

Rapport de Projet SI : **Projet SEO Compatibilité**

Encadrant :

Mathieu Delalandre

Étudiants :

ROUSSEAU Yohan

LAFONT Paul

PONS Maël

ROUSSEAU Joshua

SOURDET Jeremy

ANDRIOT Axel

Sommaire

Contexte du projet	3
Caractérisation de la démarche SEO	3
Présentation du projet	4
La répartition des tâches	5
Travail effectué	6
Agrégation des données	6
Extraction des mots clés	8
Interrogation Google Trends	12
Résultats - Analyse des données Google Trends	15
Axes d'améliorations - Conclusion	18

Contexte du projet

Caractérisation de la démarche SEO

Afin de se positionner sur le web, une simple démarche d'essai et d'analyse, efficace dans les premières années d'Internet, n'est plus suffisante. Aujourd'hui, une démarche SEO (signifiant Search Engine Optimization, soit Optimisation pour les Moteurs de Recherche en français) est nécessaire pour réussir. Cette démarche se définit en plusieurs étapes :

- **Définition du contexte de création** : l'entreprise, ou la personne, effectuant la démarche doit commencer par identifier clairement son audience cible à l'aide d'objectifs.
- **Identification des intentions de recherche** : l'idée que l'audience cible s'adaptera à ce que l'émetteur du site voudra partager est fautive. Internet est vaste et les utilisateurs ont déjà ce qu'il veulent dedans, se positionner sur le web nécessite alors de trouver ce que l'audience veut pour pouvoir s'adapter et l'attirer
- **Recherche de mots-clés** : pour satisfaire la partie précédente, il faut ensuite trouver les bons mots-clés à utiliser dans la publication pour que l'audience soit redirigée facilement vers le site.
- **Rédaction du contenu** : L'audience étant ciblée et les mots à utiliser aussi, il faut maintenant rédiger l'article, ou le site, selon les objectifs fixés dans la première partie.
- **Accessibilité du site aux moteurs de recherches** : Une page sans interaction ne sera pas mise en avant par les moteurs de recherches et l'audience ne pourra donc pas y accéder facilement. En insérant l'article, la page internet, dans un site ou en créant un site avec plusieurs redirections vers d'autres pages, les moteurs de recherches valoriseront cette nouvelle entrée sur le web en la mettant en avant dans les recherches.
- **Mesure des résultats** : Cette étape permet d'améliorer ou de conserver une bonne visibilité. En étudiant, grâce à divers outils, les résultats que génère la publication, l'entreprise pourra se repositionner et améliorer la qualité de ses prochaines publications.

Présentation du projet

Pour le contexte du projet, une station TV est installée à Polytech et enregistre les émissions télévisuelles passant sur diverses chaînes pour créer une base de données de descriptions d'émissions. Quatre utilisations de cette station TV sont à différencier :

- “Scrapping” de programmes télévisuelles : repérage des émissions en cours de diffusion avec association de descriptions de l'émission
- Reconnaissance automatique de contenu : analyses des images pour reconnaître ce qui est affiché à l'écran
- Capture intelligente des images TV : association en temps réel de mots-clés aux images diffusées
- Vérification de faits : détection des propos diffusés et recherche parallèle pour déterminer la véracité des propos

Notre partie se situe donc sur la partie “Scrapping” et pour cela, nous avons accès à 4 mois de données TV, composées d'émissions TV et de descriptions de chacune d'elles.

Le but de ce projet est de déterminer la compatibilité SEO de ces données dans un but de publication sur Internet d'articles liés aux émissions télévisuelles. Ce projet a donc un but expérimental pour déterminer une potentielle utilisation future. Les différents objectifs sont :

- Agrégation de résumés
- Extraction automatique de requêtes
- Extraction des données de trafic sur la plateforme Google Trends
- Caractérisation modèle longue traîne
- Extraction des métriques d'autorité

La répartition des tâches

Dans le périmètre qui nous a été défini au début du projet, nos missions pour découvrir les outils de compatibilité SEO s'articulent autour de trois axes. Sur chacune des tâches deux étudiants ont travaillé en coopération pour éviter le manque d'activité ou la surcharge sur une partie.

Le premier axe est le pré-traitement des données à analyser : le but initial était de traiter une base de données contenant des descriptions d'épisodes TV. Cette base était fournie par notre encadrant projet et est le résultat du travail de stagiaires de l'école.

Ces documents csv étaient parfois mal triés, des informations étaient redondantes et les descriptions souhaitées pour études étaient dispersées dans la base.

Il nous fallait alors agréger les données des différents épisodes d'une même série pour condenser les textes des descriptions à étudier. Ce sont Axel ANDRIOT et JérémY SOURDET qui se sont occupés de cette tâche du projet.

Le deuxième axe était basé sur l'étude du contenu de ces bases de données ainsi que sur l'extraction de mots-clés. Déterminer les paramètres d'extraction, le ou les algorithmes à utiliser et définir les paramètres à utiliser sur les textes étudiés furent les enjeux de cette partie. Pour cette seconde partie, ce sont Paul LAFONT et Maël PONS qui se sont chargés de travailler dessus.

Le troisième axe étudié sur ce projet était l'interrogation de Google Trends et l'automatisation des requêtes. Cette partie était la plus déterminante et nous permettait d'obtenir les rapports qui serviraient de résultats d'analyse. Enfin ce troisième axe a été développé par Yohan ROUSSEAU et Joshua ROUSSEAU.

Travail effectué

Agrégation des données

La première partie de ce projet est d'agréger les données d'entrées. Pour cela, nous nous sommes inspirés du stage de Pierre-Louis CHAN qui avait extrait des descriptions d'émissions télévisées. Nous avons alors utilisé son fichier de retour pour agréger les données car certaines émissions (comme les séries télévisées) étaient répliquées (il y avait une ligne par épisode et il nous fallait une ligne par série) ; de plus, certaines émissions n'avaient pas de description.

