuvent représenter une part significative du trafic.

Figure I.I

De façon analogue, dans le monde du commerce, l'application du modèle longue traine s'illustre par la stratégie consistant à vendre de petites quantités de produits très divers. On retrouve l'application de cette stratégie dans des entreprises telles qu'Amazon, ou Ebay.
Ce modèle est donc pertinent pour un grand nombre d'applications sur le web, tels que le e-commerce, le référencement sur Youtube, ainsi que sur les moteurs de recherche.

A l'inverse, une requête dite de courte traine sera constituée de peu de mots-clés, elle représente une requête générale, peu ciblée, et génerant un volume de recherche conséquent.

Dans le cadre de ce stage, différents types de requêtes ont été étudiés, principalement des titres de films, séries, émissions télévisées, mais également des Entités nommées, expressions linguistiques référentielles souvent associées aux noms propres et aux descriptions finies. Pour faciliter la compréhension dans la suite de ce rapport, l'appellation « base xmltv » fera référence à la liste des titres liés aux programme TV, obtenus par capture.
L'objectif thématique SEO de cette approche était de montrer, en comparant aux volumes de recherche de mots-clés de courte traine du même secteur, l'intérêt de se positionner sur un grand nombre de requêtes de plus longue-traine, plutôt que sur un mot-clé très populaire mais très compétitif, avec par exemple : « programme TV », qui génère un trafic mensuel d'environ 30 millions de recherches en France.

DONNÉES DE TRAFIC, NOTION DE TRAFIC ABSOLU ET RELATIF

Une métrique importante dans le domaine du référencement web est le volume de recherches, aussi appelé trafic. Il est caractérisé par le nombre de recherches effectuées sur les moteurs de recherche pour un mot-clé spécifié, sur une période donnée, ainsi qu'une localisation géographique si nécessaire.

Ce trafic peut s'exprimer de deux manières, on trouve dans un premier temps une valeur relative comme celle exprimée par la plateforme Google Trends, qui est représentée par une valeur normalisée entre 0 et 100 et qui correspond à une comparaison entre les volumes de recherches de différents mots-clés.

La valeur en trafic absolue est l'objet de ce stage, elle correspond au nombre effectif de recherches d'un mot-clé. Elle est sous cette forme directement exploitable pour un traitement SEO.

SOURCES DE DONNÉES

Pour une réalisation optimale du stage et la justesse du processus développé, il fut important de sélectionner des sources de données les plus pertinentes possibles.

Les données de trafic sont majoritairement détenues par Google, car c'est le moteur de recherche qui domine très largement son marché, ainsi une part très importante du trafic des internautes est capturée par Google et son moteur de recherche.

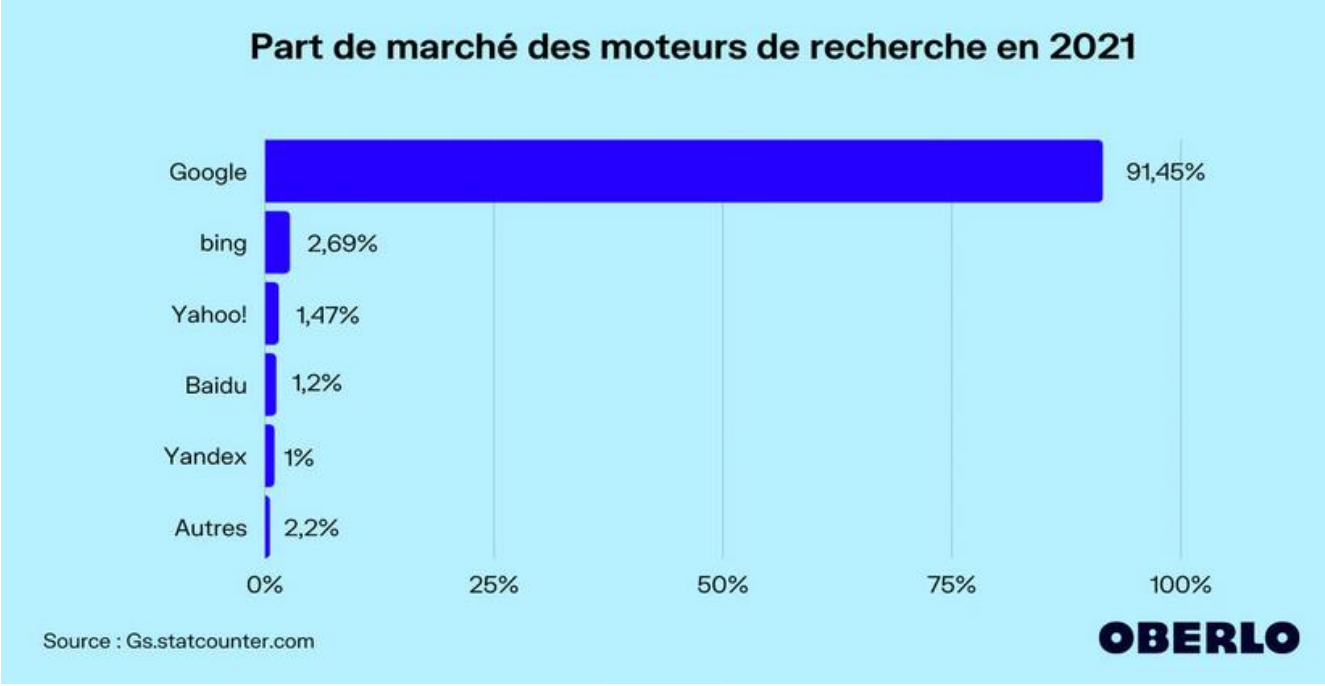

Figure I.II

L'accès gratuit à ces données reste très restreint et inexact, car il est plus généralement nécessaire de passer par des services publicitaires payants mis à disposition par Google, c'est le principe de Google AdWords, qui permet d'accéder aux valeurs de trafic relatives propres aux mots-clés sur lesquels nous portons notre attention. C'est l'outil publicitaire principal de Google qui permet aux utilisateurs de cibler ses publicités en fonction de certains mots-clés, notamment dans le cadre de campagnes marketings.

GOOGLE TRENDS

L'outils que nous utilisons est Google Trends, il permet d'avoir des informations sur les tendances de recherches sur une période donnée et sur une zone géographique précisée.
Les résultats reflètent la proportion de recherches portant sur un mot clé donné dans une région et pour une période spécifique, par rapport à la région où le taux d'utilisation de ce mot clé est le plus élevé (valeur de 100). Ainsi, une valeur de 50 signifie que le mot clé a été utilisé moitié moins souvent dans la région concernée, et une valeur de 0 signifie que les données pour ce mot clé sont insuffisantes.
La normalisation des valeurs de Google Trends en fonction des différences démographiques par zone géographique permet de comparer des valeurs normalisées dans des proportions comparables.

Il est possible de comparer jusqu'à 5 mots-clés simultanément sur Google Trends, ce qui nous permet ainsi d'augmenter l'efficacité de notre algorithme.

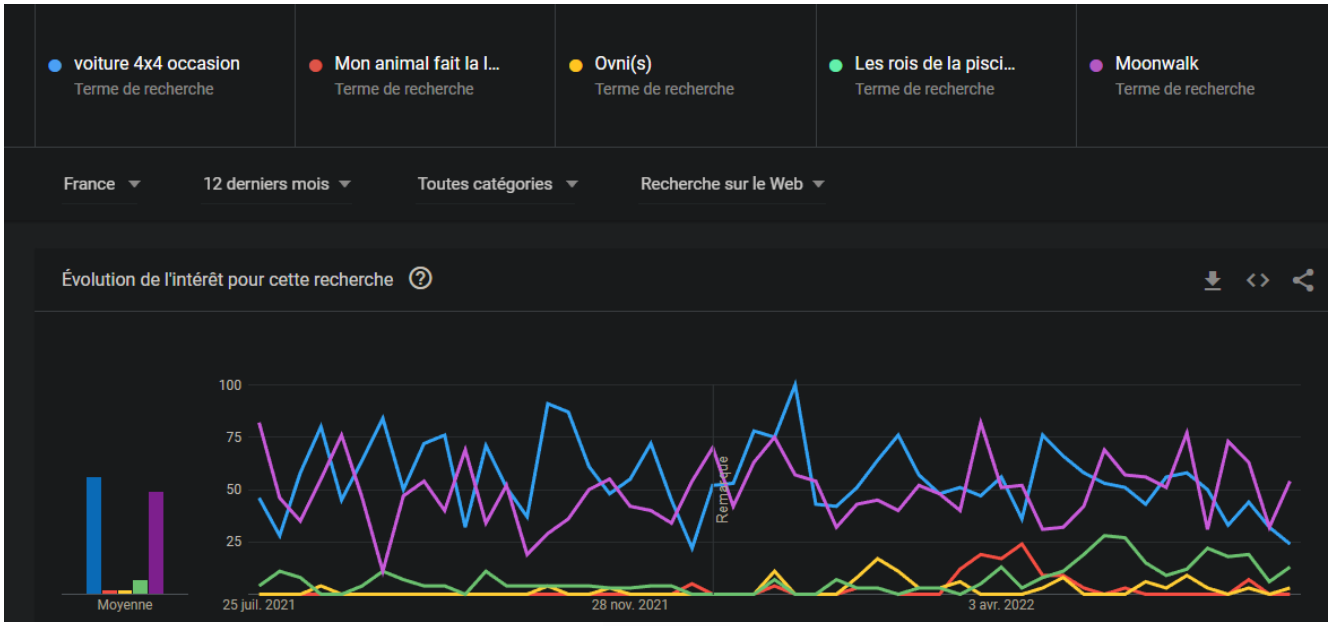

Figure I.III

Il existe sur internet des entreprises qui permette d'accéder à des outils de SEO. Si la majorité de leurs services est payante, elles mettent parfois des outils à la disposition des utilisateurs, gratuitement.
C'est le cas des plateformes Ahrefs et Wordstream.
Ces deux services permettent d'avoir accès à des valeurs relatives de trafic pour des mots-clés.
L'intérêt de ces services réside surtout dans le fait d'obtenir une liste de mots-clés associés à la requête dont on souhaite obtenir les informations, dans une optique de SEO, afin de cibler des mots-clés de longue traine, car, si l'exactitude des valeurs de volume données peut-être remise en question, il est toujours intéressant d'avoir un aperçu des recherches réelles effectuées par des utilisateurs autour d'un sujet donné.

Par exemple :

Mot-clé	KD	Le volume
chaussure	52	162K
chaussure nike	2	134K
chaussure femme	22	110K
chaussure homme	11	88K
meuble chaussure	10	61K
magasin chaussure	85	45K

Figure I.IV

Exemples de résultats obtenus par le « Keyword Generator Tool », (Outils de génération de mot clé) de la plateforme Ahrefs.

Dans le domaine de la SEO, il est crucial de connaître précisément les recherches des utilisateurs ciblés. L'accès à ces informations peut révéler certains intérêts des utilisateurs indiscernables sans ces outils.

