h e g

Haute école de gestion
Genève

Recensement de mots pour la prédiction de nouvelles tendances

Travail de Master réalisé en vue de l'obtention du Master en Sciences de l'Information

par :

Steeve GERSON

Conseiller au travail de Master :

Jean-Luc SARRADE, Chargé de Cours HES

Carouge, 01.10.2020

Haute École de Gestion de Genève (HEG-GE)

Filière Information Documentaire

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce travail de Master est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre Master en Sciences de l'Information.

L'étudiant atteste que son travail a été vérifié par un logiciel de détection de plagiat.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Master, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Master, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Carouge, le 01.10.2020

Steeve Gerson

Remerciements

Tout d'abord, je souhaite remercier mon directeur de mémoire, Monsieur Sarrade Jean-Luc, Chargé de cours à la Haute Ecole de Gestion de Genève, pour son encadrement, sa disponibilité et ses conseils avisés tout au long de l'élaboration de ce travail. Je souhaite également remercier mes amis, à savoir Madame Palma Zambrella Deborah, Messieurs Sahiti Kastriot et Wintermantel Maxim ainsi que ma copine, Madame Rocourt Daphnee, tous ayant participé à la relecture de ce travail et m'ayant apporté un soutien important lors de moments de doute au cours de ce travail. Je remercie finalement ma famille pour le soutien et l'encouragement dont ils ont fait preuve durant ces années d'études ainsi que les membres de mes différents groupes de musique pour m'avoir permis de me ressourcer en leur compagnie.

Résumé

Aujourd'hui, Internet regorge d'informations et de données en tout genre. La pratique du *web mining* est de plus en plus fréquente depuis quelques années dans l'objectif d'analyser les données présentes sur le web. L'objectif de ce travail est d'étudier la possibilité de modéliser et prédire la « vie numérique » des mots appartenant à un domaine donné, en analysant le contenu de sites web. Le thème de la pandémie du Covid-19 est un sujet vivant sur Internet depuis quelques mois et permet une variété de contenus relativement différents jour après jour. C'est pourquoi les données de ce projet ont pour thème cette pandémie. En analysant temporellement l'importance attribuée à un mot appartenant à ce domaine sur une collection de site web, il est étudié la possibilité de visualiser et de modéliser sa « vie numérique » et dans quelle mesure cela peut amener à effectuer des prédictions sur ses comportements futurs. Afin de pouvoir porter cette analyse, il est nécessaire de récolter des données depuis le web. Ces données doivent pouvoir être récoltées automatiquement grâce à un algorithme de pondération du mot pour chaque site, permettant d'identifier un terme émergeant sur un site web ainsi qu'un site émergeant dans le domaine. Ce travail décrit la méthodologie appliquée ainsi que les outils utilisés pour tenter de modéliser cette « vie numérique ».

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

Actuellement, toute entreprise a besoin de gérer et maîtriser son identité numérique / image sociale fournie au travers d'Internet. Cette identité peut être représentée à un instant donné avec des outils d'analyse statistique.

Parallèlement, tout concept (mot, expression, hashtag) a une « vie » propre sur Internet au travers d'une visibilité représentée par des occurrences et des lieux (URL) d'apparition.

1.1 Contexte

Ce projet a pour but de prédire les comportements, à savoir cette « vie numérique » de ces concepts, en se basant sur un suivi temporel de leurs occurrences telles qu'elles apparaissent sur les résultats d'un moteur de recherche (Google par exemple).

Dans un premier temps, une recherche de toutes les occurrences d'un sujet va être effectuée sur Google. On analyse le contenu de chaque page des 300 (par exemple) premiers résultats de la recherche et on sauvegarde dans une base de données toutes les données susceptibles d'être exploitables. L'outil de recherche répètera cette tâche selon la périodicité souhaitée. Les données de l'historique peuvent être recherchées de manière libre (sur l'entièreté d'Internet) ou dirigée selon une sélection de certains sites.

Dans un second temps, on analyse les régularités ou irrégularités des occurrences du sujet donné afin de prédire son comportement.

1.2 Problématique

À partir de ces constats, il apparaît intéressant de modéliser temporellement cette « vie numérique » des entités. Cela permettrait, d'une part d'accéder à l'historique de l'entité, et d'autre part de prédire, dans une certaine mesure, son évolution. Pour cela, il est possible d'interroger des acteurs divers d'Internet (Google, Twitter, Facebook, Qwant, Bing,…) et d'utiliser et de comparer les résultats de leurs tris de données.

La problématique proposée pour ce travail porte sur la modélisation temporelle de la « vie numérique » d'une entité. Plus précisément, il faut apporter des éléments de réponses aux questions suivantes :

- Dans quelle mesure et comment peut-on utiliser les traces numériques (résultats issus de différents moteurs de recherches et/ou de réseaux sociaux) propres à une entité donnée pour modéliser temporellement sa « vie médiatique » ?
- A partir d'une modélisation spécifique ou générale, comment construire une prédiction de comportement de cette entité à court ou moyen terme ?

1.3 Objectif

L'objectif de ce travail est de fournir un outil permettant la prédiction de nouvelles tendances pour un sujet donné, ceci se faisant par l'analyse de la vie des mots sur Internet. De plus, le recensement de mots pourra également permettre de savoir « qui » a le plus d'influence dans un domaine.

Pour ce faire, il est nécessaire de connaître les méthodes de quantification ainsi que les méthodes de stockage de données dans le domaine des Sciences de l'Information. L'état de l'art ci-après a pour objectif d'approfondir les différents travaux effectués dans le domaine.

2. Etat de l'art

La revue de littérature de ce projet se divise en trois parties distinctes.

Une première partie consiste à étudier les écrits concernant ce qui est fait dans le domaine de la recherche et quantification de l'Information et plus particulièrement dans le cadre de la détection et prédiction de nouvelles tendances.

Une deuxième partie détaille une mesure utilisée en Sciences de l'Information concernant le calcul de fréquence des termes contenus dans un document.

La dernière partie se focalise sur le fonctionnement théorique des bases de données de type graphe et fait une comparaison entre cette structure et la structure plus traditionnelle apportée par les bases de données relationnelles.

2.1 Recherche et quantification de l'information

2.1.1 Métriques existantes en Sciences de l'Information

Dans le domaine des Sciences de l'Information, il existe différentes métriques ayant chacune leur domaine d'analyse :

- Bibliométrie
- Scientométrie
- Cybermétrie
- Infométrie
- Webométrie

(Ibekwe-SanJuan, Fidelia, 2007)

Les objectifs de chacune de ces métriques sont expliqués ci-dessous.

Bibliométrie

La Bibliométrie est une analyse quantitative des publications scientifiques. Cette science permet de mesurer la productivité scientifique d'un établissement de recherche ou d'un chercheur par le comptage des unités bibliographiques décrivant les publications (auteurs, pages etc.).

(Heilbron, 2002)

Scientométrie

La Scientométrie est une science pouvant être considérée comme étant une réduction ou une extension de la bibliométrie. Elle n'applique les techniques bibliométriques qu'au

champ des études Scientifiques et Technologiques, mais analyse en plus des publications, leurs financements, leurs ressources humaines, leurs brevets, etc.

(Suraud, 1996)

Cybermétrie

La Cybermétrie regroupe l'ensemble des méthodes et outils permettant l'analyse de données provenant du Web. Ce domaine, aussi vaste soit-il, peut englober l'analyse des habitudes des internautes sur un site web, l'analyse du trafic sur Internet, l'optimisation de campagnes publicitaires, etc.

(Sen, 2004)

Infométrie

L'Infométrie est la science étudiant des données quantitatives économiques, humaines ou encore bibliographiques. Dans ce dernier cas, il s'agit plutôt de Bibliométrie. Sur ces données sont appliqués des traitements principalement descriptifs et classificatoires.

(Coadic, F, 2005)

Webométrie

La webométrie est une science d'analyse quantitative du web. Björneborn l'a définie en 2004 comme étant :

> *« The study of quantitative aspects of the construction and use of information ressources, structures and technologies on the Web drawning on bibliometrics and informetrics approches ».* *(Björneborn, Ingwersen, 2004, p. 2)*

Cette science aborde les thématiques d'analyse du contenu des pages web, d'analyse de la structure hypertextuelle du web, d'analyse des usages du web et d'analyse des technologies du web.