C192.api.télé	Et si on se réinventait	[' Magazine ']	Ce programme vous emmène à l		
C192.api.télé	Météo	[' Météo ']	None		
C192.api.télé	Le doudou	[' Film ']	Michel a perdu le doudou de sa t		
C192.api.télé	Esprits criminels	[' Série ']	A deux reprises, un inconnu a tu		
C192.api.télé	Esprits criminels	[' Série ']	Billy Flynn a encore sévi en kidn		
C192.api.télé	Esprits criminels	[' Série ']	Plusieurs corps ont été retrouvé		
C192.api.télé	Les experts : Miami	[' Série ']	Un futur marié ne se présente p		
C192.api.télé	Les experts : Miami	[' Série ']	Trois joueurs de volley-ball meu		
C192.api.télé	Programmes de la nu	[' Programmi	None		

Ainsi, nous avons commencé par regrouper les descriptions multiples. Nous avons alors récupéré la classification de Pierre-Louis de description générale (la description de l'émission dans sa généralité) et description spécifique (la description de l'épisode de l'émission à un instant donné) pour regrouper dans une description la description générale puis toutes les descriptions spécifiques. Comme certaines émissions ont des descriptions similaires, nous avons fait en sorte d'éliminer les doublons dans les descriptions. Cela nous permet de n'avoir qu'une seule fois la description générale ainsi que toutes les descriptions spécifiques (qui sont donc uniques) dans la description finale.

Ensuite, pour qu'un texte soit SEO compatible, il faut qu'il comporte entre 300 et 3000 mots. En effet, hors de ces bornes, la recherche de mot-clé pourrait ne pas être pertinente. Nous avons alors automatisé le compte des mots de chaque descriptions pour ne garder que les textes SEO compatibles. Grâce à cela, les émissions sans description n'apparaissent plus dans le fichier de retour, tout comme les émissions ayant une description trop longue ou trop courte car cela ne sera pas utile pour la suite.

	A	B	C	D	E	F
	Channel ID	Title	Categories	Description	Taille en nombre de mot	
2	C192.api.tele	Esprits criminels	[' Série ']	A deux repri	706	
3	C192.api.tele	Les experts : Miami	[' Série ']	Un futur mar	399	
4	C192.api.tele	Ici tout commence	[' Série ']	Saisie par la	996	
5	C192.api.tele	Demain nous apparti	[' Série ']	La police de	951	
5	C192.api.tele	Les feux de l'amour	[' Feuilleton	Profitant d'u	904	
7	C192.api.tele	Je te promets	[' Série ']	Florence et F	615	
3	C192.api.tele	New York Unité Spéc	[' Série ']	Comme Crag	1845	
3	C192.api.tele	Chicago Police Depar	[' Série ']	Des agents p	313	
0	C192.api.tele	New York, section cri	[' Série ']	Une série de	445	
1	C4.api.telera	Pays et marchés du n	[' Document	Xavier Petit	548	
2	C4.api.telera	Les petits meurtres d	[' Série ']	""Deux jour	325	
3	C4.api.telera	Les aventures du jeu	[' Série ']	""Roturier c	336	
4	C4.api.telera	Speakerine	[' Série ']	Juillet 1962.	596	
5	C4.api.telera	Une planète parfaite	[' Document	Dans l'univer	274	

Étant donné qu'il n'y avait seulement 50 articles SEO compatibles, M. Delalandre nous a transmis les fichiers XML du 17 Décembre 2021 au 10 Janvier 2022 afin d'extraire plus d'articles.

Pour cela nous avons repris la partie parser du code de Pierre-Louis, que nous avons modifié afin de pouvoir avoir un fichier CSV du même type que le fichier de retour que Pierre-Louis avait pour ensuite pouvoir réutiliser notre code pour faire l'agrégation.

Channel ID	Title	Categories	Description	Taille en nombre de mot
C192.api.tele	Les feux de l	['feuilleton s	C'est le jour	796
C192.api.tele	Ici tout comr	['série dram	Saison:2 - L'é	983
C192.api.tele	Demain nou:	['série dram	Saison:5 - Wi	957
C192.api.tele	New York Un	['série polici	"Saison:18 - l	424
C4.api.telera	Amour, gloir	['feuilleton s	Thomas tom	843
C4.api.telera	Un si grand s	['feuilleton p	Alors que My	871
C80.api.teler	Titeuf	['jeunesse : c	"Saison:4 - E	2261
C80.api.teler	Rex	['série polici	Saison:9 - Ep	1767
C80.api.teler	Plus belle la	['feuilleton r	Luna doute c	706