Concernant la précision des valeurs de volume retournées par Ahrefs et Wordstream, l'incertitude semble être réduite pour des mots-clés populaires générant un trafic très important.
Le manque de transparence vis-à-vis des méthodes de captation des données ne permet qu'une interprétation des résultats, à l'instar d'une valeur juste qui proviendrait des résultats de recherche de façon exhaustive.

On peut également remarquer l'indicateur KD, abréviation de Keyword Difficulty, traduisant la notion de « difficulté », qui indique la compétition pour être référencé avec un certain mot-clé. Est mise en évidence ici l'importance du ciblage des bons mots-clés pour obtenir un référencement efficace, mais également en limitant son budget marketing dans le référencement SEO.
Il est possible de confirmer cela avec Wordstream, qui se définit comme étant un partenaire de Google et fournit des informations supplémentaires concernant le prix des enchères publicitaires pour les mots-clés ciblés.
En effet, dans les systèmes publicitaires de Google et Youtube (possédé par Google depuis 2006), le système d'enchère permet de choisir parmi différents clients en fonction de la valeur de leur enchère pour un mot-clé. Par exemple, si un client souhaite apparaitre dans un emplacement publicitaire en ciblant le mot-clé « chaussure », il aura ainsi plus de chances d'être sélectionné si son enchère est plus élevée que celle d'un autre client.
Les valeurs listées ci-dessous par le site Wordstream illustrent bien cet aspect :

Keywords	Monthly search volume	Top of page bid (low range)	Top of page bid (high range)	Competition
chaussure	368,000	$0.20	$0.67	HIGH
basketes	246,000	$0.17	$0.50	HIGH
sandale	60,500	$0.16	$0.55	HIGH
new balance femmes	60,500	$0.06	$0.24	HIGH
timberland hommes	60,500	$0.06	$0.41	HIGH

Figure I.V

Il est aujourd'hui difficile de définir les sources précises des données fournies par des outils tels que Ahrefs, Moz, ou bien Wordstream. Ce dernier annonce être un partenaire de Google, ce qui peut conforter la confiance portée en ses résultats, mais là encore se pose la question de la pertinence des données, le service étant gratuit. En effet, la majorité des services renseignant des données vérifiées est payante, on peut donc remettre en question cet accès gratuit aux données 'officielles'.
Pour la grande majorité des sites, y compris SemRush, spécialisé dans la SEO, une partie des données provient de leur analyse des recherches des utilisateurs utilisant leur barre d'outils navigateur dédiée à la SEO.
Les résultats semblent donc être plutôt biaisés et inexacts. Peut-être peuvent-ils compléter ou croiser leurs données avec celles de Google, mais là encore, nous n'avons pas d'information disponible à ce sujet.

Ainsi, les valeurs de trafic que nous sélectionneront en tant que référence pour nos calculs proviennent de sources difficilement vérifiables, on peut donc prendre en compte cette incertitude, qui, dans notre utilisation reste acceptable, notre intérêt se portant sur les ordres de grandeur de volumes de recherches et non sur des valeurs précises.

LIMITES DE GOOGLE TRENDS

L'utilisation de Google Trends en tant que source de données principale n'est pas sans défauts.
Il est impossible, pour un utilisateur du service, d'avoir le détail de l'intégralité des éléments pris en compte pour l'obtention d'un résultat donné. Notre modèle présente donc certaines limites.
Il semble en effet que nous n'ayons que peu d'informations quant aux termes qui sont effectivement englobés par Google Trends lorsque l'on effectue une recherche.

Pour illustrer ces propos, en sélectionnant des mots-clés similaires, mais exprimant des idées différentes, on constate qu'il est possible que les mots-clés interprétés par Google Trends soient considérés de manière commune.

Plus explicitement, on peut le constater sur le graphique suivant :

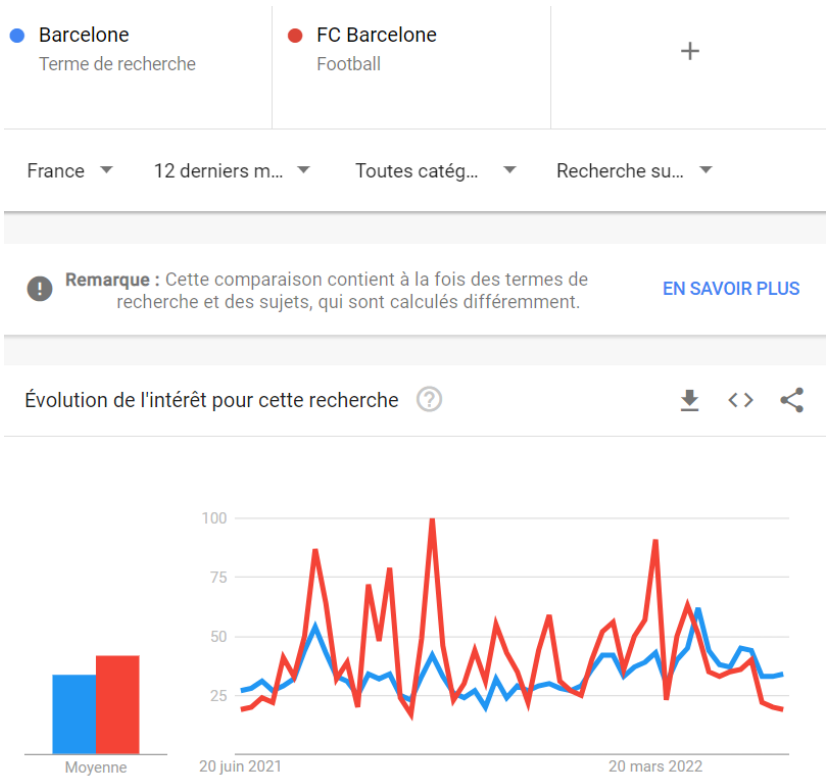

Figure I.VI

Avec les termes de recherches « Barcelone » et « FC Barcelone », mais également avec le « topic », thème associé (ici en lien avec le « Football »), on peut remarquer la similarité entre les deux les courbes, on retrouve exactement les mêmes pics aux mêmes périodes, ce qui implique que certains mots-clés sous-jacents pris en compte par Google Trends soient communs entre les deux requêtes.

Mise en garde : Lors de la recherche de termes sur Google Trends, il faut veiller à comparer les termes de recherche spécifiquement, et non pas le « topic » qui peut lui être associé, car ce dernier englobe les résultats de recherche associés, comme détaillé ci-dessous.

Les « topics », ou thèmes, sont un groupe de termes qui partagent un concept commun dans n'importe quel langage. Ils sont parfois proposés par le service en fonction du mot-clé recherché. Dans le cas d'une recherche pour le mot-clé "London" ou l'on sélectionne le thème correspondant, les résultats de recherche incluront des éléments tels que « Capitale du Royaume-Uni », ou encore « Londres », traduction du mot en Français et Espagnol.

Il faut donc veiller à restreindre la recherche au mot-clé qui nous intéresse et non pas au topic associé, qui retournerait des résultats non pertinents dans notre cas d'utilisation.

Exemple :

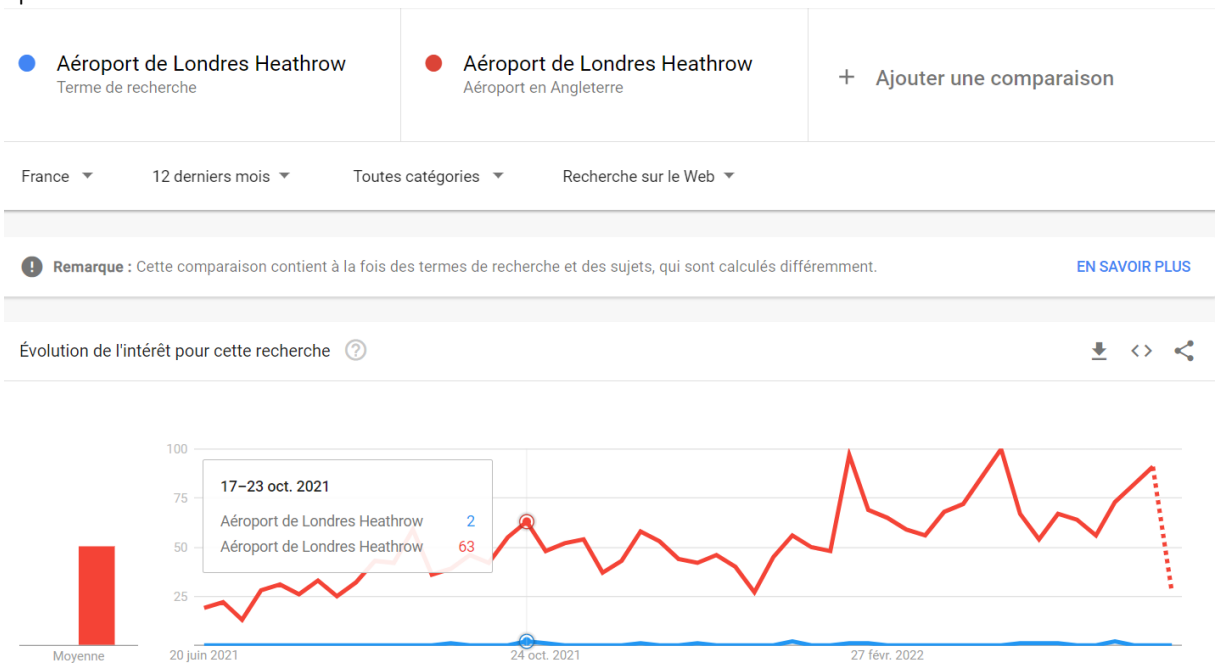

Figure I.VII

On constate effectivement ici qu'une erreur dans le type de recherche retourne des résultats complètements différents. (topic ou « simples » termes de recherche) .
Exemple ici le type « Aéroport en Angleterre » engloberait les recherches liées à ces mots-clés et à ce sujet, on retrouverait donc potentiellement des requêtes du type « date de construction aéroport de Londres », « aéroport de Londres », « London Airport », ou même des termes dans d'autres langues.
A l'inverse, pour le type « terme de recherche », cela retourne uniquement les mots-clés explicitement mentionnés, et uniquement dans notre langue.
En vérifiant sur un ensemble de mots-clés, on constate bien que l'API que nous utilisons extrait les données des « termes de recherches » uniquement, et non pas les « topics ».

PARTIE 2

REALISATION

PYTHON ET PYTRENDS

Dans le cadre de la réalisation du stage, l'utilisation du langage de programmation python est particulièrement adaptée puisque l'API (Application Programming Interface) que nous utilisons se nomme « Pytrends ».
Cet outil n'est pas développé par Google, mais par l'utilisateur « GeneralMills » et disponible sur GitHub. En effet, Google apportant régulièrement des modifications au site Google Trends, parfois dans le but de bloquer ces accès par API non officielles, le code source est accessible et des modifications peuvent y être apportées afin de garder une API fonctionnelle. Par manque de connaissances, je n'ai pas pu interpréter correctement le contenu du code source de l'api, car elle contient du javascript avancé.
Pytrends permet d'effectuer des recherches Google Trends depuis un script et de récupérer les données sous forme de Dataframe du module Pandas sur Python. La Dataframe est une structure de données qui organise les données en lignes et en colonnes, ce qui en fait une structure de données bidimensionnelle.