Nous nous intéressons dans ce qui suit au *web mining* qui repose sur la webométrie

2.1.2 Web Mining et prédiction de nouvelles tendances

Le *web mining* est l'adaptation des techniques de *data mining* pour des données se trouvant sur le *World Wide Web*. Cela consiste à récupérer des données en masse aux travers de différents moteurs de recherche, réseaux sociaux et autres outils fournissant de l'information grâce au web.

(Cooley et al., 1997)

Nous avons repéré que très peu d'études ont été faites concernant la détection et prédiction de nouvelles tendances en utilisant le Web Mining. Le peu de travaux de

détection de nouvelles tendances ont été effectués par analyse temporelle des liens hypertexte. En effet, Einat Amitay et al. ont publié en 2004, dans le « Journal of the American Society for Information Science and Technology » (JASIST), un article étudiant la possibilité d'ajouter une dimension temporelle lors de l'analyse structurelle des liens, permettant ainsi de détecter les tendances actuelles.

(Ibekwe-SanJuan, Fidelia, 2007)

Dans cet article, Amitay et ses compères mentionnent le travail de Kleinberg paru en 2000, lequel porte sur l'analyse temporelle des données dans le contexte des e-mails. Pour ce travail, Kleinberg est parti du principe que les e-mails présents lors d'un échange font tous partie du même contexte.

Selon Amitay, les données nécessaires à l'analyse temporelle des liens sont les suivantes :

- La date de création et de dernière modification de la page
- La date de détection de suppression de la page
- Les différentes dates de création et de suppression des liens

Sur ce concept, Kraft and Hastor ont lancé un projet en 2003 afin de créer une infrastructure permettant l'étude de l'évolution des liens à travers le temps, utilisant des *snapshots* du Web, c'est-à-dire en enregistrant l'état dans lequel se trouve le Web à un instant *t*. Un *snapshot* correspond à un instantané de l'état d'une entité ou d'un objet à un instant précis.

Partant des données nécessaires mentionnées ci-dessus, Amitay propose le concept de « *Timestamped Links* ». Il s'agit ici de tout d'abord définir un thème de recherche, puis de récolter une collection des pages *P* retournées par le moteur de recherche. *P* étant une collection de pages, un « *Timestamped Link* » d'une page *p* de *P* sera une paire (*u,t*) où *u* est une URL de *p* dont la date de dernière modification a été faite au temps *t*.

Basé sur les « *Timestamped Link* », Amitay propose également un outil utile à la mesure des activités temporelles au sein d'une communauté sur le Web. Cet outil, nommé « *Timestamped Link Profile* », est une projection normalisée des « *Timestamped Links* » d'un thème sur une ligne temporelle.

Quelques exemples de mise en application sont mentionnés dans ce même article :

- Comparaison du niveau d'activité des communautés pour les thèmes relevés
- Comparaison du même thème à différents points temporels
- Suivi de l'évolution d'un thème à travers le temps
- Ajouter une dimension temporelle aux autorités du Web

(Amitay et al., 2004)

Prédiction des liens

L'article de Jon Kleinberg et David Liben-Nowell, paru en 2007 dans le « Journal of the American Society for Information Science and Technology », nommé « The Link-Prediction Problem for Social Networks », mentionne différents outils de prédiction de liens dans un réseau donné.

Cette étude se base sur la théorie des graphes pour comparer différentes techniques de prédiction des liens, dans ce cas pour un réseau social. La technique la plus commune et la plus intuitive est celle des voisins en communs. Selon cette technique, une forte probabilité que deux nœuds (x, y) dans un réseau (R) n'étant pas liés à l'instant T, le soient à l'instant T+1 si ces deux nœuds ont des voisins en communs. Un voisin du nœud x est un nœud ayant un lien avec x.

(Liben-Nowell, Kleinberg, 2007)

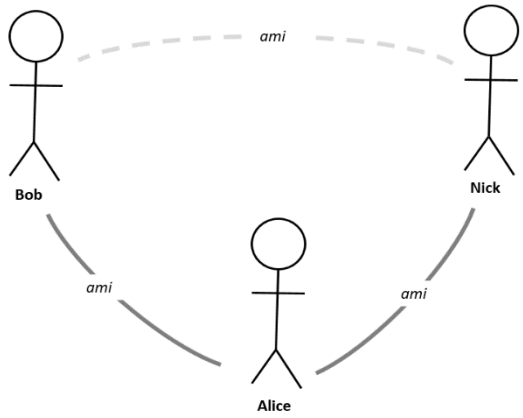
Plus récemment, en janvier 2014, Peng WANG publie un article de recherche concernant la prédiction de nouveaux liens au sein des réseaux sociaux (*Link Prediction in Social Network : the State-of-the-Art*).

En effet, WANG mentionne une difficulté à prédire des nouveaux liens au sein d'un réseau social. Il nous présente l'exemple suivant :

> *« A l'instant t, Alice et Bob sont amis. Alice est également amie avec Nick. A l'instant t+1, peut-être qu'Alice aura présenté Bob à Nick et qu'ils deviendront amis également. »* *(Wang et al., 2014, p.4)*

La représentation graphique de cet exemple serait un graphe, chaque nœud représentant une personne, associé à ses relations, comme le montre la figure ci-dessous :

Figure 1 : Liens de relations



(Wang et al., 2014, p.5)

Avec cet exemple, il n'est pas question de prédiction concernant une future amitié entre Bob et Nick, mais simplement d'une supposition. Différents facteurs humains (que je n'aborderai pas ici) entreront en considération concernant une amitié entre ces deux personnes. Cet exemple montre que prédire des futures relations, dans le cas de relations humaines, est difficile.

Mesure de fréquence des mots

Un des objectifs de ce travail étant d'être capable de définir quel média détient le plus d'autorité sur un thème donné, il est nécessaire pour cela d'analyser le contenu de ce média. Pour ce faire, il existe la mesure « *Term Frequency-Inverse Document Frequency* » (TF-IDF) qui permet d'évaluer l'importance d'un terme dans un document. Cette méthode de pondération ajoute et augmente le poids d'un terme en fonction du nombre de fois qu'il apparaît sur ledit document.

(Christian et al., 2016)

Une variante de cette mesure statistique est utilisée, par exemple, par les moteurs de recherche afin d'évaluer la pertinence du document retourné par rapport à la recherche de l'utilisateur.

Cette mesure se découpe en deux parties. La partie « *Term Frequency* » (TF), est le calcul de fréquence du mot sur le document (ou la page).

Figure 2 : Formule de « *Term Frequency* »

$$TF = \frac{Total\_appearance\_of\_a\_word\_in\_document}{Total\_words\_in\_a\_document}$$

(Christian et al., 2016, p.289)

La seconde partie de cette métrique, « *Inverse Document Frequency* », consiste à calculer l'importance du mot en question sur l'ensemble des documents présents dans le corpus, soit, dans le cadre de notre projet, sur l'ensemble des pages web récupérées.

Figure 3 : Formule de « *Inverse Document Frequency* »

$$IDF_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

(Christian et al., 2016, p.289)

Où :

- $|D|$ représente le nombre total de documents dans le corpus.
- $|\{d_j : t_i \in d_j\}|$ représente le nombre de documents où le terme t_i apparaît.