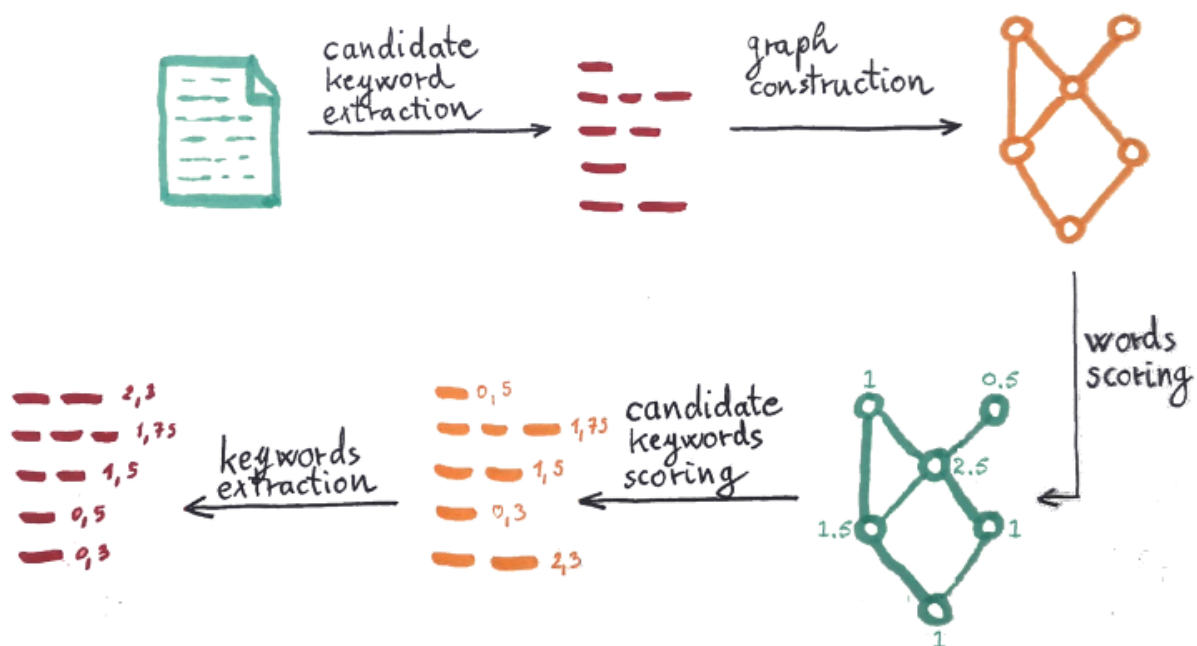
Nous passons alors de 50 articles à 167 articles dans le fichier de retour.

Extraction des mots clés

Pour l'extraction des mots-clés des descriptions d'épisodes, nous avons étudié les différents algorithmes existants pour effectuer cette tâche. Après un rapide benchmark nous nous sommes penché sur l'algorithme RAKE, qui est l'un des plus récent parmi ses concurrents.

L'algorithme RAKE se base sur des graphes. L'algorithme est basé sur l'observation que les mots-clés sont souvent composés de plusieurs mots et n'incluent généralement pas les mots vides ou les ponctuations.

Il comprend les étapes suivantes :



1. **Extraction des mots-clés candidats** — le texte est divisé en mots-clés en fonction des mots vides et du délimiteur d'expression. Un mot-clé candidat est une phrase qui se trouve entre deux mots vides ou délimiteurs de phrase. A titre d'exemple, nous pourrions classer de mot vide des verbes conjugués ("avait", "sont"), des conjonctions ("mais", "or"), etc., qui sont des mots dénués de sens syntaxique. Pour les délimiteurs de phrases, nous retrouvons tous les éléments de ponctuation.
2. **Construction d'un graphe de cooccurrence de mots-clés** : l'algorithme crée à partir de là un graphe avec des mots comme sommets, et ces mots sont alors connectés s'ils apparaissent ensemble dans les mots-clés candidats (un mot-clé peut être composé de 1 à n mots, selon le paramétrage de l'algorithme). Pour la pondération de ce graphe, le nombre de fois que des mots apparaissent ensemble dans les mots-clés candidats en étant connecté. Le graphe comprend également des connexions au sommet lui-même (chaque mot apparaît dans un mot-clé candidat avec lui-même).
3. **Notation des mots** — chaque mot du graphique est noté avec l'un des scores :
 - a. **Degré du mot** : nombre de mots avec lesquels ce mot coexiste (somme des poids des arêtes, y compris une arête qui pointe vers le sommet lui-même).

Plus un mot apparaît souvent, ou plus il existe dans des mots-clés plus longs, plus son degré sera important.

- b. **Fréquence des mots** : nombre de fois où ce mot apparaît dans un mot-clé candidat. La fréquence favorise les mots qui semblent plus fréquents.
 - c. **Rapport du degré à la fréquence** : cette métrique favorise les mots qui apparaissent principalement dans les mots-clés candidats plus longs. Il est suggéré d'utiliser soit le degré de mot, soit le rapport du degré à la fréquence. À partir de ces deux-là, le degré favorise les mots-clés plus courts.
4. Score des **mots-clés candidats** : le score de chaque mot-clé candidat est la somme des scores de ses mots membres.
5. **Mots-clés adjacents** : les mots-clés candidats n'incluent pas les mots vides. Étant donné que parfois les mots vides peuvent faire partie du mot-clé, ils sont ajoutés à cette étape. L'algorithme trouve des paires de mots-clés associés à un mot vide dans un texte et les ajoute à l'ensemble de mots vides existants. Ils doivent apparaître au moins deux fois dans le texte pour être ajoutés. Le score du nouveau mot-clé est la somme de ses mots-clés .
6. **Extraction de mots-clés** — en conséquence, 1/3 des meilleurs mots-clés de notation sont extraits.

Par la suite nous avons quand même décidé de comparer cette méthode à un autre algorithme et de croiser les résultats pour une meilleure analyse.

Nous voulions comparer l'algorithme RAKE à un concurrent plus connu et répandu : TF-IDF.

TF-IDF est un algorithme qui prend en compte l'importance des mots dans le document par rapport à l'ensemble d'un corpus. Pour chaque terme, il calcule sa fréquence dans le document, et la pondère par l'inverse de la fréquence du terme dans l'ensemble du corpus. Au final, les termes avec les scores les plus élevés sont sélectionnés comme mots-clés.

L'équation du TF-IDF est :

$$\text{TF-IDF} = \text{TF}(t, d) \cdot \text{IDF}(t)$$

term frequency

$$tf = \frac{\text{count of } t \text{ in document}}{\text{number of words in document}}$$

inverse document frequency

$$idf = \log \left(\frac{\text{number of documents in corpus}}{\text{number of documents where } t \text{ appears}} \right)$$

où "t" est le terme observé.

L'équation est appliquée à chaque mot dans la description d'épisode : la partie bleue de l'équation est la fréquence du terme et la partie orange est la fréquence inverse du document.