LIMITES DE PYTRENDS

Un des problèmes les plus importants rencontrés lors de l'utilisation de pytrends est la détection anti-bot de Google.
En effet, lorsque des requêtes effectuées par Pytrends sont enchainées trop rapidement, l'api retourne un code d'erreur 429. La réception de ce message d'erreur induit un blacklistage de l'adresse ip utilisée pour effectuer les requêtes, pendant une durée indéterminée. Ce blocage peut durer de quelques minutes à plusieurs heures, et plus rarement une journée.
Il est donc nécessaire de trouver un délai minimal qui permette une bonne exécution du programme sans interruption.
En ayant effectué de nombreux tests, il semble qu'un délai de 8 s entre les requêtes soit nécessaire afin d'éviter de déclencher l'erreur 429. Il est également possible d'ajouter des délais supplémentaires toutes les x requêtes par précaution. En effet, lors d'un test effectué sur 5 000 requêtes et 3 jours d'exécution, le système anti-bot a bloqué les requêtes, ce qui signifie que pour une vitesse moyenne de téléchargement de 1,15 requête par minute, l'erreur 429 est survenue. Cette vitesse de requêtage est due aux délais rajoutés toutes les 10 et 30 requêtes, et un délai minimal de 11 s était ajouté entre chaque requête. Le téléchargement des données depuis Google Trends reste donc expérimental, car aucune donnée juste n'est fournie par Google concernant la vitesse de récupération des données.
Également, il est possible d'agir sur deux paramètres de l'api pytrends :
« retries » et « backoff_factor ».
Lorsque pytrends effectue une tentative de téléchargement des données depuis Google Trends, il se peut qu'une erreur quelconque survienne, liée au réseau, au site, à l'api, à la forme de la requête à soumettre, ou encore l'encodage, parmi les sources d'erreurs les plus fréquentes.
Ainsi, grâce au premier paramètre, il est possible de définir le nombre de tentatives de téléchargement pour une requête donnée, en cas d'échec initial.
Il est possible ensuite d'agir sur le délai entre chacune de ces tentatives avec le second paramètre.

Le délai est calculé selon la formule suivante :
{backoff factor} * (2 ^ ({valeur de retries} - 1)).

Ainsi, si le backoff_factor est de 0.1, alors le délai sera de [0.0s, 0.2s, 0.4s,…] entre les tentatives.
Il est important de noter que le programme n'ajoute pas de délai entre la requête initiale et la seconde tentative, car cela est en général suffisant pour corriger l'erreur. En revanche, il est probable que Google Trends interprète cette double demande instantanée comme une connexion malveillante.
La reproductibilité de l'évènement est complexe, et cet aspect dépend certainement du type d'erreur, et si elle provient en amont ou lors de l'interaction avec Google Trends.

En revanche, il peut arriver que des blocages se déclenchent après plusieurs milliers de requêtes, et l'unique moyen de déterminer le délai idéal semble être les tests à grande échelle. (Plusieurs milliers de requêtes)
De plus, si Google décide de renforcer son blocage anti-bot, la recherche d'un nouveau délai entre requêtes peut être obligatoire.

Il est difficile d'estimer un taux de téléchargement théorique idéal sur Google Trends, car cela dépend tout d'abord de la méthode employée pour télécharger les fichiers CSV.

Il existe plusieurs façons de limiter les blocages. En effet, lorsque l'on utilise dans le script des headers simulant une connexion d'un utilisateur avec un compte Google, le site est plus tolérant.

Lorsque l'on utilise l'outil pyautogui (programmation graphique) pour répliquer une interaction humaine sur un navigateur, Google Trends détecte moins facilement une automatisation de l'extraction des données. Cela peut encore être optimisé lorsque l'utilisateur est connecté au préalable sur le navigateur utilisé pour le processus.

Lorsque l'on utilise l'api, y compris en essayant d'ajouter des headers sensés simuler une connexion par navigateur avec utilisateur connecté, google Trends détecte ce processus et le bloque plus rapidement. En effet, le téléchargement des données est au moins 2x plus lent lorsqu'il est effectué par l'api.

En essayant de répliquer cette méthode dans le cas actuel avec pytrends, il apparait être ainsi plus compliqué, car l'utilisation de l'api ne nécessitant pas d'ouverture de navigateur, il plus difficile de « tromper » la détection anti-bot, car donc impossible de se connecter manuellement.
Pour résoudre ce problème, il faut essayer de répliquer la partie headers obtenus lors d'une utilisation du navigateur en étant connecté. Pour se faire, en se connectant au site Google Trends en étant connecté, sur un navigateur, nous allons observer la requête effectuée d'un point de vue réseau, lors d'une demande de résultat, et nous allons utiliser ces headers dans le cas de l'utilisation avec pytrends. Le système de détection va ainsi penser que la demande de requête émanant de pytrends est effectuée depuis un navigateur avec l'utilisateur connecté.

Les headers, obtenus dans la requête réseau depuis l'outil d'inspection réseau intégré aux navigateurs, sont copiés au format cUrl puis convertis sous forme de code de la forme suivante :

```
headers = {
    'authority': 'trends.google.fr',
    'accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9',
    'accept-language': 'fr-FR,fr;q=0.9,en-US;q=0.8,en;q=0.7',
    'cache-control': 'max-age=0',
    # Requests sorts cookies= alphabetically
    # 'cookie': '__utma=108793543.338554105.1635232609.1655132842.1655213348.33; __utmc=108793543; __utmz=108793543.1655213348.33.5.utmcsr=google|utmccn=(organic)|
    'referer': 'https://trends.google.fr/trends/?geo=FR',
    'sec-ch-ua': '" Not A;Brand";v="99", "Chromium";v="101", "Opera";v="87"',
    'sec-ch-ua-mobile': '?0',
    'sec-ch-ua-platform': '"Windows"',
    'sec-fetch-dest': 'document',
    'sec-fetch-mode': 'navigate',
    'sec-fetch-site': 'same-origin',
    'sec-fetch-user': '?1',
    'upgrade-insecure-requests': '1',
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.67 Safari/537.36 OPR/87.0.4390.45',
}
```

Figure II.I

On y retrouve des éléments qui permettent d'identifier le type de navigateur, des identifiants de session et de connexion utilisateur.

Ensuite, il pourrait être possible d'utiliser un système d'IP rotative qui changerait toutes les x requêtes, car c'est en effet l'adresse IP qui est détectée par le système anti-bot, afin de cibler l'origine des requêtes.
Si l'on se trouve bloqué à la suite d'une erreur 429 de Google Trends, il est possible par exemple d'utiliser un vpn afin d'obtenir une adresse IP différente et contourner ce blocage.

Lors de tests préliminaires, un temps d'exécution de 10 h avait été suffisant pour effectuer 2698 requêtes, soit environ une requête toutes les 12 secondes, en moyenne.
La tolérance du système de détection est donc variable du point de vue de l'utilisateur effectuant les requêtes. Opter pour une fréquence de téléchargement plus lente est donc une mesure de précaution qui permet, en règle générale, d'assurer la non-interruption du programme.

Il existe également d'autres erreurs qui peuvent survenir durant l'exécution du programme, avec notamment l'erreur 423 (Timeout), provenant en général d'une erreur « SSL Handshake », protocole permettant entre autres l'identification mutuelle entre le serveur et le client.

Il n'est pas possible de fragmenter l'exécution du programme sur plusieurs machines, en disposition locale, car les adresses IP sont identiques ou similaire (ex : 192.168.206.25 – 192.168.206.26), et Google Trends bloque ainsi les exécutions en parallèle.
Afin d'optimiser le processus de téléchargement, la méthode préférable à utiliser est, après définition fonctionnelle de la vitesse de téléchargement en requêtes/heures, la séparation sur plusieurs Machines Virtuelles, chacune utilisant un VPN pointant vers une adresse/pays différent. Dans notre cas, nous avons utilisé le VPN gratuit urban.me, qui permet d'utiliser des adresses IPV4 situées dans des pays du monde entier.
L'utilisation d'un vpn avec des adresses géographiquement éloignées pourrait en général sembler être un facteur ralentissant dans l'exécution d'un algorithme, en raison de la vitesse de transmission des données, mais dans notre cas le facteur qui bride la vitesse d'exécution de l'algorithme est l'attente entre chaque soumission de requêtes sous Google Trends (>10s).

On se soucie donc peu de l'emplacement géographique et de la vitesse de transmission de l'adresse fournie par le VPN, dans la mesure ou la connexion reste stable et continue.

En termes d'optimisation, on notera le grand intérêt d'un vpn, qui permet d'éviter l'attente liée aux blocages, qui semble être pour l'instant indéterminée.

Un autre élément qui rentre en jeu est la mise à jour régulière de Google Trends par Google, il est donc nécessaire de surveiller l'évolution du site vis-à-vis de l'api, car pytrends, les temps de téléchargements et autres paramètres peuvent devenir obsolètes.

Selon le développeur de l'API, le taux de téléchargement est défini par défaut à 100 requêtes sur 100 secondes, et peut être ajusté à une valeur de 1000. Le nombre maximum de requêtes est donc limité par l'API à 10 requêtes par seconde par utilisateur.
Dans les faits, nous n'avons jamais pu exploiter l'API avec une telle vitesse de téléchargement, en raison des blocages mentionnés précédemment.

Un des avantages de pytrends est l'optimisation des téléchargements, en effet l'api gère automatiquement les requêtes qui ne génèrent pas de résultats par manque de données, et évite donc les téléchargements non pertinents, vérification que nous devions faire manuellement, précédemment en utilisant un outil OCR pour détecter en amont l'absence effective de résultats pour une requête donnée.

L'ALGORITHME

En ce qui concerne l'algorithme mis au point durant ce stage, une optimisation continue était une priorité, dans une optique de diminution de la complexité du programme.
Cependant, le principe de fonctionnement général reste identique.

En effet, l'objectif est de trouver cinq mots-clés pour chaque recherche à effectuer sous pytrends :
Un mot-clé de référence, dont on connaît le trafic en nombre quantitatif / jour (par exemple), et quatre mots-clés, par exemple, des titres provenant de la base xmltv.
En connaissant ensuite le volume de recherches du mot-clé de référence, et en s'assurant que sa valeur maximale atteigne les 100% dans les résultats Google Trends, on peut calculer par un produit en croix la quantité de trafic quantitative des quatre mots-clés associés.

Attention, il faut bien s'assurer que le mot-clé de référence soit à 100%, et que les autres mots-clés soient dans un ordre de grandeur suffisamment proche de celui du mot clé de référence. (ex : 10k requêtes vs 1k, et ne pas comparer un mot clé à 1M requêtes quotidiennes avec un mot à 2k, qui serait alors évidemment représenté par une valeur de 0 sur le graphique Google Trends).
On se contentera de regarder les requêtes françaises dans un premier temps.

Pour effectuer une requête, il apparait plus pertinent de sélectionner un mot-clé de référence parmi la liste des mots-clés de référence en fonction de leur valeur croissante de trafic plutôt que par dichotomie, car en ayant une approche par valeur croissante, nous avons la certitude de prendre le mot-clé de référence qui sera dans un ordre de grandeur <10 en termes de quantité de trafic par rapport aux 4 autres mots-clés de la requête. Cela permet d'obtenir des courbes exploitables, avec des valeurs régulièrement supérieures à zéro.
La constitution de la base de mots-clés de référence est évoquée en partie 3 de ce rapport.