Cette formule peut ainsi être représentée plus simplement par :

Figure 4 : Formule de « *Inverse Document Frequency* » simplifiée

$$IDF = log \frac{All\_document\_number}{Document\_frequency}$$

(Christian et al., 2016, p.289)

Finalement, ces deux indicateurs doivent être multipliés entre eux afin d'obtenir le poids attribué au terme :

Figure 5 : Formule de « *TF-IDF* »

$$TF - IDF = TF * IDF$$

(Christian et al., 2016, p.289)

2.2 Stockages de l'information

2.2.1 Théorie des Graphes

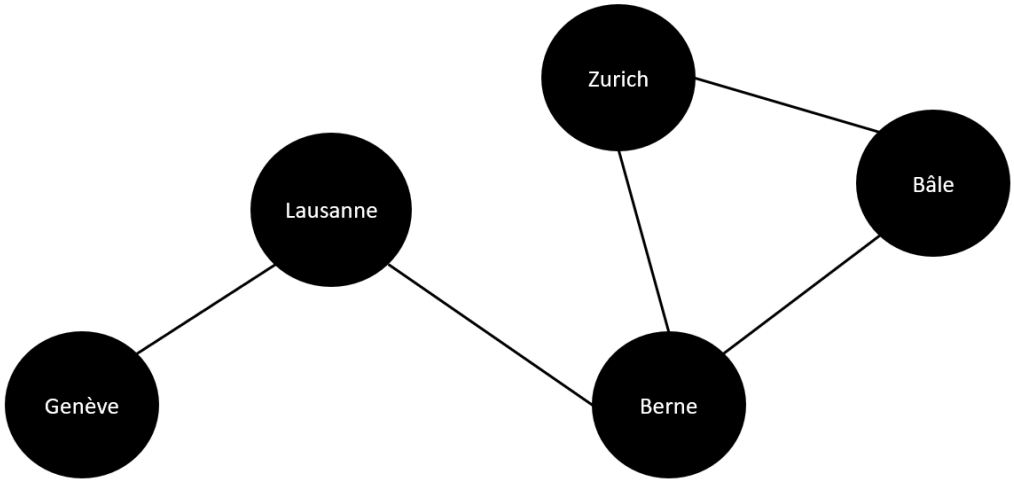
Comme vu ci-dessus, plusieurs chercheurs, dont les travaux ont pour base la pratique du *web mining*, ont structuré leurs données sous forme de graphe.

Un graphe est un ensemble de nœuds, reliés entre eux par des arêtes. Chaque nœud représentant une entité du domaine étudié, les arêtes symbolisant un lien entre ces nœuds.

Un exemple, parmi d'autres, d'une mise en application des graphes, serait le réseau routier d'un pays. Chaque ville d'un pays peut être représentée par un nœud, tandis que chaque route reliant ces différentes villes peut être représentée par une arrête.

La *Figure 6 : Exemple de graphe de villes suisses* illustre la liaison routière entre cinq villes suisses, à savoir Genève, Lausanne, Berne, Zurich et Bâle.

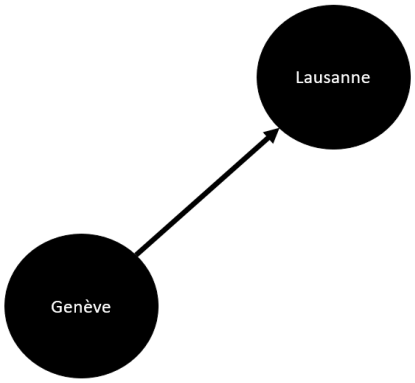
Figure 6 : Exemple de graphe de villes suisses



Ce graphe est considéré comme un graphe « non-orienté ». Cela signifie que les arêtes n'ont pas de sens défini. Dans ce cas, il peut être admis que chaque route entre les villes est bidirectionnelle.

Un graphe dit « orienté » est un graphe dont les arêtes ont un sens défini. Cela signifie que la relation entre les nœuds est unidirectionnelle. En reprenant l'exemple ci-dessus, un graphe orienté contiendrait des relations comme celles-ci :

Figure 7 : Exemple d'un graphe dirigé



Cette relation unidirectionnelle signifierait, dans cet exemple, que le trajet entre Genève et Lausanne n'est possible que dans un sens.

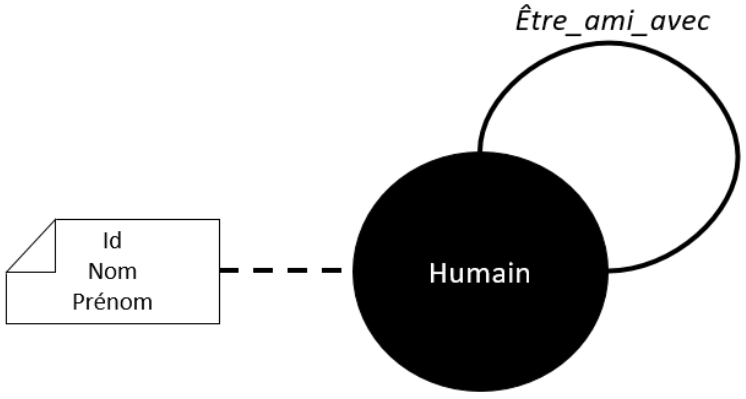
(Philippe Lacomme et al., 2003)

2.2.2 Base de données NoSQL de type Graphe

On retrouve des bases de données orientés graphe principalement dans la gestion des réseaux sociaux tels que Facebook, LinkedIn ou Instagram, mais également dans tout domaine où les données peuvent être représentées sous forme de réseau (ou de graphe).

Pour un réseau social, par exemple, chaque nœud représente une personne (cf Figure 1 : : Liens de relations) tandis que les arêtes représentent une relation (« être ami », dans le cas de Facebook). On peut ainsi générer le modèle de base de données en graphe suivant :

Figure 8 : Exemple conceptuel d'une relation entre personnes sous forme de graphe



Neo4J étant un système de gestion de base de données basé sur la théorie mathématique des graphes, son équipe donne une définition claire et précise du concept de nœud et de relation présents dans leur base de données :

Noeud

> *« Les nœuds sont les entités du graphique. Ils peuvent contenir n'importe quel nombre d'attributs (paire clé-valeur) appelés propriétés. Les nœuds peuvent être balisés avec des étiquettes, représentant leurs différents rôles dans le domaine à analyser. Les étiquettes de nœud peuvent également servir à attacher des métadonnées (telles que des informations d'index ou de contrainte) à certains nœuds. »*

(Neo4J, 2020b)

Relation

> *« Les relations fournissent des connexions dirigées, nommées et sémantiquement pertinentes entre deux entités de nœud(…). Une relation a toujours une direction, un type, un nœud de départ et un nœud de fin. Comme les nœuds, les relations peuvent également avoir des propriétés. Dans la plupart des cas, les relations ont des propriétés quantitatives, telles que les poids, les coûts, les distances, les évaluations, les intervalles de temps ou les forces. En raison de la manière efficace de stocker les relations, deux nœuds peuvent partager n'importe quel nombre ou type de relations sans sacrifier les performances. Bien qu'elles soient stockées dans une direction spécifique, les relations peuvent toujours être parcourues efficacement dans les deux sens. »*
>
> (Neo4J, 2020b)

Ce type de base de données est accompagné d'un langage de requête (Cypher), permettant ainsi le parcours complet du graphe. Un nœud d'une personne sera relié à autant d'autres nœuds que le nombre d'amis que cette personne possède.

Cypher est un langage de requête, au même titre que SQL pour une base de données relationnelle, permettant ainsi la lecture et l'écriture dans une base de données de type graphe.

SQL permet la création de tables ainsi que leurs diverses contraintes d'intégrité, l'ajout, la modification et la suppression de données et la lecture de ces données. Cypher a le même rôle dans Neo4J. Grâce à ce langage, il va être possible de créer, modifier et supprimer des nœuds, leur donner des attributs et créer les relations entre ces nœuds.