L'idée de cet algorithme est que les mots qui apparaissent le plus fréquemment dans le document ne sont pas forcément les plus pertinents. Cet algorithme privilégie les termes fréquents dans le document texte et peu fréquents dans les autres documents.

Cet algorithme est relativement rapide à implémenter, mais il nécessite un corpus assez grand pour la comparaison des termes, or nous ne pouvons créer des corpus contenant des descriptions de différents. Une alternative serait de prendre en compte les descriptions des différents épisodes en tant que corpus, mais cette solution fausserait les résultats car certaines séries n'ont pas plusieurs épisodes recensées en base, et cela reviendrait à revenir sur le travail de l'autre groupe.

Nous avons donc décidé d'implémenter TF-IDF pour des corpus d'un texte unique.

Le processus d'extraction suit l'ordre suivant :

- Nous prenons le fichier csv contenant les informations voulues (les descriptions d'épisodes).
- Nous sélectionnons les descriptions en fonction du nombre de mots qu'elles contiennent. La norme pour une bonne description ou texte à publier sur Internet, est un nombre de mots entre 50 et 3000 mots. Cet intervalle est une norme pour assurer un bon référencement.
- Nous appliquons l'algorithme TF-IDF
- Nous appliquons l'algorithme RAKE
- Nous comparons les mot-clés extraits et nous gardons les 5 meilleurs (plus récurrents, meilleurs scores...)
- Nous les extrayons et les ressortons dans un autre fichier CSV

Voici un exemple du traitement d'extraction avec des mot-clés de taille 1 :

Channel ID	Title	Categories	Description	Taille en nor	Mot-Cle
C192.api.tele	Les feux de l	['feuilleton s	C'est le jour	796	arrêtées,accusées,témoignage,côté,procès
C192.api.tele	Ici tout com	['série dram	Saison:2 - L'é	983	démène,charlène,détacher,début,défendre
C192.api.tele	Demain nou	['série dram	Saison:5 - Wi	957	célébrer,désespéré,révélations,dépassée,dévoilés
C192.api.tele	New York Un	['série polici	"Saison:18 - l	424	déroulée,débute,département,touchée,traumatisée
C4.api.telera	Amour, gloir	['feuilleton s	Thomas tom	843	dévastée,découvert,bébé,dernière,réinstalle
C4.api.telera	Un si grand s	['feuilleton	Alors que My	871	décolère,carrière,père,colère,décision
C80.api.teler	Titeuf	['jeunesse : r	"Saison:4 - E	2261	déléguée,accélération,thérèse,dévouées,télévision
C80.api.teler	Rex	['série polici	Saison:9 - Ep	1767	téléphone,côtés,légende,témoins,propriété
C80.api.teler	Plus belle la	['feuilleton r	Luna doute c	706	délégués,cérémonie,léa,découvre,découvrir
C34.api.teler	Pen15	['série humo	"Saison:1 - E	779	répétitions,père,réaliser,répliques,réjouissent
C34.api.teler	Schitt's Cree	['série humo	"Saison:5 - E	552	crémaillère,première,anxiété,découvre,après
C111.api.tele	360°-GEO	['documenta	"Le caviar fai	720	célèbrent,vétérinaire,véhicules,noël,siècles

Voici un même exemple pour des mots-clés de taille 3 :

Channel ID	Title	Categories	Description	Taille en nor	Mot-Cle														
C192.api.tele	Les feux de l	['feuilleton s	C'est le jour	796	dernière sème encore,inspecteur north enquête,révèle accablant,très inquiètes,trois jugées coupable														
C192.api.tele	Ici tout com	['série dram	Saison:2 - L'é	983	voyant célia douter,très stressés,révélér être,charlène remuent ciel,gaëtan décide														
C192.api.tele	Demain nou	['série dram	Saison:5 - Wi	957	détail interpelle roxane,semble très impliqué,grande décision concernant,dernière doit garder,cédric doit prendre														
C192.api.tele	New York Un	['série polici	"Saison:18 - l	424	révèle particulièrement problématique,pédophiles préalablement appâtés,unité spéciale reçoivent,conseillère municipale musulmane,alicia harding présente														
C4.api.telera	Amour, gloir	['feuilleton s	Thomas tom	843	steffy sait précisément,zoe insiste auprès,inquiéter lorsque liam,excuse auprès,fréquentait ze														
C4.api.telera	Un si grand s	['feuilleton	Alors que My	871	goût très amer,sent désemparée face,années lumières,décision à contrecœur,réveillon imprévu														
C80.api.teler	Titeuf	['jeunesse : r	"Saison:4 - E	2261	fête costumée organisée,stratégie bien huilée,après avoir bâclé,prénomme thérèse,même traitement après														
C80.api.teler	Rex	['série polici	Saison:9 - Ep	1767	reportages particulièrement polémiques,jeunes femmes célèbres,autres animaux délaissés,bébés volés,fidèle rex mènent														
C80.api.teler	Plus belle la	['feuilleton r	Luna doute c	706	caractère héroïque,réveillons contrastés,famille castel résistera,apprêtent à fêter,délégués														
C34.api.teler	Pen15	['série humo	"Saison:1 - E	779	figurines animales préférées,première soirée dansante,entraîne plusieurs malenten,chacune refusant catégoriquement,fou rire incontrôlable														
C34.api.teler	Schitt's Cree	['série humo	"Saison:5 - E	552	david tente désespérément,conseils peu avisés,david découvrent rapidement,fête prénatale,patrick emmène david														
C111.api.tele	360°-GEO	['documenta	"Le caviar fai	720	viticulteurs français privilègie,toutes très motivées,cérémonie rituelle marque,fleuve zambèze avant,bébé hérisson														
C111.api.tele	Merveilles d	['documenta	Saison:1 - Ep	319	rivière caño cristales,espèces animales endémiques,vaste forêt abritant,îles galápagos abritent,plus importantes réserves														

Interrogation Google Trends

Le dernier maillon de la chaîne du projet consiste à récupérer le fichier généré par les deux parties précédentes afin de les consolider avec des données analytiques provenant de l'outil Google Trends.