Le programme développé durant ce stage se décompose principalement en 4 parties.
La première partie consiste à récupérer une liste de titres d'une base xmltv, constituée à partir de la capture de programme TV (restreint ou non aux chaînes TNT) assurée par le LIFAT.
Ensuite, un processus de « standardisation » des mots-clés/titres est nécessaire.
L'objectif étant d'avoir une valeur de trafic la plus proche de la réalité, il peut être important d'adapter la requête à sa « version » la plus populaire saisie par les utilisateurs.
De plus, l'interprétation de Google Trends de différentes variations d'un titre reste très abstraite et difficilement généralisable.
Pour clarifier ces éléments, si nous utilisons le titre « Ma sorcière bien-aimée », il est possible de faire varier la casse avec les majuscules, les accents, et la présence du trait d'union.
Ainsi, on obtiendra des résultats différents pour les mots-clés : « ma sorcière bien aimée », « ma sorcière bien-aimée », ou encore « ma sorciere bien aimee ».

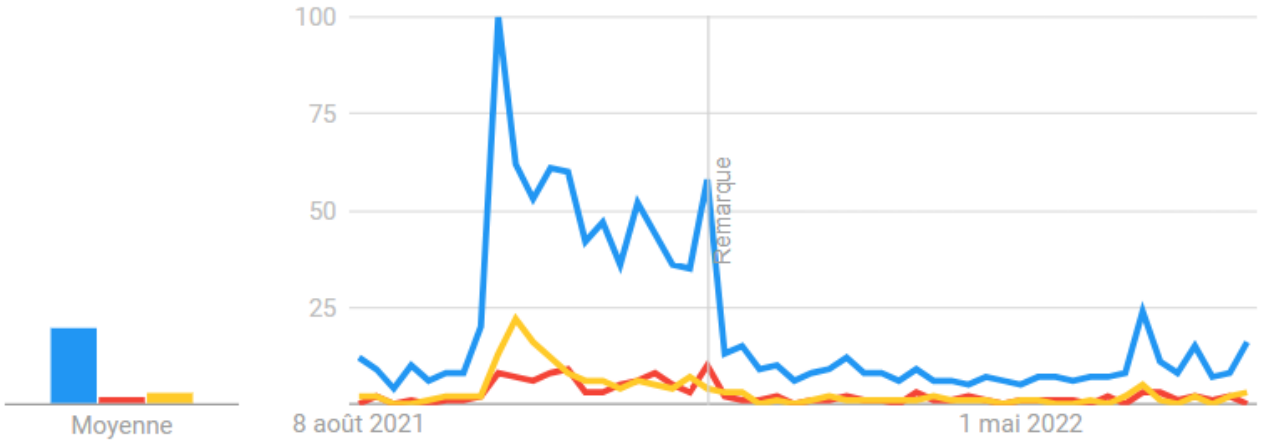

Figure II.II

Chaque variante retourne des résultats différents sous Google Trends. Il pourrait sembler intéressant de généraliser le même traitement à chaque mot-clé de la base de titres, en revanche il semblerait que ce soit du cas par cas pour Google, et ce qui est applicable pour un mot-clé ne l'est pas nécessairement pour un autre. L'exemple ci-dessous illustre bien ce propos :

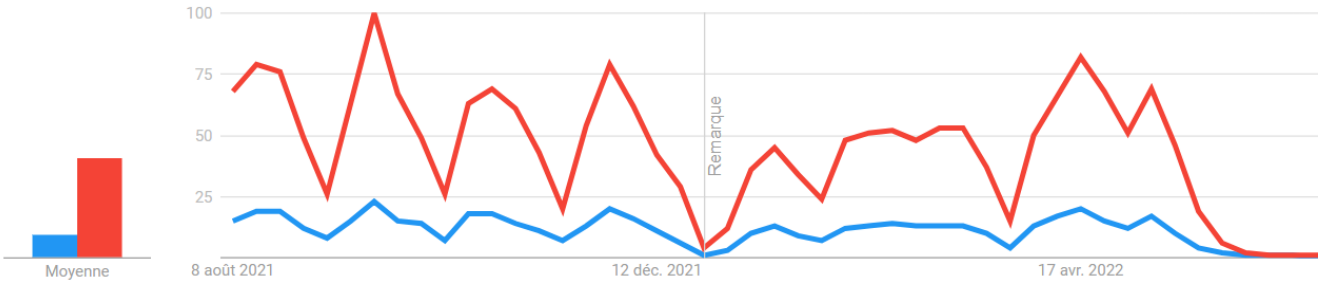

Figure II.III

Cette fois-ci, le mot-clé qui semble être le plus recherché est la version sans accent : « resultat ligue 1 » en rouge, contre « résultat ligue 1 » en bleu.

Une règle que l'on peut constater à plus forte tendance sur Internet, de la part des utilisateurs, est une saisie minimaliste des mots-clés, la plus simple et rapide, soit une version sans majuscule, accents, ni traits d'union, cependant, cette tendance semble être plus adaptée aux noms communs.
Google Trends considère le mot-clé avec ou sans majuscule(s) de façon identique.

Il est donc relativement difficile de trouver une forme « standard » que l'on pourrait appliquer à tous les mots-clés.

L'ensemble de l'algorithme est construit autour de pytrends, et particulièrement des requêtes au format suivant :

```
kw_list = ["Blockchain",”Ethereum”,”Bitcoin”]

pytrends.build_payload(kw_list, cat=0, timeframe='today 5-y', geo='', gprop='')
```

Figure II.IV

On constate que l'on y retrouve effectivement les différents paramètres de recherche Google Trends : la zone géographique, la période, mais il existe également la possibilité de rechercher un « topic », ou thème, plutôt qu'un simple mot-clé.
Par exemple, pour le mot-clé « London Heathrow », il est possible d'obtenir uniquement les résultats sur ce mot-clé, ou bien de recueillir des résultats des recherches liées au sujet aéroport « London Heathrow ». Cependant, cette partie « sujet » n'est pas très pertinente pour notre application, car, de nouveau, nous n'avons aucune information sur les termes englobés par cette notion de « topic ».

La première partie du programme consiste donc à assurer le bon encodage des caractères, par exemple pour des chaînes allemandes telles que Arte avec le caractère « ß », la suppression des « - » qui donnent des résultats sous-représentatifs, et tout élément de syntaxe qui pourrait perturber le formatage de la requête Google Trends. En-dehors du formatage utf-8, une quarantaine de règles de correction ont été ajoutées manuellement afin de s'assurer d'éviter toute erreur d'encodage.
En effet, un caractère erroné dans la liste de mots d'une requête pytrends induit une erreur.

Le cœur du programme est ensuite dépendant d'une base de mots-clés de référence.
En effet, le principe du processus est d'utiliser des mots-clés de référence dont on aura une valeur absolue de trafic approximée au mieux à partir des outils Ahrefs et Wordstream, et auxquels nous compareront la liste de mots-clés/titres afin d'en déterminer le volume de recherche absolu.

La constitution de la base de mots-clés de référence est une étape importante, puisque c'est à partir des valeurs de trafic de ces mots que nous allons déterminer le trafic des titres qui nous intéressent.

Il est important de constituer la liste avec suffisamment de mots-clés, puisqu'il est nécessaire de comparer des mots d'ordre de grandeur similaire sur Google Trends.

ReferenceKeyword	AverageTraffic
facebook	50 250 000
youtube	43 250 000
programme tv	35 600 000
gmail	23 950 000
google	19 200 000
Paypal	2 170 000
tiktok	1 410 500
iphone 11	743 000
iphone 12	675 000
zidane	275 000
doctolib mon compte	212 000
recette quiche lorraine	105 500
avengers endgame	83 500
mini frigo	43 750
carte anniversaire gratuite	38 250
accordeur guitare	21 000
pink lady	9 000
peluche disney	7 600
voiture 4x4 occasion	5 000
chaise en osier	4 000
gamelle pour chien	3 450
petit perroquet	1 900
pneu lisse	950
petit chien roux	415

Figure II.V

En effet, si l'on compare le mot-clé Facebook avec un titre suscitant peu d'intérêt sur internet, les résultats vont être peu exploitables, car le mot-clé de référence est bien trop populaire, par exemple :

Figure II.VI

On constate bien ici qu'il est difficile d'analyser le volume de recherche du terme « zig & sharko » s'il est comparé au mot-clé « Facebook ».

En revanche, si l'on compare ce titre avec un autre mot-clé de la base de référence dont le volume de recherche est similaire, ou d'un ordre de grandeur environ dix fois supérieur au maximum, la comparaison est plus intéressante :

Figure II.VII

Il est donc crucial d'avoir une liste de mots de référence qui couvre une plage de trafic importante.

Une première étape du programme permet de filtrer les doublons de la base, mais également les mots-clés/titres qui ne suscitent pas d'intérêt sur Internet.
En effet, pytrends envoie un DataFrame vide lorsque le mot-clé retourne des valeurs de recherche insuffisantes.

Au cours de cette étape, chaque mot-clé de la liste des titres va être associé à un mot-clé de référence dont le volume de recherche est du même ordre de grandeur.
Cependant, un nombre important de mots constituant la longue traîne génère un faible trafic. Ainsi, afin de réduire la complexité et la durée d'exécution du programme, ces mots vont être associés aux 4 plus petits mots-clés de la liste de référence, si le volume de recherche est d'un ordre de grandeur similaire. Cette optimisation permet une plus grande mise à l'échelle du programme, car plus de 40% des requêtes sont généralement filtrées par ce processus.