L'équipe de Neo4J donne cette définition de Cypher :

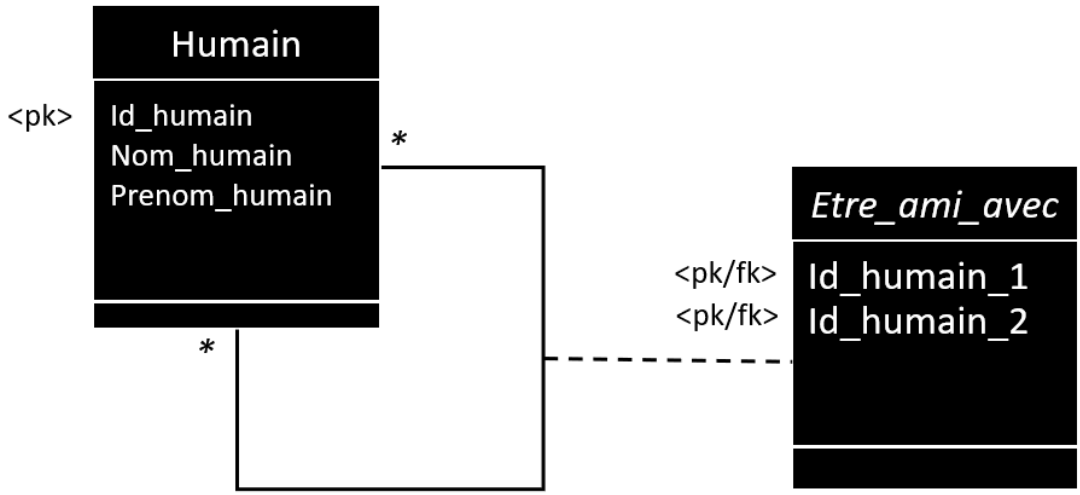
> *« Cypher est un langage de requête graphique de Neo4J qui permet aux utilisateurs de stocker et de récupérer des données de la base de données graphiques. Neo4J souhaitait rendre les requêtes de données graphiques faciles à apprendre, à comprendre et à utiliser pour tout le monde, mais aussi intégrer la puissance et les fonctionnalités d'autres langages d'accès aux données standards. C'est ce que vise Cypher.*
>
> *La syntaxe de Cypher fournit un moyen visuel et logique de faire correspondre les modèles de nœuds et de relations dans le graphique. Il s'agit d'un langage déclaratif inspiré de SQL pour décrire les modèles visuels dans les graphiques à l'aide de la syntaxe ASCII-Art. Il nous permet d'indiquer ce que nous voulons sélectionner, insérer, mettre à jour ou supprimer de nos données graphiques sans une description exacte de la façon de le faire. Grâce à Cypher, les utilisateurs peuvent créer des requêtes expressives et efficaces pour gérer les fonctionnalités de création, de lecture, de mise à jour et de suppression nécessaires.*
>
> *Cypher n'est pas seulement le meilleur moyen d'interagir avec les données et Neo4J – il est également open source ! Le projet openCypher fournit une spécification de langage ouvert, un kit de comptabilité technique et une*

implémentation de référence de l'analyseur, du planificateur et du runtime pour Cypher. Il est soutenu par plusieurs sociétés du secteur des bases de données et permet aux développeurs de bases de données et aux clients de bénéficier, d'utiliser et de contribuer librement au développement du langage openCypher. »

(Neo4J, 2020a)

L'équivalent de cette structure de base de données serait représenté comme suit pour une base relationnelle traditionnelle :

Figure 9 : Exemple conceptuel d'une relation entre personnes sous forme relationnelle



Le modèle ci-dessus représente le fait qu'un être humain peut être ami virtuellement avec plusieurs autres êtres humains. Afin de stocker la liste des amis de chaque humain dans la base de données, nous utilisons la table d'association « Etre_ami_avec », stockant uniquement les IDs des deux humains étant amis. Ceci permet ainsi d'avoir une infinité de possibilités de couple d'amis.

En revanche, pour pouvoir avoir accès aux noms et prénoms de ce couple d'ami, une liaison entre la table « Humain » et la table d'association « Etre_ami_avec » sera nécessaire, ce qui coûte cher en temps de calcul.

2.3 Conclusion de l'état de l'art

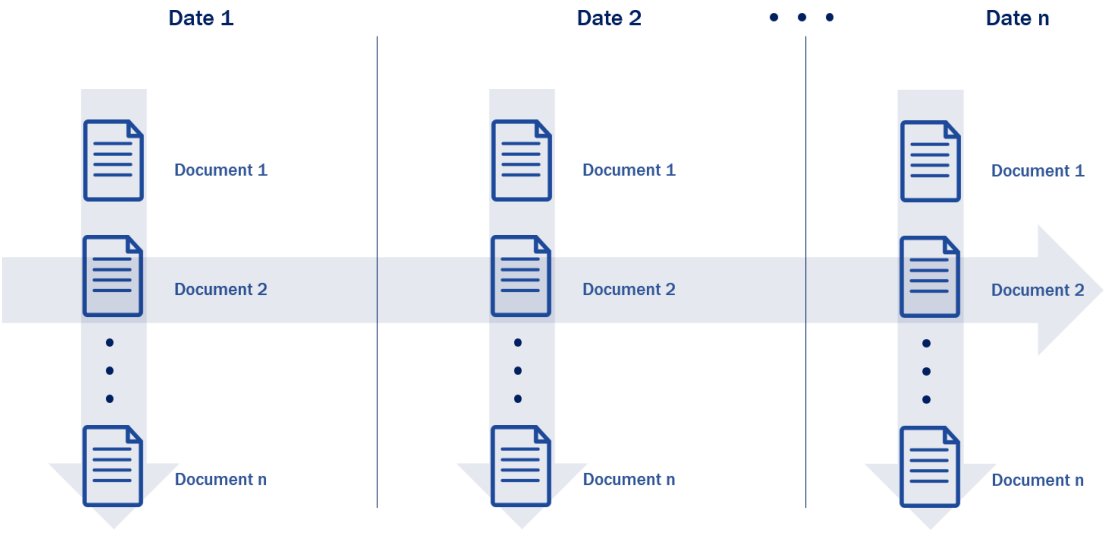
Afin de conclure cette revue de la littérature, cette section a pour objectif de présenter quels outils seront utilisés dans la suite de ce projet et à quelle fin.

Les données seront stockées dans une base de données relationnelle traditionnelle. Nous verrons plus tard quelle architecture a été choisie pour la construction de cette base de données. A partir de là, le système de Neo4J, s'appuyant sur la théorie des

graphes, permettra une modélisation ainsi qu'une visualisation de la « vie numérique » d'une entité, tandis que l'importance d'une entité répartie dans le temps sera calculée grâce à la mesure TF-IDF.

Lors des différentes recherches effectuées, je me suis aperçu que les différents travaux effectués, et présentés ci-dessus, se basent sur une analyse verticale des documents. C'est-à-dire que l'analyse se fait en comparant les différents documents, mais pas de façon horizontale en analysant un document. L'analyse que je souhaite porter se fera donc de manière horizontale, afin de pouvoir analyser un même document à travers le temps.

Figure 10 : Méthode d'analyse



La *Figure 10 : Méthode d'analyse* ci-dessus illustre la méthode d'analyse que je souhaite adopter au travers de ce travail en comparaison avec la méthode d'analyse de l'état de l'art. En plus de conserver l'axe vertical de comparaison des documents à une temps *t*, je souhaite ajouter la dimension horizontale en analysant un même document sur plusieurs dates, soit en ajoutant une notion temporelle.

3. Méthodologie

Ce chapitre a pour objectif de présenter les différentes étapes mises en place afin de mener à bien ce projet.

3.1 Récolte de données

La récolte de données s'est faite manuellement durant le mois de mars, puis entre juillet et août. Le manque de données entre mars et juillet s'explique par le fait de ma mobilisation à la Protection Civile de Genève. Ce travail a donc dû être mis en arrêt à partir du mois d'avril jusqu'à fin juin. Des fichiers HTML, basés sur des recherches Google dont le mot clé est un interne au domaine à analyser, ont été enregistrés et classés journalièrement, ceci afin de conserver une trace temporelle du contenu de chaque page web ressortie par la recherche Google.

Un sujet d'actualité, le thème du Covid-19, a été choisi comme thème de recherche de test. Ce thème présentant la faculté d'avoir des résultats de recherche relativement différents tous les jours, l'analyse temporelle devait, à priori, être facilitée. J'ai pris en compte arbitrairement quatre axes de recherches à effectuer sur Google, tous les jours, afin d'évaluer et analyser les résultats retournés selon ces quatre mots-clés :

- Covid-19
- Coronavirus
- Gel hydroalcoolique
- Masque de protection

Cette récolte peut être faite de façon automatisée, en utilisant les poids attribués au différents termes (cf 3.8 Calcul de pertinence) afin d'exécuter une requête sur Google avec des mots dont le poids est de plus en plus important au cours du temps. Estimant que la mise en place d'une récolte automatique n'était pas un aspect prioritaire de ce travail, cette dernière n'a pas été effectuée afin de conserver du temps aux autres aspects présentés dans ce chapitre.