Google Trends est un outil donnant accès à des informations sur la fréquence d'utilisation de mots clés sur le célèbre moteur de recherche Google. Il permet notamment d'avoir accès à une évolution de l'intérêt porté au mot clé ainsi que des données géographiques sur celui-ci (taux de recherche pour chaque zone géographique). Une de ces fonctionnalités les plus intéressantes est aussi qu'il permet de comparer jusqu'à 5 mots clés entre eux permettant d'évaluer facilement leur pertinence en termes de recherche et de référencement, et donc de SEO.

Plus exactement, nous prenons les mots clés de chaque article et nous les envoyons à Google Trends afin d'en sortir des "pourcentages d'intérêt" pour chaque mot clé. En effet, malgré sa quantité d'informations impressionnante, Google refuse de partager les données de volumétries exacte concernant les recherches des mots clés. Il fournit en substitution une valeur d'intérêt pour les mots clés.

Voici ce qui est indiqué sur le site de Google Trends :

"Les résultats reflètent la proportion de recherches portant sur un mot clé donné dans une région et pour une période spécifiques, 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."

Ainsi avec ces données, il est alors possible et intéressant dans le cadre de notre projet de :

- Visualiser l'intérêt porté à un mot clé au cours du temps
- Comparer la pertinence de mots clés entre eux (5 mots clés extrait d'un article)
- Visualiser si un mot clé est pertinent ou non (valeur d'intérêt faibles voir nulles)



Exemple d'évolution pour "Coupe du monde de football" sur 30 jours

En débutant le projet, nous nous sommes heurté à une première problématique : Google Trends ne fournit aucune API publique permettant de récupérer automatiquement les données qu'il fournit. Google Trends permet de télécharger des extractions des données fournies sur son interface sous forme de fichiers CSV. Cependant, il n'y a aucun moyen pour un développeur de le récupérer depuis une API ou une URL dédiée.

date	mère	elevée	père	amy	morte
12/08/2021	70	1	49	11	33
13/08/2021	73	0	62	15	26
14/08/2021	71	0	55	15	37
15/08/2021	77	0	50	21	37
16/08/2021	67	6	52	27	28
17/08/2021	66	0	62	14	46
18/08/2021	93	1	53	22	27
19/08/2021	66	0	60	13	25
20/08/2021	79	0	55	18	20
21/08/2021	88	0	72	21	28
22/08/2021	92	0	54	18	66
23/08/2021	79	4	49	15	29
24/08/2021	98	0	52	9	37
25/08/2021	68	3	43	14	22
26/08/2021	83	3	45	13	35
27/08/2021	67	1	40	29	32
28/08/2021	66	0	87	15	27
29/08/2021	62	0	65	26	30
30/08/2021	63	1	41	13	21
31/08/2021	74	2	54	11	17

Exemple de fichier CSV retourné par Google Trends

Un ancien code source d'un stagiaire ayant travaillé avec M. Delalandre permettant de récupérer automatiquement les données depuis Google Trends nous avait été fourni. La solution se basait sur des simulations d'entrées souris afin de cliquer automatiquement sur le bouton de téléchargement. Cependant, la solution se basant sous des coordonnées de l'écran en pixel (pouvant varier) et nécessitant d'avoir un navigateur d'ouvert à la bonne URL, nous avons choisi de l'abandonner la trouvant trop rudimentaire.

Nous nous sommes donc mis en tête de trouver un moyen de récupérer de simuler un clic utilisateur non pas depuis l'interface mais en simulant les requêtes effectuées par le navigateur au moment du clic (ou même en arrivant sur la page si besoin).

Nous avons donc observé les requêtes envoyées au navigateur en observant la console réseau de notre navigateur. Grâce à cela, nous avons découvert que Google Trends disposait en réalité d'une API REST interne au service et non documentée.

Après avoir étudié la structure de l'API, nous avons réussi à créer un programme Python permettant de télécharger automatiquement un fichier CSV contenant les données d'évolution d'un mot clé sur 3 mois.

Cependant, en testant plus amplement notre code, nous nous sommes aperçu que de nombreux cas particuliers étaient présents et que la tâche serait encore complexifiée si nous souhaitions utiliser des fonctionnalités plus avancées comme l'ajout des comparaisons entre mots clés.

Nous avons donc recherché des bibliothèques libres permettant d'interroger directement cette API. Après plusieurs recherches, nous sommes tombés sur une bibliothèque Python méconnue mais pourtant très complète et maintenue depuis plusieurs années nommée PyTrends.

Après avoir remplacé notre code par des appels à la bibliothèque PyTrends, nous avons généré des exports de fichiers CSV pour chaque programme TV de la base de données en fournissant à Google Trends les 5 mots clés les plus pertinents retournés par la partie extraction des mots clés.

Pour chaque article, nous créons un fichier CSV contenant la comparaison des 5 mots clés. Le nom du fichier CSV est basé sur le titre du programme. Cependant certains noms de programme portant des caractères spéciaux, il nous a été nécessaire de choisir un "encodage" pour les noms des fichiers.

Note : Il est possible d'obtenir des données temporelles et par régions (pays ou région d'un pays ou ville) par mot-clés à l'aide du paramètre 'geo' de PyTrends. Il s'agit d'une liste de code de pays ou de régions à fournir. Exemple : ['US', 'FR'] pour USA et France ou ['FR-F'] pour Centre-Val de Loire.