Une requête correspondrait ainsi à cette recherche sous Google Trends :

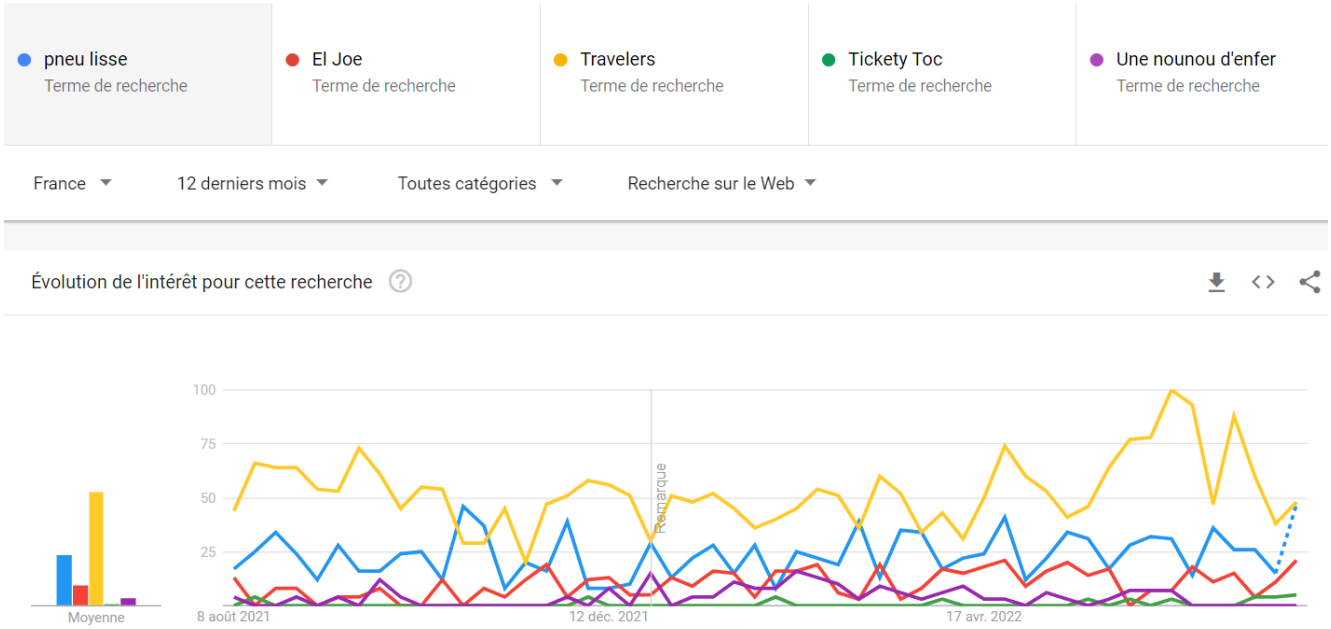

Figure II.VIII

Ici, le mot-clé de référence est « pneu lisse », et les 4 autres mots-clés sont des titres de la base de titres xmltv, dont on souhaite déterminer le volume de recherche cumulé des éléments qui la composent.

Ensuite, chaque mot-clé/titre étant associé à un mot-clé de référence, il est possible de calculer son volume de recherche estimé à partir de la comparaison Google Trends.
En effet, puisque l'on connaît la valeur de trafic absolue (estimation Ahrefs/Wordstream) du mot-clé de référence, sa valeur relative (comparaison Google Trends), ainsi que la valeur relative du mot-clé titre (comparaison Google Trends), il est alors possible de calculer sa valeur de trafic absolue.

On utilise en effet pour le calcul les valeurs moyennes indiquées que l'on peut observer en bas à gauche de l'illustration ci-dessus. La figure ci-dessous permet de schématiser le processus de calcul.

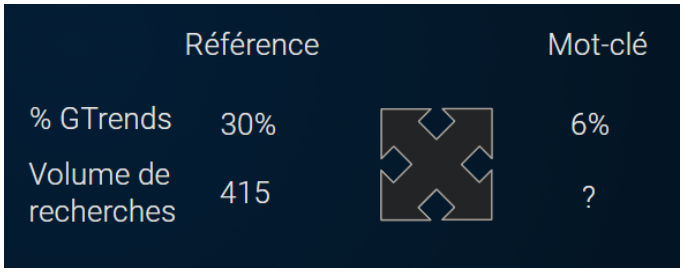

Figure II.IX

En utilisant cette méthode du produit en croix, on constate que la variable qui est potentiellement inexacte est la valeur moyenne de trafic absolue du mot-clé de référence.
En effet, cette valeur est calculée à partir des données incomplètes et difficiles à sourcer des sites Ahrefs et Wordstream, il est donc important de réduire l'erreur/incertitude.

Notre approche consiste à obtenir un ordre de grandeur pour le volume total de recherches associées aux mots-clés, ainsi, une précision affinée n'est pas nécessaire dans notre application.
En revanche, il reste pertinent d'essayer de corriger cette erreur pour d'autres applications ou la précision dans les résultats serait particulièrement recherchée.

L'affinage de la précision de cette base de référence et de valeurs de trafic associées sera évoqué en partie 3 de ce rapport.

NOTION DE COMPLEXITÉ

La notion de complexité est très importante dans notre programme. En effet, puisqu'un délai de plusieurs secondes est requis entre chaque requête, il faut veiller à en effectuer le moins possible.
Cela est d'autant plus illustré par un nombre important de requêtes, car plusieurs milliers de requêtes à traiter peuvent ainsi requérir plusieurs jours d'exécution.
En intégrant l'étape de 'filtrage' développée dans la partie suivante, la complexité de l'algorithme peut être décomposée en 3 parties :

- Le 'filtrage' des requêtes en complexité O(n/4)
- L'affectation mot-clé titre/mot-clé de référence en complexité O(log(m+1)n'/4) avec n'<<n
- La caractérisation du trafic en complexité O(n'/4)

La variable m représentant le nombre de mots-clés de référence, et n représentant le nombre de mots-clés/titre dans la base xmltv à analyser.

ETAPE DE "FILTRAGE" POUR LA MISE À L'ÉCHELLE

L'étape de sélection des mots-clés à faible trafic permet d'optimiser grandement la durée d'exécution du programme, car une seule itération suffit à associer ces mots aux 4 plus petits mots-clés de base, contre 24 itérations correspondantes aux 24 mots-clés de référence de base.

Ainsi, pour la mise à l'échelle sur une base de plusieurs milliers de titres/mots-clés, il est possible de retrouver plus de 40% des mots-clés qui soient ainsi associés au plus petit volume de recherche.

Sur une liste de 2000 mots dont 40% vont être filtrés comme générant un petit volume de recherche, cette méthode permet de réduire le nombre d'itérations par 23*800 = 18400.

Avec un délai d'attente de plusieurs secondes entre chaque requête, on constate ainsi une augmentation considérable en performance, et un gain de temps de plus de 24h.

INDEPENDANCE CONTEXTUELLE DES REQUETES ET LABELLISATION

Un autre élément à considérer lorsque l'on souhaite analyser le volume de recherche d'un mot-clé titre est ce que l'on pourrait caractériser comme étant son indépendance du contexte télévisuel.
En effet, la portion de trafic qui nous intéresse pour un mot-clé/titre donné est celle associée à l'intérêt pour le programme TV en question, et non le reste du trafic indépendant de ce contexte.

Ainsi, il est possible de distinguer deux types de mots-clés, ceux que l'on pourrait qualifier d'« indépendants » et ceux qui seraient « non-indépendants ».

A titre d'exemple, un mot-clé indépendant serait :
« 100 jours avec les gendarmes de Bourgogne ».

Dans le cadre de cette requête, on peut déterminer aisément que la proportion de trafic internet associée à ce mot-clé est proche des 100% pour l'émission tv. En effet, aucun utilisateur n'est susceptible de rechercher ces termes en dehors du cadre de l'émission TV.

Dans le cas opposé, avec l'émission intitulée « H », ou « maison à vendre », les utilisateurs sont susceptibles de rechercher ces termes sans les associer aux émissions tv.
La difficulté réside donc dans le fait de déterminer la proportion de trafic associée aux programmes TV pour ces termes de recherche.

La labellisation des requêtes est manuelle, basée sur une certaine « expertise » et difficilement automatisable, car des outils de traitement du langage avancés seraient nécessaires pour interpréter les termes, mais il serait également difficile de se procurer les résultats indiquant les proportions associées aux idées et concepts différents représentés par un même mot-clé.

Ainsi, lors de la première exécution du programme, le volume de recherche total pour environ 700 termes était de 31 millions. En retirant les requêtes qualifiées de non-indépendantes, le total s'élevait à 18 millions. L'ordre de grandeur reste ainsi similaire, mais on constate l'importance de prendre en compte cet aspect lié au contexte d'un mot-clé.

Parmi ces requêtes, la plus grande quantité de trafic « imprécis » concernait des mots-clés dont les titres étaient des prénoms, « Sam », « Claude », pour lesquels il est très difficile de quantifier la part de volume de recherche associée à ces émissions TV uniquement, parmi toutes les recherches effectuées pour ces noms.
Il serait donc peu pertinent de considérer 100% du trafic calculé, pour ces termes qui ont une forte dépendance contextuelle.

PARTIE 3

AMELIORATION

La problématique des mots-clés de référence pourrait se résumer ainsi :
Comment estimer le volume de recherches des titres afin de les associer à un mot-clé de référence qui soit dans un ordre de grandeur d'environ 10x leur volume de recherche, au maximum ?

SELECTION DES MOTS-CLES PAR EXPERTISE UTILISATEUR ET SCORING/NOTATION

La sélection des mots-clés de référence est l'étape la plus importante du processus étudié dans cette partie. En effet, le calcul des valeurs de trafic des mots-clés titres est explicitement basé sur les valeurs des mots-clés de référence. Il est donc important de sélectionner uniquement ceux dont l'incertitude est minimale.
En effectuant une première vérification préalable, on peut constater une différence non-négligeable entre les valeurs obtenues sur le site Ahrefs et les courbes équivalentes attendues sur Google Trends. Exemple :

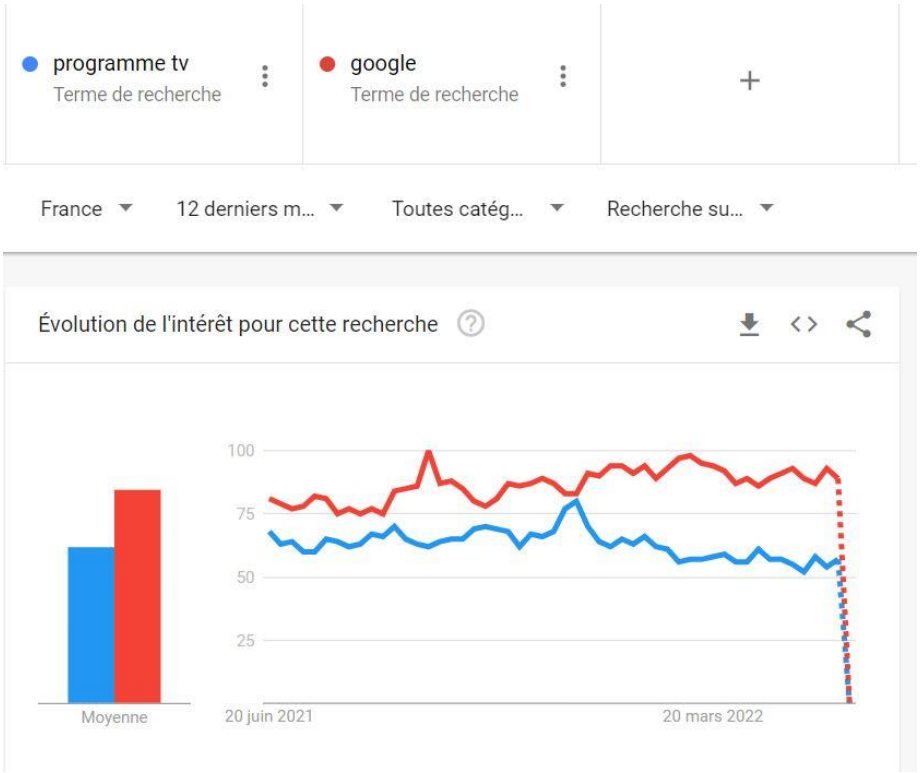

Figure III.I

Dans ce premier exemple, on constate un intérêt plus prononcé pour le mot-clé « google » que le mot-clé « programme tv », or selon les données de trafic renseignées par l'outil Ahrefs, nous obtenons : 18 Millions de recherches mensuelles pour le mot-clé google, et 34 Millions de recherches mensuelles pour le mot-clé « programme tv ». On constate donc ici que pour un écart théorique de 16 Millions de requêtes, on obtient un résultat qui ne correspond pas à la comparaison effective retournée par Google Trends.