3.2 Nettoyage des URLs

Chaque fichier HTML récupéré retourne un ensemble de pages web à ouvrir et à *crawler*. Un *web crawler* est un outil informatique permettant de récupérer toutes les données présentes sur Internet. Il analyse une page ou une collection de pages, récupère les données présentes sur la(les) pages(s) en question et en retourne l'ensemble.

(Thelwall, 2001)

Cependant, lors de la récupération des URLs, une phase de nettoyage est nécessaire. En effet, l'ouverture des fichiers HTML retourne toutes les URLs présentes sur la page.

Une table dans une base de données MySQL a été construite, contenant toutes les URLs récupérées afin d'en faciliter le filtrage. L'*annexe 7.1 Nettoyage des URLs* illustre l'utilisation d'une requête SQL filtrant les liens utiles à analyser, excluant ainsi les liens redirigeant sur les paramètres Google par exemple.

Le système de stockage basé sur une table de base de données a été choisi afin de pouvoir facilement sélectionner les URLs à analyser. Certaines URLs contenant le mot clé « google », qui sont à priori à exclure du dataset, sont tout de même utiles. En effet, l'ouverture d'un fichier HTML ne retourne que les liens présents sur la première page de résultat de Google. Or, il est important de pouvoir analyser des pages web présentes sur les pages suivantes de résultat de Google. L'*annexe 7.2 Récupération des liens redirigeant sur d'autres pages Google* illustre la requête SQL exécutée afin de les récupérer.

Lors de la récupération des liens URLs présents dans un fichier HTML, certaines données de valeur NULL ont été retournées, de même que la requête de recherche dans la barre de recherche Google. Ces données ont été supprimées, également grâce à une requête SQL afin de ne pas surcharger la base de données.

3.3 Nettoyage des mots

Les données utilisées dans ce projet sont des données de type texte. Une phase de nettoyage a été nécessaire afin d'harmoniser les données entre elles.

Le processus de *crawling* des sites web retourne l'entièreté du contenu de la page sous forme d'un texte. Ce texte comprenant de nombreux caractères inutiles à l'analyse du contenu, il a fallu les supprimer des données récoltées. Pour ce faire, la librairie open-source « *Natural Language Toolkit* » (NLTK), disponible sous Python, a été utilisée car elle permet un nettoyage rapide et efficace des données de type texte.

(Loper, Bird, 2002)

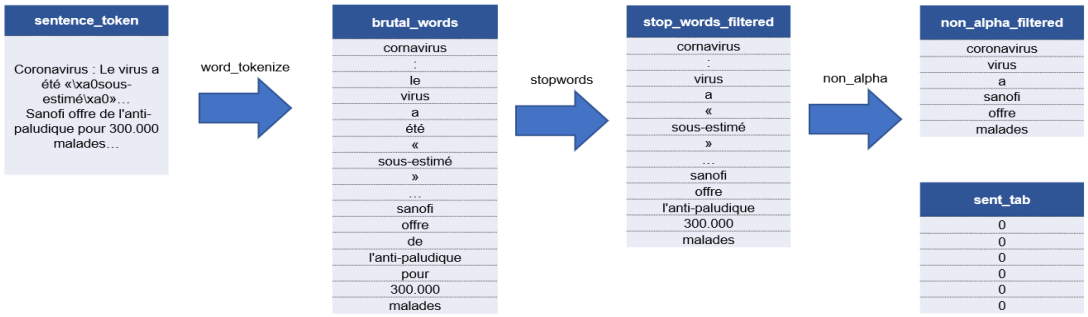
3.3.1 Phase de nettoyage par NLTK

La première étape de ce nettoyage fut le fait de séparer tout le texte en phrase. NLTK parcourt ainsi entièrement le texte, et sépare le texte en « *Sentence Tokens* ». Une fois les phrases identifiées, chacune de celles-ci a été découpée en mots, dans une structure différente, permettant ainsi de garder une structure contenant les phrases et une autre

structure contenant tous les mots. Pour chaque mot associé à une phrase, un « flag » lui a été attribué permettant ainsi de savoir dans quelle phrase ce dernier se trouve.

Il est important de préciser que chaque mot conservé est un mot utile à l'analyse. En effet, le traitement de base de NLTK, qui effectue une « tokenisation » des mots, garde absolument tout. C'est pourquoi il a fallu affiner ce nettoyage sur l'ensemble des mots ressortis, afin de supprimer tous les mots parasites (*stopwords*) ainsi que tous les signes de ponctuations.

Figure 11 : Processus de nettoyage des mots



La *Figure 11 : Processus de nettoyage des mots* illustre le processus de nettoyage des mots en utilisant la librairie NLTK. Les données présentes dans cette figure sont des données récoltées durant le processus de *crawling* d'une page web.

La librairie NLTK étant certes élaborée, il y a tout de même quelques dysfonctionnements, notamment lors du filtre des *stopwords*. En effet, NLTK propose plusieurs corpus de langues différents, contenant chacun une liste plus ou moins exhaustive de *stopwords* correspondants. Cependant, certains *stopwords* ne sont pas présents dans le corpus proposé par la librairie. Il est donc nécessaire de les ajouter manuellement si ceux-ci alourdissent trop le jeu de données.

Pour plus de détails quant à l'algorithme, l'*annexe 7.3 Processus de nettoyage des mots* présente la fonction python développée utilisant la librairie NLTK.

3.3.2 Phase de filtrage des mots à analyser

La première phase de nettoyage des mots grâce à la librairie NLTK a permis de faire ressortir uniquement du texte prêt à être utilisé et analysé. Ces données, pourtant nettoyées, contiennent toutefois des mots inutiles au domaine à analyser.

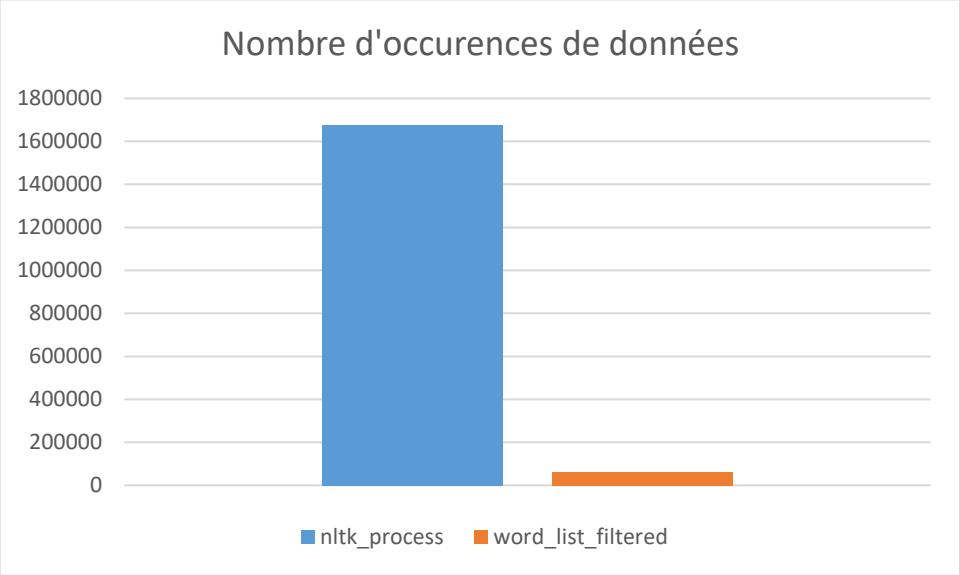
Le domaine de test d'analyse pour ce projet étant l'épidémie du Covid-19, il a été nécessaire d'appliquer un filtre de mots portant sur ce thème sur l'ensemble du jeu de données, afin de pouvoir analyser le contenu des pages web en se basant uniquement sur des termes liés au domaine à analyser.

Il a donc fallu créer une liste de mots, sur le thème du Covid-19, pour pouvoir l'appliquer comme filtre sur le jeu de données. Pour définir les mots présents dans cette liste, une phase de « brainstorming » a été mise en place. J'ai donc effectué cette phase accompagné d'un collègue afin d'avoir une meilleure variété de mots. L'*annexe 7.4 Liste de mots à analyser* illustre cette liste, non-exhaustive, de mots.