Documentation : <https://github.com/GeneralMills/pytrends#common-api-parameters>

Résultats - Analyse des données Google Trends

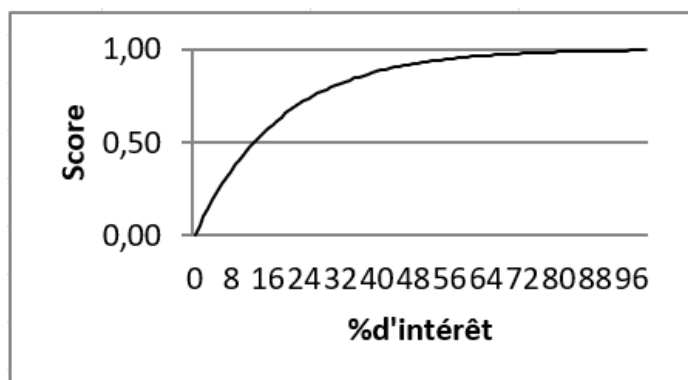
Chaque mot clé contenu dans les fichiers CSV est analysé par Google Trends, et comme évoqué précédemment il renvoie une métrique (un pourcentage) qui caractérise une évolution d'intérêt de ce mot-clé à une date donnée, et ce sur les 90 jours précédant la requête Trends.

Le problème qui a été rencontré à ce sujet est que cette métrique ne nous permet pas de mesurer exactement le trafic qui a lieu sur ce mot-clé, il peut simplement nous servir d'estimateur de la tendance.

Pour pallier ce problème il nous a été proposé d'appliquer un algorithme de scoring à ces pourcentages d'intérêt afin de pouvoir déterminer parmi les mots clés retenus lesquels sont réellement SEO compatibles.

Pour cela il a donc fallu implémenter dans notre script la fonction de scoring suivante :

$$score = e^{\frac{-(100 - \% \text{ intérêt})}{\alpha}}$$



Ainsi, en fonction du pourcentage d'intérêt relevé par jour par mot-clé, le résultat de cette fonction va permettre de normaliser ces pourcentages pour obtenir un score entre 0 et 1 qui va globalement représenter la compatibilité SEO de chaque mot-clé. Dans cette fonction, α est un facteur de pondération de l'exponentielle qui permet d'obtenir exactement un score de 1 lorsque le pourcentage d'intérêt est de 100%. Cette valeur est ici fixée à 18,8.

Ensuite nous avons moyenné tous ces scores obtenus par le nombre de jours analysés par Trends afin de définir quel mot-clé parmi les 5 retenus est le plus "populaire". Pour finir nous avons calculé une dernière moyenne sur les scores moyens des 5 mots clés afin d'obtenir un "score final" déterminant le SEO compatibilité (ou non) de l'article étudié.

Dans un premier temps nous avons traité les mots-clés de taille 3 afin de tester et de vérifier si notre programme fonctionnait bien.

Nous avons rassemblé toutes ces données dans un fichier au format CSV dont voici un exemple :

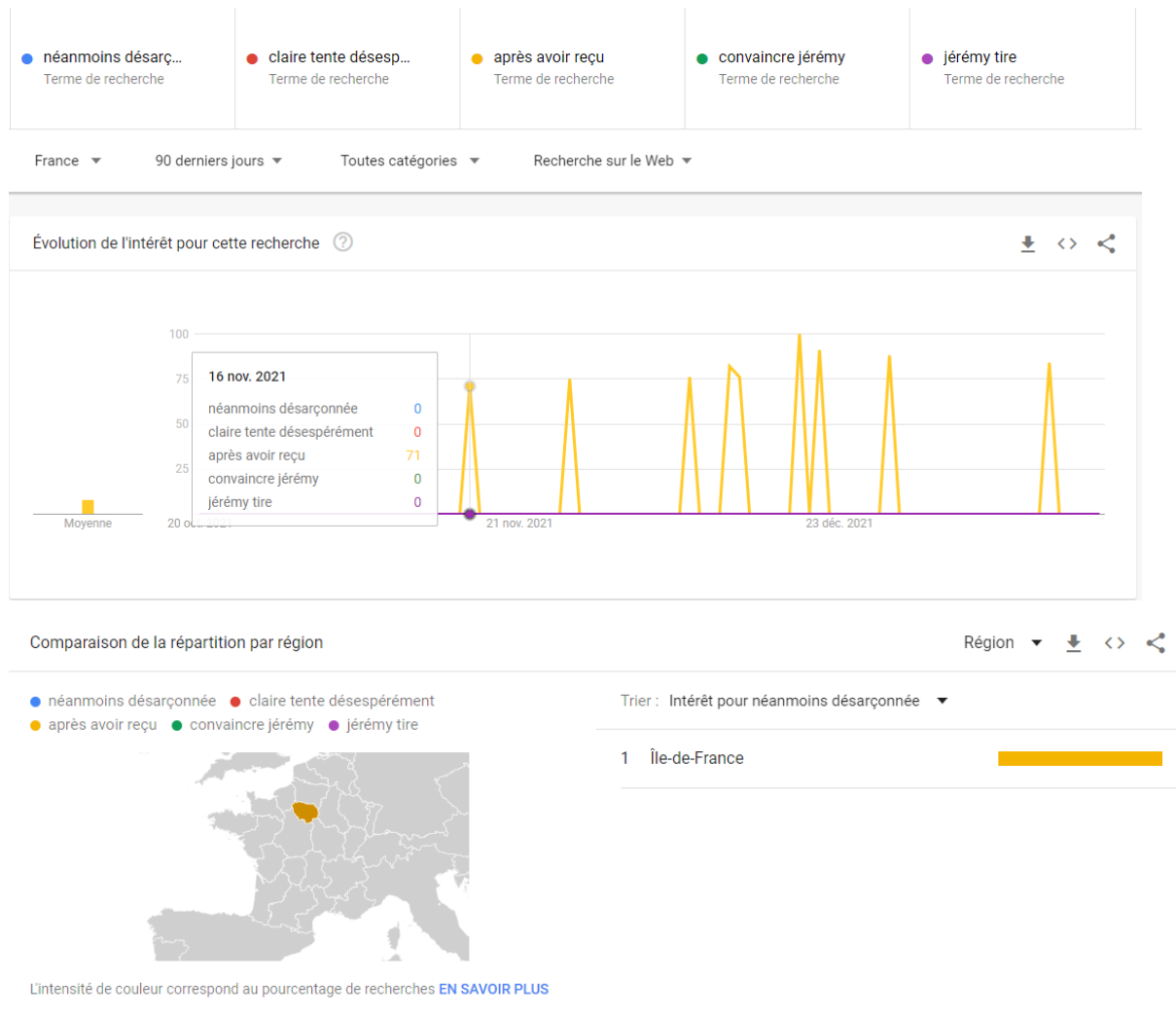
Article	Mot-Cle1	Mot-Cle2	Mot-Cle3	Mot-Cle4	Mot-Cle5	Score Moyen	
Les experts : Miami	0.0	0.0	0.0	0.0	0.0	0.0	
Ici tout commence	0.0048969195	0.0048969195	0.0592688217	0.0048969195	0.0048969195	0.015771299991827976	
Demain nous appartient	0.0393820171	0.0048969195	0.0048969195	0.0048969195	0.0048969195	0.011793939057481584	
Les feux de l'amour	0.0274117428	0.0048969195	0.0048969195	0.0048969195	0.0150070317	0.011421906643762653	
Je te promets	0.0	0.0	0.0	0.0	0.0	0.0	
New York Unité Spéciale	0.0	0.0	0.0	0.0	0.0	0.0	
Chicago Police Department	0.0	0.0	0.0	0.0	0.0	0.0	
New York, section criminelle	0.0048969195	0.0048969195	0.0272666710	0.0048969195	0.0048969195	0.009370869853872166	
Pays et marchés du monde	0.0	0.0	0.0	0.0	0.0	0.0	
Les petits meurtres d'Agatha Christie	0.0	0.0	0.0	0.0	0.0	0.0	
Les aventures du jeune Voltaire	0.0	0.0	0.0	0.0	0.0	0.0	
Speakerine	0.0	0.0	0.0	0.0	0.0	0.0	
Une planète parfaite	0.0048969195	0.0048969195	0.0048969195	0.0048969195	0.0402948162	0.011976498883636185	
La faute à Rousseau	0.0048969195	0.0048969195	0.0069392635	0.0048969195	0.0474309268	0.013812189820480807	
Tropiques criminels	0.0	0.0	0.0	0.0	0.0	0.0	
C'est pas le bout du monde	0.0048969195	0.0048969195	0.0469594580	0.0482353746	0.0048969195	0.0219771182543746	
Grantchester	0.0	0.0	0.0	0.0	0.0	0.0	
Les enquêtes de Morse	0.0048969195	0.0493012446	0.0048969195	0.0048969195	0.0048969195	0.013777784559339878	
Grizzy et les lemmings	0.0	0.0	0.0	0.0	0.0	0.0	