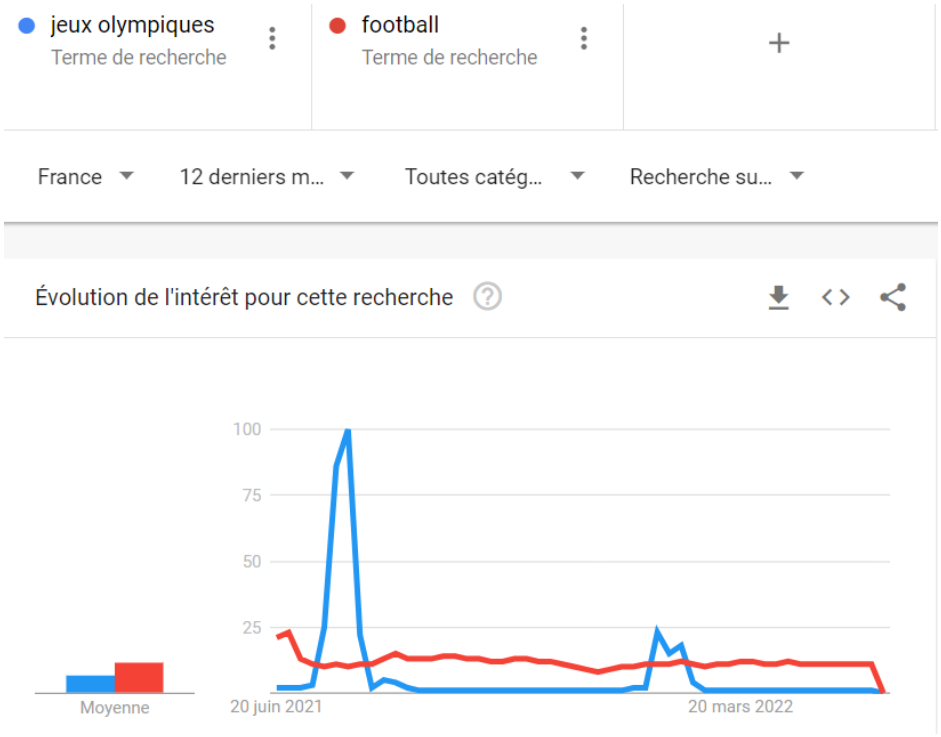

Figure III.II

Autre exemple flagrant sur la figure ci-dessus avec la comparaison des mots-clés : 'football' et 'jeux olympiques'.
On constate une popularité plus importante pour le mot-clé « football » sur la grande majorité de l'année, en comparaison du mot-clé « jeux olympiques ». En revanche, selon les valeurs fournies par Ahrefs, le mot-clé 'jeux-olympiques' génère un volume de recherche de 313 000 contre 223 000 pour le mot-clé « football ».
On constate donc de nouveau une différence non-négligeable entre les données fournies par Ahrefs et celles retournées par Google Trends.

La liste de mots-clés de référence constituée initialement s'est avérée être imprécise, certaines valeurs des mots-clés invalides étant mentionnées ci-dessous.

Valeurs théoriques :

Keyword	Number of Searches (Ahref)
journal télé	2000
journal télévisé	3100
flex css	4400
jt	5400

Figure III.III

Valeurs réelles :

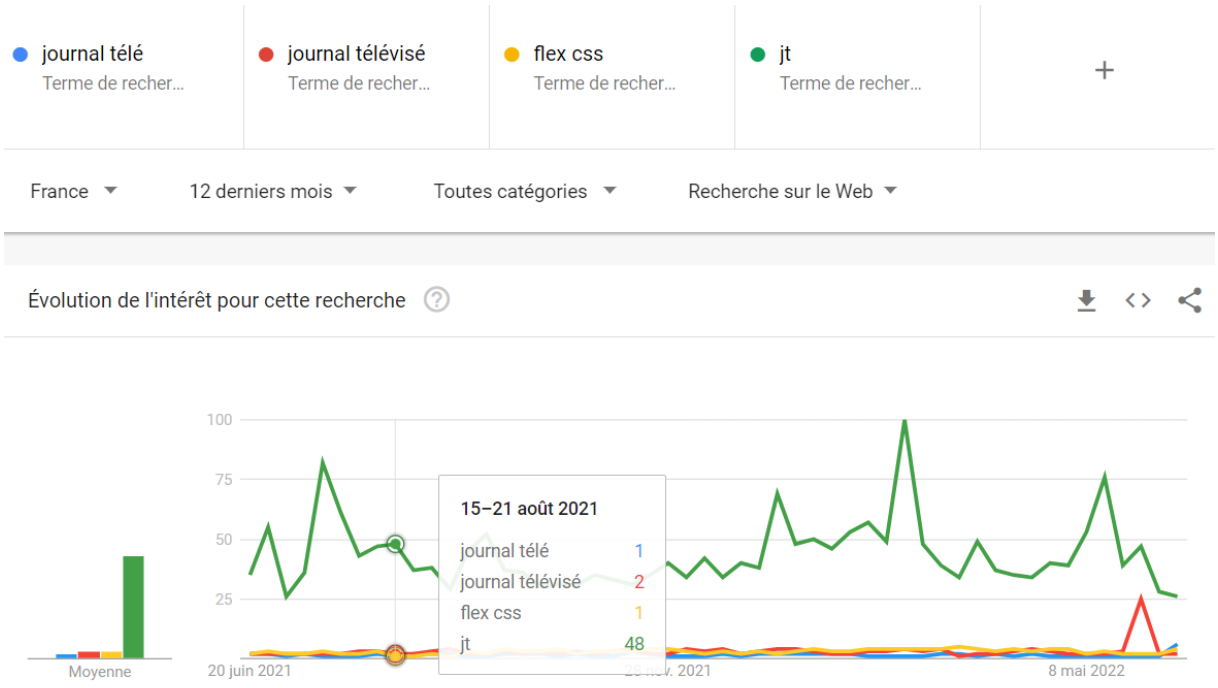

Figure III.IV

Les courbes fournies par Google Trends ne correspondent pas aux valeurs ci-dessus, retournées par Ahrefs.

PRÉSENTATION DU PRINCIPE DE VALIDATION DES MOTS-CLÉS

La validation des mots-clés de référence est une étape importante dans la diminution de l'incertitude liée aux calculs, ce qui a été mis en évidence précédemment.
En effet, la source d'erreur sur le trafic des mots-clés de base étant assez difficilement quantifiable avec Ahrefs et Wordstream, il est possible d'améliorer cette base en utilisant Google Trends, qui procure toujours une valeur juste.

Le processus consiste à trouver des mots-clés, de courte traîne si possible, car potentiellement plus précis, qui génèrent un volume quantitatif connu de requêtes (ex : 38K/jour), et en sélectionner plusieurs pour constituer une base de mots-clés de référence. Le but étant de constituer des « paliers » de volume de recherche, correspondant à environ 10x maximum le nombre de requêtes attendues du/des mots-clés auxquels il sera associé.
On souhaite limiter le ratio à 10x pour éviter d'avoir des nombres proches de 0 ou 1 dans Google Trends qui risqueraient de ne pas être représentatifs car d'une échelle trop inférieure.

Il est assez compliqué de trouver des mots-clés de références qui conservent une certaine cohérence dans les données entre les 3 plateformes utilisées, WordStream, Ahrefs, et Google Trends.
En effet, il suffit parfois d'un accent ajouté ou manquant pour obtenir un mot-clé de référence qui soit éligible, ou non.

Du fait des différences que l'on peut retrouver en raison des accents dans la langue française, il est préférable de sélectionner des mots-clés de référence qui ne sont pas sujets à cette ambiguïté.
Avec l'exemple ci-dessous en figure III.V, on pourrait penser que le mot-clé sans accent intègre, dans une certaine mesure, le volume de recherche du mot-clé « avec » accent, en raison de l'évolution de la courbe presque identique, et qui semble simplement présenter un « décalage » positif.
L'opacité dans la présentation des résultats et l'interprétation des recherches effectuées par Google ne permet qu'une analyse partielle et sujette à un bruit plus ou moins conséquent.

De plus, le site Ahrefs permet d'avoir des valeurs de trafic, avec ou sans accent, lorsque Wordstream confond les deux recherches (on ne sait pas si le résultat est donc l'agrégation du mot-clé avec ET sans accent).
Il est donc nécessaire de bien choisir ses mots-clés de référence, pour éviter une double erreur potentielle, au niveau de la valeur de calcul de trafic entre Ahrefs et Wordstream, et dans la comparaison avec les valeurs de Google Trends.

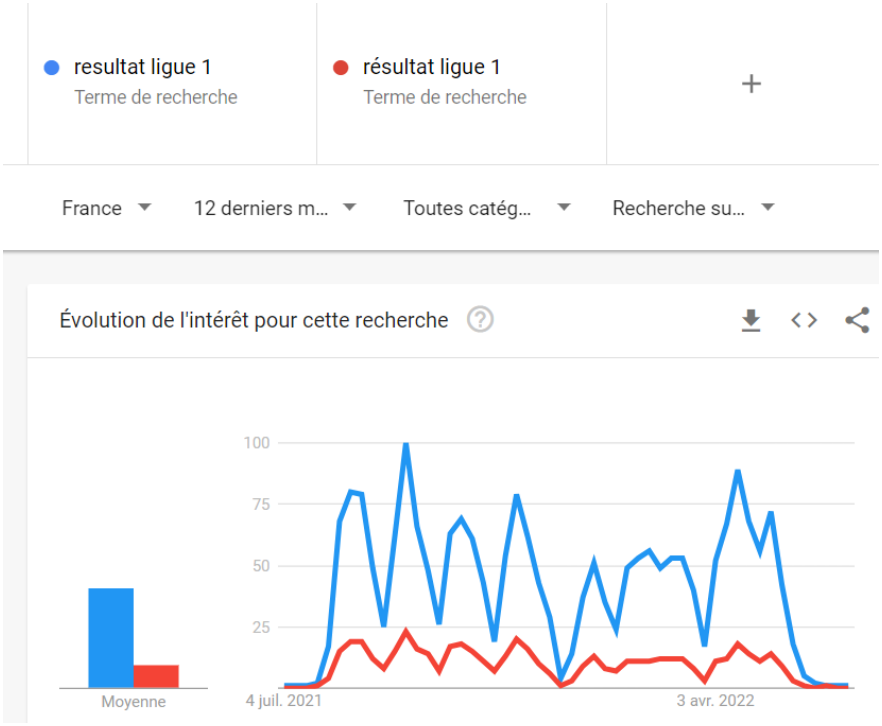

Figure III.V

Historiquement, la gestion et l'interprétation des mots-clés par le moteur de recherche Google ont été modifiées à plusieurs reprises.

En ayant une certaine expérience dans le domaine des mots-clés, il est possible d'éviter certains critères qui nuiraient à la justesse de la base de mots-clés de référence dans le temps.