Afin d'appliquer le filtre, la façon la plus simple qui m'est apparue fut de créer une table dans la base de données contenant précisément ces mots, puis de créer une vue en exécutant une requête SQL récupérant uniquement tous les liens URLs contenant un des mots de la liste.

La *Figure 12 : Nombre d'occurrences de données entre juillet et août 2020* présente la différence de nombre de données entre avant l'application du filtre de la liste de mots (c'est-à-dire, ayant uniquement appliqué le nettoyage avec NLTK), et après l'ajout du filtre, sur les fichiers HTML récupérés entre juillet et août 2020.

Figure 12 : Nombre d'occurrences de données entre juillet et août 2020



Le Tableau 1 présente le delta de nombre de mots différents entre avant et après l'ajout du filtre de mots.

Tableau 1 : Différence du nombre de mots

Nltk_process	Word_list_filtered
39001	55

Ce filtre nous donne ainsi une liste de 55 mots possibles pour l'analyse des termes liés au Covid-19. J'ai décidé de vous présenter dans ce rapport uniquement les résultats pour trois d'entre eux, à savoir « Confinement », « Pandémie » et « Epidémie ». Ces trois termes ont été sélectionnés consciemment. C'est-à-dire qu'il était voulu de ne pas analyser les mêmes termes que lors de la récolte de données, à savoir « Covid-19 », « Coronavirus », « Masque » et « Gel hydroalcoolique » afin de ne pas être complètement dépendant de l'algorithme d'indexation fourni par Google. Cette méthode permet ainsi d'analyser des termes faisant partie du même domaine d'analyse que les termes de recherche sans pour autant analyser les mots recherchés, lesquels ayant une forte probabilité de ressortir sur les documents fournis par les résultats de Google.

3.4 Regroupement des données

Afin d'identifier la prédictibilité d'un média, il a été décidé de regrouper les données récupérées en fonction du site auquel appartient l'URL parcourue. Chaque mot provenant d'une ou plusieurs URLs connues, celles-ci ont été regroupées en fonction du site web auquel elles appartiennent. Ceci favorise une analyse par site et non une analyse indépendante par URL.

Le tableau ci-dessous présente le nombre d'occurrences de données après application du filtre de mots ainsi que le nombre de sites différents présents dans le jeu de données.

Tableau 2 : Occurrences de données et nombre de sites

Occurrences de données	Nombre de sites différents
62'414	119

Afin d'ajouter un axe d'analyse supplémentaire, ces 119 sites différents ont été classifiés manuellement pour permettre l'étude des différences de résultats entre les différentes

classes de site prises en compte. Ces sites se répartissent ainsi parmi les 11 catégories suivantes :

Tableau 3 : Nombre de sites par catégorie

Catégories	Nombre de sites
News	61
Etat	12
Health	9
Social Media	3
Utilities	1
Research Engine	1
Encyclopedie	2
Online Shop	27
Assurance	1
Telecom	1
Gaming	1

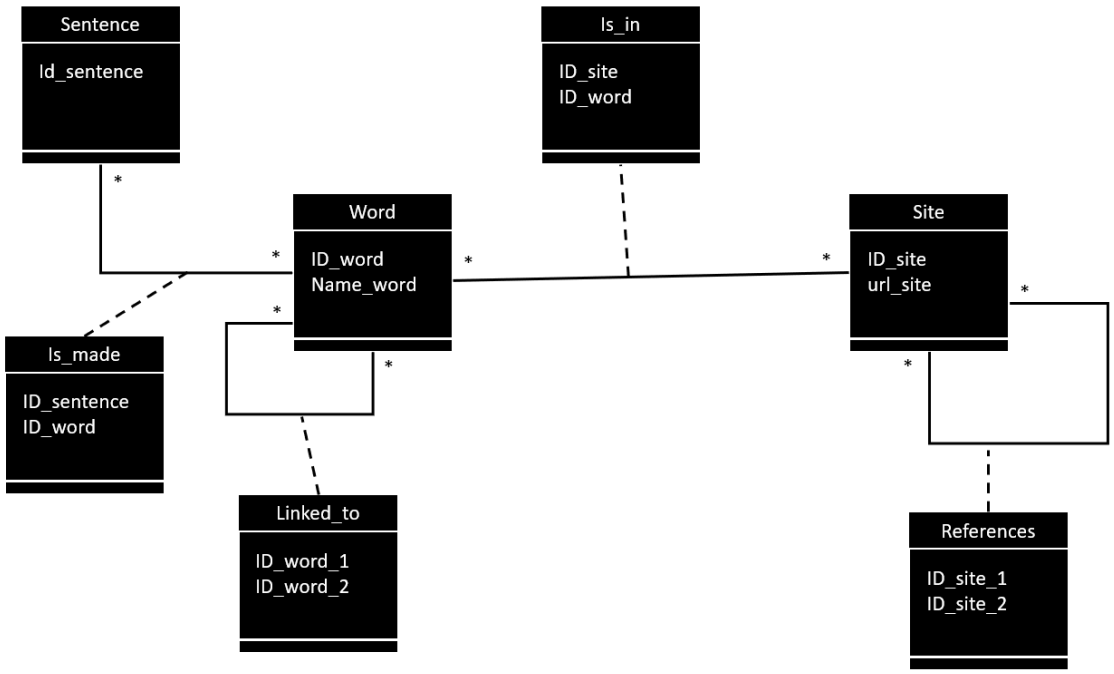
3.5 Architecture de la base de données

Ce chapitre a pour but d'illustrer l'évolution de l'architecture de la base de données. Savoir comment structurer correctement les données récupérées avant de les mettre en forme a été un travail conséquent et une étape qui a dû être revue à plusieurs reprises.

Pour commencer, il a été décidé de stocker les données dans une base de données relationnelle. Cette base de données n'a pas pour objectif d'analyser et de comprendre les données, mais simplement de les stocker, pour ensuite pouvoir les réutiliser. Le modèle de la base de données s'est construit petit à petit, en lien avec les différentes étapes de nettoyage mentionnées ci-dessus. Le modèle conceptuel de données (MCD) ci-dessous illustre la structure visée pour la base de données relationnelle. Il est à noter que le modèle est déjà lourd. En effet, celui-ci contient uniquement des relations « *many-to-many* » car :

- Une phrase contient plusieurs mots et un mot se trouve dans plusieurs phrases
- Un site contient plusieurs mots et un mot peut se retrouver sur plusieurs sites
- Un mot peut être relié à plusieurs mots
- Un site peut être référencé par plusieurs sites et un site peut référencer plusieurs sites

Figure 13 : Modèle conceptuel de la base de données



Cette architecture étant très lourde en lecture et écriture, il a été décidé de stocker toutes les données dans une unique table afin de gagner en performance lors de l'écriture et de la lecture des données.

Figure 14 : Table contenant toutes les données

Cette table unique, contenant toutes les données, ne respecte évidemment pas le concept d'une base de données relationnelle, visant à ne pas avoir de duplication de données. Or, dans ce cas précis, il est plus simple de répertorier pour chaque mot, l'url sur lequel il a été trouvé, le nom du site correspondant à l'URL en question, ainsi que sa date d'apparition. En adoptant une vision verticale de cette table, c'est-à-dire en prenant un seul attribut (soit une seule colonne), il y aura des doublons de données. Cependant, en optant pour une vision horizontale, c'est-à-dire en prenant en compte une seule donnée par attribut (soit une ligne complète), le risque de doublons est moindre.