Nous retrouvons donc dans ce tableau les différents scores calculés pour chaque mot clé dans chaque article, ainsi qu'un score moyen qui nous permet de résumer sur le SEO compatibilité ou non des mots clés dans l'article.

Un des problèmes que l'on peut relever sur ces résultats est que les scores obtenus sont très faibles, et notamment tous les scores qui sont positionnés à "0.0". Pour ces scores cela signifie que Google Trends ne nous a absolument rien envoyé comme donnée sur ces mots-clés.

Un autre problème, cette fois lié aux scores très proches de 0 montre que sur cette analyse, ayant été faite très récemment, il est fort probable que les mots-clés définis comme étant pertinents auparavant ne sont au final pas du tout SEO compatibles à l'heure actuelle. L'une des potentielles raisons liées à ceci se trouve dans l'ancienneté des articles TV fournis en entrée.

Si l'on regarde un peu plus précisément en prenant les mots-clés retenus pour l'article sur "Ici tout commence", Google Trends nous renvoie les résultats suivants :



Ces résultats nous prouvent donc bien que les mots-clés relevés pour cet article ne sont au final pas SEO compatibles. Nous constatons également que 100% des recherches liées à ce mot-clé en France sont situées dans la région Ile-de-France.

Dans le même temps, nous avons également pu tester notre programme d'analyse sur les mots-clés de taille 1, afin de voir si les mots-clés relevés sont plus pertinents pour une SEO compatibilité.

Le fichier CSV de scoring qui en est sorti est le suivant :

Article	Mot-Cle1	Mot-Cle2	Mot-Cle3	Mot-Cle4	Mot-Cle5	Score Moyen
Les experts : Miami	0.0	0.01	0.04	0.16	0.01	0.04
Ici tout commence	0.02	0.01	0.01	0.14	0.01	0.04
Demain nous appartient	0.05	0.04	0.01	0.04	0.01	0.03
Les feux de l'amour	0.0	0.01	0.0	0.0	0.24	0.05
Je te promets	0.01	0.01	0.01	0.25	0.01	0.06
New York Unité Spéciale	0.01	0.01	0.22	0.01	0.0	0.05
Chicago Police Department	0.0	0.0	0.01	0.24	0.01	0.05
New York, section criminelle	0.31	0.01	0.01	0.01	0.0	0.07
Pays et marchés du monde	0.01	0.01	0.01	0.08	0.02	0.03
Les petits meurtres d'Agatha Christie	0.0	0.22	0.01	0.01	0.15	0.08
Les aventures du jeune Voltaire	0.0	0.0	0.03	0.0	0.02	0.01
Speakerine	0.01	0.12	0.01	0.01	0.01	0.03
Une planète parfaite	0.01	0.01	0.0	0.18	0.0	0.04
La faute à Rousseau	0.0	0.0	0.25	0.0	0.0	0.05
Tropiques criminels	0.01	0.01	0.01	0.01	0.19	0.05
C'est pas le bout du monde	0.01	0.09	0.01	0.01	0.01	0.03
Grantchester	0.0	0.01	0.16	0.01	0.01	0.04
Les enquêtes de Morse	0.01	0.01	0.01	0.16	0.01	0.04
Grizzy et les lemmings	0.02	0.2	0.0	0.01	0.01	0.05

Les scores obtenus ici, bien qu'étant toujours très faibles, sont meilleurs qu'avec des mots-clés de taille 3, ce qui pourrait indiquer que des mots-clés de taille 1 permettent d'obtenir de meilleurs résultats. Cependant, ce qui a pu être relevé également, c'est que la qualité des mots-clés trouvés joue énormément sur le score qui sera obtenu.