En effet, dans le but de garder une certaine stabilité dans les résultats, il est préférable d'éviter de cibler des mots-clés qui contiennent une date non-significative, mais qui va grandement faire varier le trafic sur la période qui nous intéresse, par exemple eurovision 2021 ou 2022, le résultat va être grandement différent d'une année sur l'autre, ce que l'on retrouvera dans une certaine mesure dans le mot-clé « base » eurovision, ce que l'on peut imputer au caractère cyclique de l'événement lié au mot-clé.
Il pourrait être également pertinent de ne pas sélectionner de noms d'acteurs, de franchises, qui peuvent redevenir populaires à l'occasion d'un nouveau film ou d'un nouvel opus.
Tout mot-clé caractérisant un évènement « cyclique », même de façon irrégulière est à éviter afin de constituer une base de mots-clés de référence saine et plus stable dans le temps.

Figure III.VI

Figure III.VII

On observe sans surprise une plus grande stabilité dans la recherche web de « carte anniversaire gratuite » que dans la recherche « Tom Cruise », ici largement influencée par le nouveau film « Top Gun », dont il en est l'acteur principal.
On retrouve également de grandes variations de popularité pour un grand nombre de mots dont le volume de recherche dépend des saisons, des fêtes, des évènements récurrents, pour ne citer que certains critères.
Le choix de mots-clés peut sembler anodin et relativement simple, mais il nécessite en réalité une recherche approfondie.

En perspective d'ouverture non abordée dans ce stage, peut-être serait-il possible d'évaluer et de comparer l'évolution du niveau d'écriture au sein d'une population en comparant la proportion de requêtes qui conservent les éléments de littérature tels que les accents et les traits d'union, (cas 1) et des requêtes qui ne respectent ni la casse, ni l'orthographe de manière stricte (cas 2).
Une augmentation du cas 2 et une baisse du cas 1 pourrait indiquer une diminution du niveau en orthographe de la population sur une période donnée, et inversement.
Les messageries instantanées sont également grandement responsables des mauvaises habitudes d'orthographe prises par les utilisateurs, et cet élément est à prendre en compte lors de la réalisation d'un projet orienté SEO, tel qu'une campagne marketing.

METHODE DE SCORING :

Un score est assigné à chaque mot-clé de référence, il permet de vérifier la correspondance des relations de proportionnalité entre ces mots-clés. Chaque mot-clé de référence est comparé aux 4 mots-clés de popularité directement supérieure, et 4 mots-clés de popularité directement inférieure, ce qui permet de rester dans un ordre de grandeur similaire pour effectuer cette comparaison. En effet, il n'est pas pertinent d'observer la relation de proportionnalité entre « Facebook » et « petit chien roux », lorsque le facteur de multiplication de trafic est de 5000+ entre les deux mots. L'exploitation graphique ne serait pas valide sur Google Trends.

Ainsi, si Google Trends nous indique que le mot-clé « Facebook » est 1.5 fois plus populaire que le mot-clé « Youtube », il faut également que cette relation soit respectée avec les valeurs Ahrefs/Wordstream entre ces mêmes mots, afin d'être certain d'utiliser des valeurs de référence qui soient pertinentes et justes. Dans cet exemple, le score s'approche de 1 si l'on obtient 30 Millions en volume de recherches pour « Facebook », et 20 Millions pour « Youtube », ce qui conserve alors cette relation de proportionnalité de 1.5.
Une comparaison entre la relation de proportionnalité théorique (Ahrefs/Wordstream) et la relation de proportionnalité réelle (Google Trends) permet d'attribuer ainsi un score à chaque mot-clé.

Matrice des relations de proportionnalités réelles sur Google Trends (tronquée) :

	facebook	youtube	programme	gmail	google	Paypal	tiktok	playstation 5	tf1	zidane	doctolib moi	recette quicl	avengers en	mini frigo	carte annive	lightroom	messenger s	peluche disr	uefa tv
facebook	1.0	1.133	1.204	2.616	0.875	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
youtube	Na	1.0	1.064	2.313	0.773	19.293	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
programme	Na	Na	1.0	2.174	0.727	18.133	14.918	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
gmail	Na	Na	Na	1.0	0.334	8.343	6.864	188.75	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
google	Na	Na	Na	Na	1.0	24.95	20.527	564.5	10.961	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
Paypal	Na	Na	Na	Na	Na	1.0	0.828	9.745	0.438	3.884	Na	Na	Na	Na	Na	Na	Na	Na	Na
tiktok	Na	Na	Na	Na	Na	Na	1.0	11.764	0.529	4.688	11.835	Na	Na	Na	Na	Na	Na	Na	Na
playstation 5	Na	Na	Na	Na	Na	Na	Na	1.0	0.045	0.399	1.006	1.634	Na	Na	Na	Na	Na	Na	Na
tf1	Na	Na	Na	Na	Na	Na	Na	Na	1.0	8.862	22.372	36.327	70.558	Na	Na	Na	Na	Na	Na
zidane	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	2.531	4.177	7.375	10.727	Na	Na	Na	Na	Na
doctolib moi	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	1.619	2.979	4.224	2.743	Na	Na	Na	Na
recette quicl	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	1.847	2.621	1.694	1.549	Na	Na	Na
avengers en	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	1.419	0.917	0.839	4.552	Na	Na
mini frigo	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	0.647	0.591	3.215	1.9	Na
carte annive	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	0.913	4.969	2.937	11.683
lightroom	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	5.443	3.217	12.796
messenger s	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	0.587	2.362
peluche disr	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0	4.022
uefa tv	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	1.0
gamelle pou	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
huile moteu	0.0	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
voiture 4x4 c	0.0	0.0	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
pneu lisse	0.0	0.0	0.0	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na
petit chien r	0.0	0.0	0.0	0.0	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na

Figure III.VIII

Matrice des relations de proportionnalités théoriques sur Ahrefs/Wordstream (tronquée) :

	facebook	youtube	programme	gmail	google	Paypal	tiktok	playstation 5	tf1	zidane	doctolib mo	recette quic	avengers en	mini frigo	carte annive	lightroom	messenger s	peluche disr	uefa tv
facebook	1.0	1.162	1.412	2.098	2.617	23.157	35.626	56.941	74.888	182.727	237.028	476.303	601.796	1148.571	1313.725	2451.22	5075.758	6611.842	7389.706
youtube	0.861	1.0	1.215	1.806	2.253	19.931	30.663	49.008	64.456	157.273	204.009	409.953	517.964	988.571	1130.719	2109.756	4368.687	5690.789	6360.294
programme t	0.708	0.823	1.0	1.486	1.854	16.406	25.239	40.34	53.055	129.455	167.925	337.441	426.347	813.714	930.719	1736.585	3595.96	4684.211	5235.294
gmail	0.477	0.554	0.673	1.0	1.247	11.037	16.98	27.139	35.693	87.091	112.972	227.014	286.826	547.429	626.144	1168.293	2419.192	3151.316	3522.059
google	0.382	0.444	0.539	0.802	1.0	8.848	13.612	21.756	28.614	69.818	90.566	181.991	229.94	438.857	501.961	936.585	1939.394	2526.316	2823.529
Paypal	0.043	0.05	0.061	0.091	0.113	1.0	1.538	2.459	3.234	7.891	10.236	20.569	25.988	49.6	56.732	105.854	219.192	285.526	319.118
tiktok	0.028	0.033	0.04	0.059	0.073	0.65	1.0	1.598	2.102	5.129	6.653	13.37	16.892	32.24	36.876	68.805	142.475	185.592	207.426
playstation 5	0.018	0.02	0.025	0.037	0.046	0.407	0.626	1.0	1.315	3.209	4.163	8.365	10.569	20.171	23.072	43.049	89.141	116.118	129.779
tf1	0.013	0.016	0.019	0.028	0.035	0.309	0.476	0.76	1.0	2.44	3.165	6.36	8.036	15.337	17.542	32.732	67.778	88.289	98.676
zidane	0.005	0.006	0.008	0.011	0.014	0.127	0.195	0.312	0.41	1.0	1.297	2.607	3.293	6.286	7.19	13.415	27.778	36.184	40.441
doctolib mo	0.004	0.005	0.006	0.009	0.011	0.098	0.15	0.24	0.316	0.771	1.0	2.009	2.539	4.846	5.542	10.341	21.414	27.895	31.176
recette quic	0.002	0.002	0.003	0.004	0.005	0.049	0.075	0.12	0.157	0.384	0.498	1.0	1.263	2.411	2.758	5.146	10.657	13.882	15.515
avengers en	0.002	0.002	0.002	0.003	0.004	0.038	0.059	0.095	0.124	0.304	0.394	0.791	1.0	1.909	2.183	4.073	8.434	10.987	12.279
mini frigo	0.001	0.001	0.001	0.002	0.002	0.02	0.031	0.05	0.065	0.159	0.206	0.415	0.524	1.0	1.144	2.134	4.419	5.757	6.434
carte annive	0.001	0.001	0.001	0.002	0.002	0.018	0.027	0.043	0.057	0.139	0.18	0.363	0.458	0.874	1.0	1.866	3.864	5.033	5.625
lightroom	0.0	0.0	0.001	0.001	0.001	0.009	0.015	0.023	0.031	0.075	0.097	0.194	0.246	0.469	0.536	1.0	2.071	2.697	3.015
messenger s	0.0	0.0	0.0	0.0	0.001	0.005	0.007	0.011	0.015	0.036	0.047	0.094	0.119	0.226	0.259	0.483	1.0	1.303	1.456
peluche disr	0.0	0.0	0.0	0.0	0.0	0.004	0.005	0.009	0.011	0.028	0.036	0.072	0.091	0.174	0.199	0.371	0.768	1.0	1.118
uefa tv	0.0	0.0	0.0	0.0	0.0	0.003	0.005	0.008	0.01	0.025	0.032	0.064	0.081	0.155	0.178	0.332	0.687	0.895	1.0
gamelle pou	0.0	0.0	0.0	0.0	0.0	0.002	0.002	0.004	0.005	0.013	0.016	0.033	0.041	0.079	0.09	0.168	0.348	0.454	0.507
huile moteu	0.0	0.0	0.0	0.0	0.0	0.001	0.002	0.003	0.004	0.01	0.013	0.027	0.034	0.064	0.073	0.137	0.283	0.368	0.412
voiture 4x4 c	0.0	0.0	0.0	0.0	0.0	0.001	0.001	0.002	0.002	0.005	0.007	0.014	0.018	0.034	0.039	0.073	0.152	0.197	0.221
pneu lisse	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.001	0.001	0.003	0.004	0.009	0.011	0.022	0.025	0.046	0.096	0.125	0.14
petit chien r	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.002	0.002	0.004	0.005	0.009	0.011	0.02	0.042	0.055	0.061

Figure III.IX

La matrice de scoring de la liste de mots-clés de référence utilisés ainsi obtenue est la suivante :

Keyword	Fiability Ratio
facebook	0,822
youtube	0,833
programme tv	0,792
gmail	0,804
google	0,502
Paypal	0,587
tiktok	0,663
playstation 5	0,208
tf1	0,179
zidane	0,575
doctolib mon compte	0,772
recette quiche lorraine	0,605
avengers endgame	0,711
mini frigo	0,63
carte anniversaire gratuite	0,629
lightroom	0,486
messenger sans facebook	0,463
peluche disney	0,53
uefa tv	0,282
gamelle pour chien	0,513
huile moteur essence	0,208
voiture 4x4 occasion	0,5
pneu lisse	0,788
petit chien roux	0,7

Figure III.X

Les mots en rouge sont peu fiables et doivent donc être retirés de la liste, afin de les remplacer par des mots-clés recevant un bon score de fiabilité.