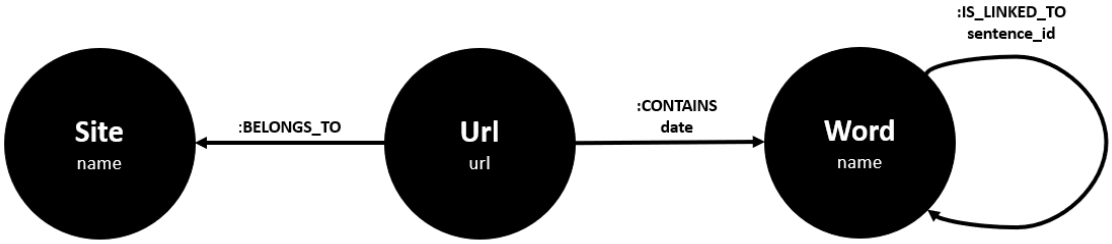
3.6 Mise en forme des données

Ce chapitre va montrer quelle structure de graphe a été mise en place afin de représenter au mieux les données. Cette structure a également été sujette à de nombreuses modifications.

3.6.1 Structure du graphe dans Neo4J

Cette section va présenter les différents éléments composant le graphe des données de ce projet ainsi que l'architecture du graphe.

Figure 15 : Architecture conceptuelle du graphe de données



Les trois nœuds ci-dessus représentent les trois entités intéressantes à étudier. Les relations, dirigées, représentent une information sémantique pertinente entre les nœuds connectés.

Quelques attributs, tels que « name » pour l'entité « Site » et « url » pour l'entité « Url » permettent la définition des données. La relation « CONTAINS » est également identifiée par un attribut « date » qui permet de connaître à quelle date un mot est présent sur une page web, représentée elle par une URL.

La lecture de ce graphe se fait de la manière suivante :

- Une URL appartient à un site
- Une URL contient un mot à une certaine date
- Un mot est lié à un autre mot par la phrase dans laquelle on les retrouve

Lors du chargement des données dans Neo4J, chaque donnée (site web, url et mot) se verra attribuer l'un des trois labels représentés par les entités ci-dessus.

3.7 Architecture du travail

Cette section a pour objectif de décrire l'architecture de l'application Python permettant la récupération des données, leur nettoyage, leur stockage et la création du graphe dans Neo4J.

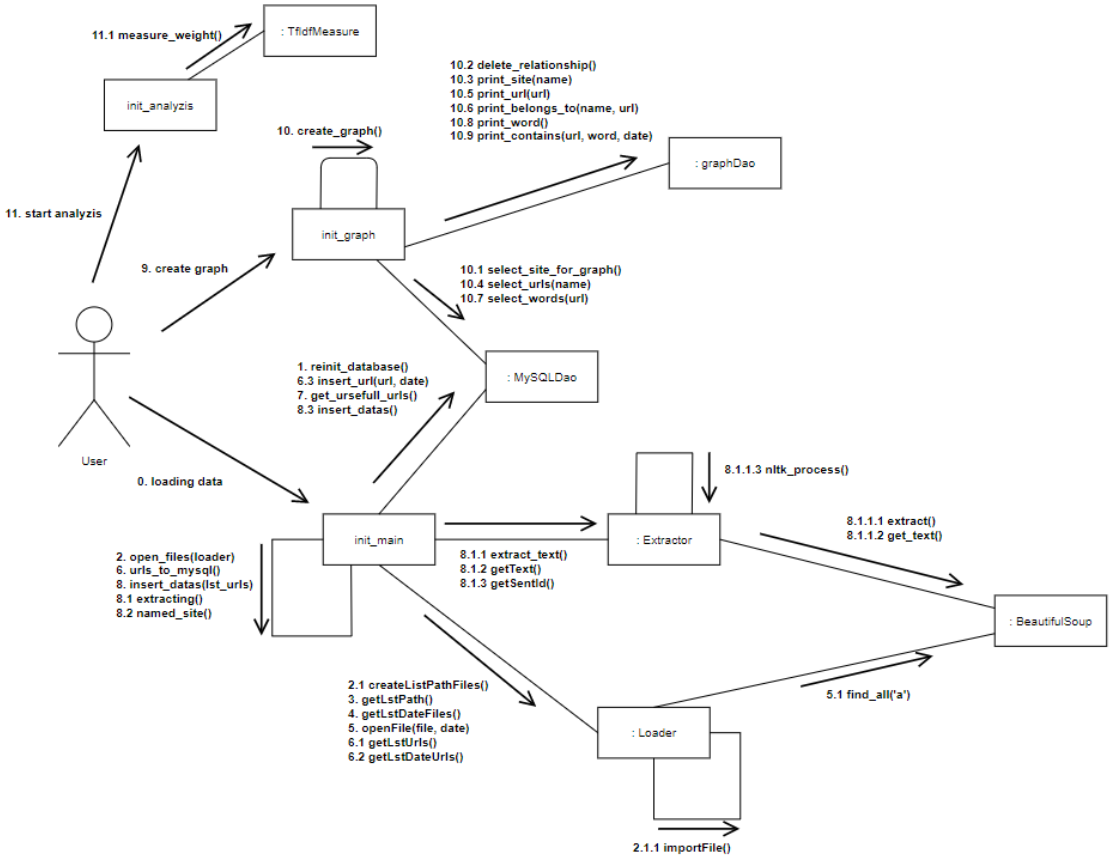
3.7.1 Description globale

L'architecture ainsi que le fonctionnement du code sont décrits par un diagramme de communication proposé par Unified Modeling Language (UML). Le diagramme de communication est une synthèse entre le diagramme de classe et le diagramme de séquence, tous deux également proposés par UML, permettant ainsi de représenter les appels de méthode entre les différentes instances de classe du programme.

(Samuel et al., 2007)

J’ai fait le choix de distinguer deux processus distincts. Un premier processus est responsable de récupérer les données depuis les fichiers HTML, en récupérer le texte et stocker le tout dans la table « datas » de la base de données (cf. Figure 14 : Table contenant toutes les données). Un deuxième est responsable de récupérer ces données et de les mettre en forme dans Neo4J.

Figure 16 : Diagramme de communication



3.7.2 Description détaillée

Init_main

Le fichier « init_main » est le fichier de lancement et point central du processus de récupération et stockage des données.

Init_graph

Le fichier « init_graph » est le fichier de lancement et point central du processus de mise en forme des données dans Neo4J.

Init_analyzis

Le fichier « init_analyzis » est le fichier démarrant le calcul du poids de chaque terme sur les différents sites présents dans la base de données.

MySQLDao

La classe « MySQLDao » contient toutes les méthodes nécessaires aux différentes requêtes dans la base de données. Chaque requête SQL est implémentée sous forme de procédure stockée dans cette classe.

GraphDao

La classe « GraphDao » contient toutes les méthodes nécessaires aux différentes requêtes à exécuter dans Neo4J. Ces requêtes, au même titre que « MySQLDao », sont implémentées sous forme de procédures stockées dans cette classe.

Loader

La classe « Loader » est responsable de lire tous les fichiers HTML enregistrés manuellement depuis un navigateur sous le répertoire de la constante « PATH ». Elle va également être responsable de retourner toutes les URLs retrouvées dans chaque fichier HTML, ainsi que leur date de consultation.

Extractor

La classe « Extractor » est responsable, quant à elle, de retourner le contenu texte de chaque URL. C'est dans cette classe que l'utilisation de la librairie NLTK, décrite dans la section *3.3.1 Phase de nettoyage par NLTK*, est nécessaire. En effet, la responsabilité de cette classe est également de nettoyer les données textes de tout le bruit présent dans le jeu de données.

BeautifulSoup

Contrairement aux autres éléments constituant le code de ce programme, la classe « BeautifulSoup » est un objet fourni par la librairie « bs4 » dans Python. Cette classe permet, entre autres, l'extraction de diverses informations depuis un fichier HTML.

TfIdfMeasure

La classe « TfIdfMeasure » est une classe utilisant la librairie « sklearn » de Python pour le calcul du poids d'un terme dans un document.

3.8 Calcul de pertinence

Le jeu de données étant construit, il s'agit maintenant de calculer la pertinence de chacun des documents (URLs) en analysant la fréquence des termes présents dans la page, permettant ainsi d'identifier quelle est la page la plus pertinente par rapport à un mot clé donné.

Pour ce faire, l'utilisation de la métrique « *Term Frequency-Inverse Document Frequency* » (TF-IDF) est choisie. Cette méthode de pondération permet d'évaluer l'importance d'un terme présent dans un document. Dans le cas de ce projet, elle permettra d'évaluer l'importance d'un mot présent sur une page web.