Pour conclure nous avons refait une analyse des données Google Trends, mais cette fois en traitant les titres des programmes TV étudiés (pour 1019 articles environ). En plus du traitement présenté précédemment, nous avons inclus le scoring du titre de l'article pour vérifier si ce titre contient des mots-clés SEO compatibles. A cause de ce rajout notre programme nécessite plus d'une heure d'exécution pour récolter toutes les données Trends nécessaires. Il ne sera ainsi pas possible d'exécuter plusieurs fois ce programme sans subir un avertissement de la part de Google Trends à cause de la demande massive de requêtes. L'exécution de notre programme nous a permis d'obtenir le fichier "score_final.csv" qui contient donc ici les informations suivantes :

	A	B	C	D	E	F	G	H	I
1	Article	Titre	Mot-Cle1	Mot-Cle2	Mot-Cle3	Mot-Cle4	Mot-Cle5	Score Moyen Mots-clés	
2	TFou	0.05	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
3	Auto Moto	0.08	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
4	Téléfoot	0.07	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
5	Les douze coups de midi	0.02	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
6	Météo	0.18	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
7	Habitons demain	0.03	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
8	Journal	0.39	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
9	Reportages découverte	0.02	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
10	Grands reportages	0.02	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
11	Les docs du week-end	0	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
12	Sept à huit Life	0.04	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
13	Handball : Championnat du monde masculin	0.02	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
14	Sept à huit	0.04	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
15	Petits plats en équilibre	0.04	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
16	TF1, rendez-vous sport	0	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
17	Et si on se réinventait ?	0	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
18	Le doudou	0.09	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
19	Esprits criminels	0.21	0.0	0.0	0.01	0.01	0.04	0.0	
20	Les experts : Miami	0.0	0.01	0.03	0.15	0.01	0.03	0.0	
21	Programmes de la nuit	0	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
22	Téléshopping	0.05	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	
23	Petits secrets en famille	0.03	0.0	0.0	0.0	0.0	0.0	Pas de mot-clé pour cet article	

Ce fichier contient la totalité des articles étudiés (1019 articles) où pour chaque ligne nous retrouvons le score obtenu par le titre de l'article, ainsi que les scores de chaque mot-clé puis le score moyen. Nous n'avons pas intégré le score du titre de l'article dans le calcul du score moyen, il ne concerne toujours que la moyenne des 5 mots-clés de l'article.

Malheureusement, nous n'avons pas pu tester à temps un autre jeu de données "plus récent" qui nous aurait permis d'obtenir des résultats plus satisfaisants en sortie d'analyse.

Axes d'améliorations - Conclusion

Dans la partie agrégation de donnée, il peut être pratique de créer une autre base de donnée, soit manuellement soit via un script afin de récupérer de vrais articles scientifiques, afin que les résultats des étapes d'après soient plus pertinents, utiles à une SEO compatibilité car ici nous avons des descriptions de film qui passent à la télévision.

De plus, si cela se fait, il peut être utile d'utiliser la partie de Pierre-Louis sur la comparaison de 2 articles avec l'indice de Jacquard qui permet de calculer un indice afin de déterminer si 2 articles sont similaires ou pas, chose qui n'a pas pu être utilisé avec les descriptions de films car lorsque 2 même film apparaissait, la description était la même à l'identique donc il suffisait simplement de retirer les doublons.

Pour la partie d'extraction des mots-clés, il serait intéressant de croiser les résultats obtenus avec un autre algorithme d'extraction de mot-clé tel que Yuke, ou via des méthodes de machine learning, qui peuvent extraire des mots-clés de manière sensiblement différente.

En complément de ces analyses, il sera intéressant de cibler l'extraction de mot-clé, en implémentant de la détection de sentiment : nous pourrions alors cibler des mot-clés relatant de la colère, de la tristesse, etc.

Un autre axe d'amélioration est l'implémentation de Part-Of-Speech, qui est une méthode qui applique un tag à chaque mot selon sa fonction dans la phrase (verbe, nom, adjectif), qui permettrait de filtrer encore l'extraction.

Concernant enfin la partie Interrogation Google Trends, et par extension l'analyse des données Google Trends, une autre solution a été proposée pour l'analyse qui serait d'intégrer les données liées aux régions ou localités. En effet, Google Trends propose, en plus des données "temporelles" étudiées précédemment, des données dites "locales" (pourcentage d'intérêt par régions, départements et villes). Il apparaît que les pourcentages d'intérêt locaux liés à ces données locales semblent être normalisés par les nombres de terminaux (smartphones, tablettes, ordinateurs, etc.) détectés localement. Ainsi, le 100% d'intérêt pour une région semble correspondre à un ratio max pour cette région "nombre de requêtes sur la région / nombre de terminaux sur la région".

Cela nous permettrait d'obtenir des pourcentages d'intérêt qui sont plus en adéquation avec l'estimation de trafic sur un mot clé donné, et d'évaluer plus précisément la popularité de chaque mot-clé par rapport à une région, un département ou une ville.

Tous ces axes d'améliorations pourront être étudiés dans une potentielle poursuite du projet.