La technique de notation utilisée pourrait être améliorée dans le cadre d'un travail annexe.

CORRECTION DES MOTS-CLÉS

La correction des valeurs de trafic pour les mots-clés de référence survient après la sélection de mots-clés efficaces et correspondants aux différents critères de « qualité » évoqués précédemment.
Cette étape complémentaire correspond à un « lissage » des valeurs de références.
En effet, la valeur du premier mot « Facebook » qui semble être précise, étant donné son score et son aspect populaire et de courte traîne, va être utilisée comme référence pour la correction des valeurs de trafic des autres mots-clés de référence.

En effet, les mots-clés vont être comparés entre eux sur Google Trends 'en cascade', en ajustant la valeur d'un mot par rapport à la nouvelle valeur corrigée du mot directement supérieur dans la liste.
Plus la valeur du mot-clé initial sera juste, ici « Facebook », plus la correction des valeurs sera précise.
En revanche, lorsque les étapes de sélection des mots-clés précédentes ont au préalable été appliquées scrupuleusement, cette correction des valeurs n'en devient que peu significative, les valeurs de départ étant ainsi plus justes.

PARTIE 4

EXPERIMENTATION

PRÉSENTATION DES BASES

Les différentes bases contenant les titres d'émissions TV, faisant l'objet d'une analyse complète étaient les suivantes :

- Base numéro 1, issue d'une capture de la station TV du 17/12/21 au 08/07/22, soit 204 jours
- Base numéro 2, issue d'une capture de la station TV du 07/04/22 au 27/06/22, soit 82 jours

La première liste contient des titres des programmes de la base numéro 1 pour 30 chaînes (TNT) avec plus de 300 mots de description par programme, et la seconde liste contient les titres des programmes de la base numéro 2 avec 300 chaînes, et plus de 300 mots dans la description de chaque programme.

Ainsi, on peut vérifier l'application du modèle longue traîne en observant les résultats obtenus :

LISTE 1 :

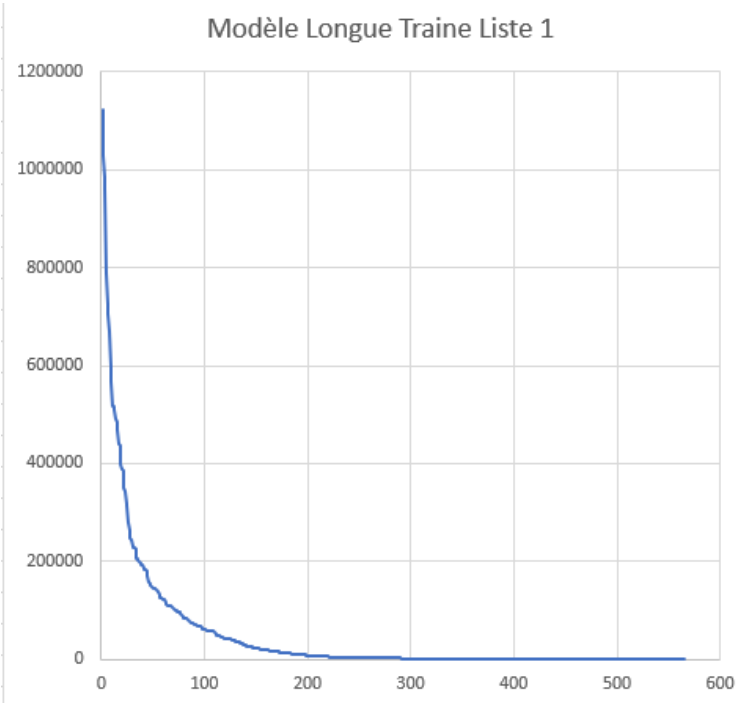

Figure IV.I

Dans cette première liste initialement constituée de 810 titres, 244 ont été supprimés, étant soit des doublons, soit des mots-clés ne générant pas assez de trafic selon Google Trends.
Cela représente environ 30 % des titres.
En observant le graphique, on reconnaît bien le modèle longue-traîne.
Le volume total de recherches associées aux mots-clés/titres de cette base est d'un peu plus de 27 Millions.

LISTE 2 :

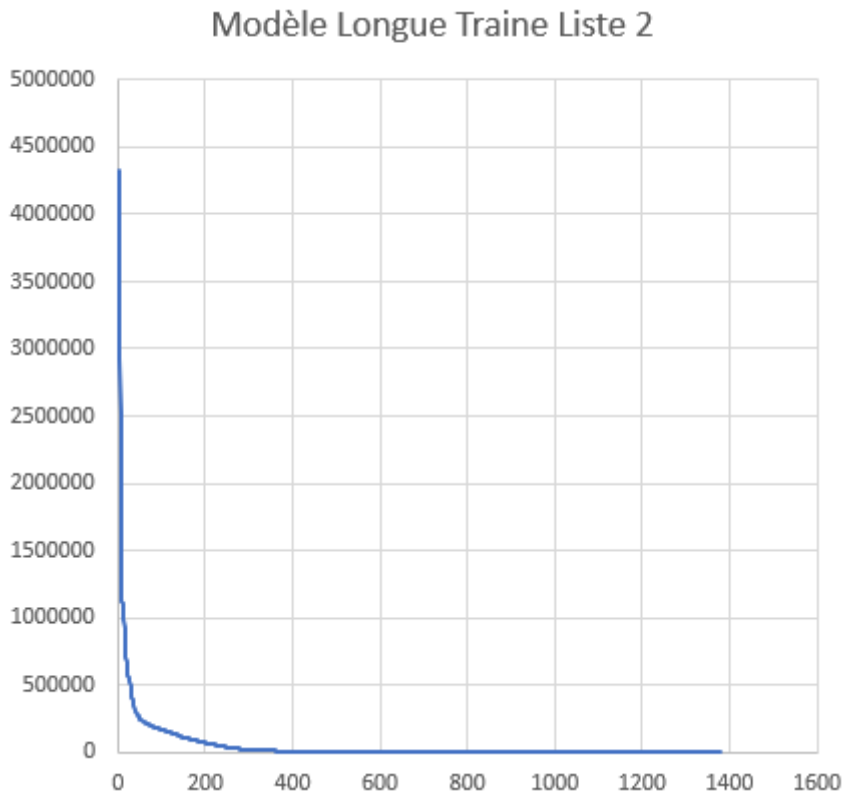

Figure IV.II

Dans cette seconde liste initialement constituée de 2920 titres, 1542 ont été supprimés, étant soit des doublons, soit des mots-clés ne générant pas assez de trafic selon Google Trends.
Cela représente environ 52 % des titres.
En observant le graphique, on reconnaît encore une fois le modèle longue-traine.
Le volume total de recherches associées aux mots-clés/titres de cette base est d'un peu plus de 72 Millions.

Ces modèles ne prennent pas en compte la labellisation des requêtes « indépendantes » et « non indépendantes » du contexte, évoquées précédemment, puisqu'elle représenterait un travail manuel trop conséquent pour une base contenant des milliers de titres.
Dans le cadre d'un travail de SEO, notamment pour un site web, on constate qu'il est intéressant de cibler une multitude de mots-clés, qui, agrégés, génèrent un volume de recherche conséquent.
Dans ce cas d'application, cela se traduirait par le fait de tenter de référencer un site web lié à un programme TV en ciblant une multitude de titres d'émissions, plutôt que de tenter d'être référencé uniquement avec des mots-clés de courte traîne très compétitifs tels que le mot-clé « programme TV », qui génère seul plus de 30 Millions de recherches mensuelles en France.

CONCLUSION

Ce stage sur la thématique de la SEO a été réellement épanouissant. Etant amené à travailler quotidiennement sur le référencement web/Youtube dans ma vie professionnelle, c'est un sujet que j'apprécie particulièrement.
Je pense que le processus de traitement développé au cours du stage est relativement complet, et l'ensemble des axes d'étude a été abordé, notamment avec la partie de sélection, validation, et correction des mots-clés.
Certaines méthodes de scoring/notation pourraient être retravaillées et améliorées, mais seule la justesse et la précision s'en trouveraient affinées.
J'ai pu acquérir des compétences en langage de programmation Python tout en découvrant l'aspect « recherche » de l'Informatique.
Je souhaite de nouveau remercier mon maître de stage Mathieu Delalandre pour son investissement et nos échanges critiques quotidiens qui ont permis de réaliser avec succès ce projet.

Services d'Analytics SEO TV

Résumé :

La Station TV du LIFAT de Tours traite différentes thématiques autour de la capture et analyse de flux TV.
Ces thématiques sont constituées principalement de traitements techniques de « bas niveau » informatique, et la problématique SEO s'inscrit dans une démarche annexe qui permet d'avoir une analyse de « haut niveau » sur certaines données récupérées par ces captures. L'outil développé ici est appliqué au traitement de ces données (titres), mais peut être réutilisé dans tout secteur souhaitant porter une analyse de popularité sur ses contenus et améliorer son référencement sur le Web.
L'objectif de ce stage est de déterminer le volume de recherche associé à des mots-clés sur Google. L'accès public à ces données sur internet est très restreint. En effet, Google, dominant majoritairement le marché des moteurs de recherche, ne permet que d'accéder à ces informations en payant pour leurs services publicitaires tels qu'AdWords. En utilisant le service Google Trends fournissant des données précises et relatives pour des mots-clés spécifiés (pourcentages compris entre 0 et 100), nous allons calculer les valeurs de trafic absolues et volumes de recherche, à l'aide d'outils SEO gratuits qui nous fourniront des données de référence.

Mots-clés :
Longue-Traine
Courte-Traine
SEO (Search Engine Optimization)
Mot-clé
Difficulté
Trafic / Search Volume
Google Trends

Abstract :
5 to 15 lines to provide an overview of your internship work

The TV Station from the LIFAT of Tours deals with several fields revolving around data capture and TV streams analysis. These subjects are mainly based on low-level technical processing, and the SEO issue arises with a high-level approach and analysis of data gathered from the different captures. The tool we developed during this internship is applied to the treatment of this data but can be used in any field where the SEO challenge would require content popularity analysis to achieve a better Web referencing.
The purpose of this work placement is to calculate search volume for specified keywords on Google. The public access to this data on the Internet is very limited. Indeed, Google owns most of the Search Engine market and thus only provides this data through some of their paid tools, such as Google AdWords.
Using the Google Trends website, which provides accurate relative values for specific keywords (percentages ranging from 0 to 100), we manage to calculate search traffic absolute values using free SEO tools, that will provide us with data used as a reference.

Keywords :
Long-tail
Short-tail
SEO (Search Engine Optimization)
Keyword
Difficulty
Traffic / Search Volume
Google Trends