4. Résultats

L'objectif de ce travail était de concevoir un outil afin de fournir une prédiction de nouvelles tendances pour un sujet donné :

Ce chapitre présente les mesures obtenues lors du calcul de poids (cf 3.8 Calcul de pertinence) pour chaque mot présent sur un site web.

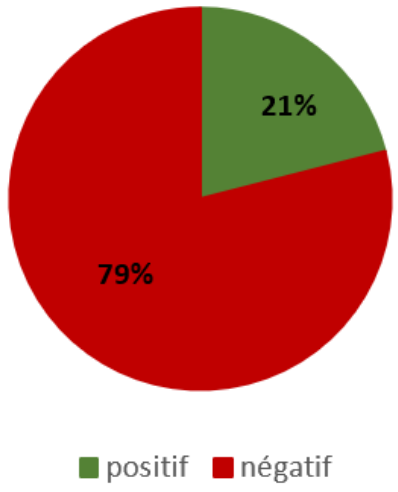
Une première partie de résultat présente la comparaison de l'évolution temporelle de l'importance d'un terme à travers différents sites. Une deuxième partie présente l'évolution temporelle de l'importance d'un mot au travers de tous les sites donnés. Et enfin une troisième partie présente l'évolution temporelle de l'importance d'un mot sur sites regroupés par catégories.

Le nombre de sites différents étant important (un peu plus d'une centaine), j'ai décidé d'exposer ici uniquement les mesures pour les sites ayant une certaine renommée dans la région genevoise et en Suisse Romande, à savoir le site de l'état de Genève, La Tribune de Genève ainsi que le site de la Radio Télévision Suisse (RTS)

L'axe des abscisses correspond à l'échelle de temps à laquelle le mot apparaît sur la page en question. L'axe des ordonnées représente le poids TF-IDF attribué au mot sur le site à une date donnée. Il est important de préciser que l'échelle des abscisses n'est pas équivalent d'un graphique à un autre. Un graphique étant propre à un site, il se peut que le mot n'apparaisse pas à chaque date identique entre les sites. La courbe montre ainsi l'évolution d'importance du terme entre une date d'apparition et la suivante.

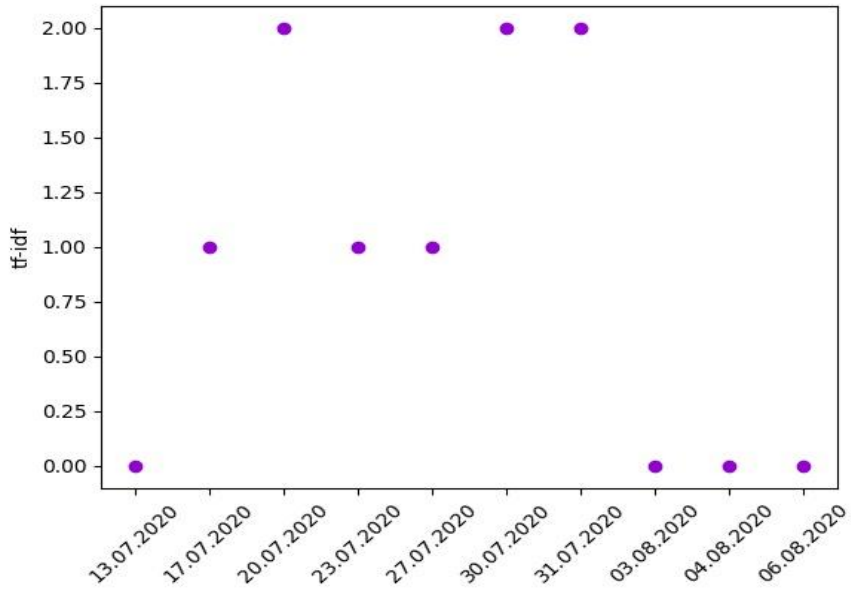
4.1 Mesures « Confinement »

Figure 17 Pourcentage de sites parlant de « Confinement »



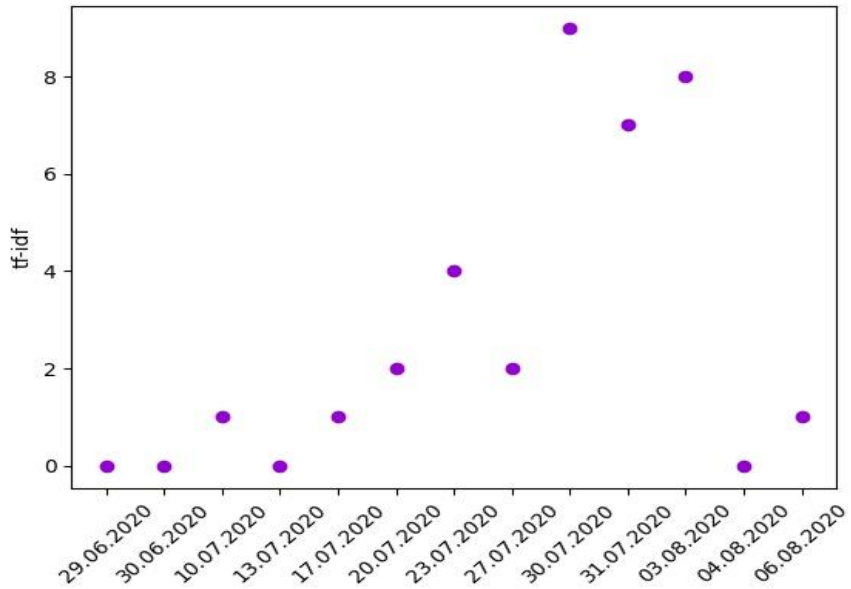
Sur les 119 sites différents présents dans la base de données, seuls 25 d'entre eux mentionnent le terme « Confinement ». Voici ci-dessous les mesures pour cinq sites parmi ces 25.

Figure 18 : « Confinement » sur www.ge.ch



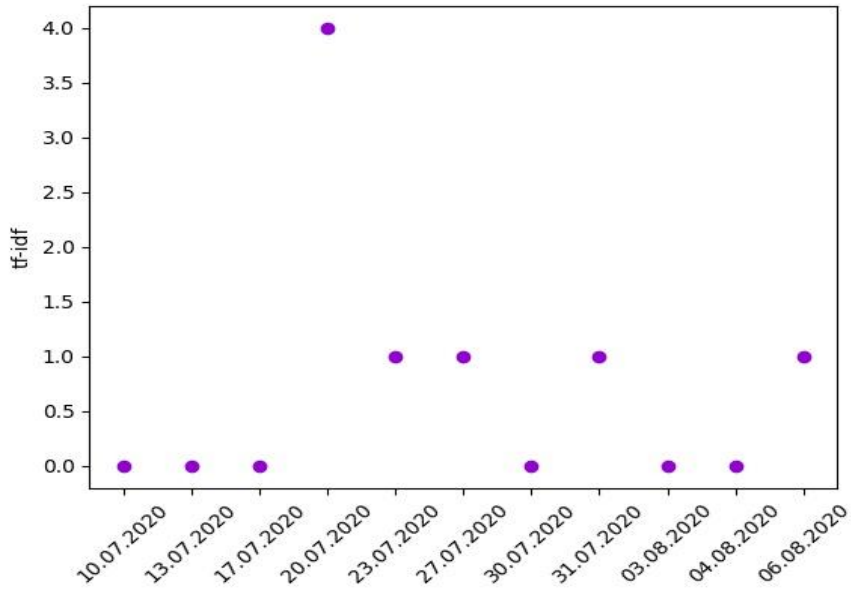
On remarque une hausse d'importance du terme « Confinement » sur le site de l'Etat de Genève à partir de la mi-juillet puis à nouveau entre fin juillet et début août avec une valeur maximale du TF-IDF à 2.00.

Figure 19 « Confinement » sur www.rts.ch



Un pic d'importance élevée du terme « Confinement » a eu lieu entre fin juillet et début août sur le site de la Radio Télévision Suisse avec une valeur maximale du TF-IDF supérieure à 8.00. Une hausse d'importance de ce terme est visible à partir des premiers jours du mois d'août.

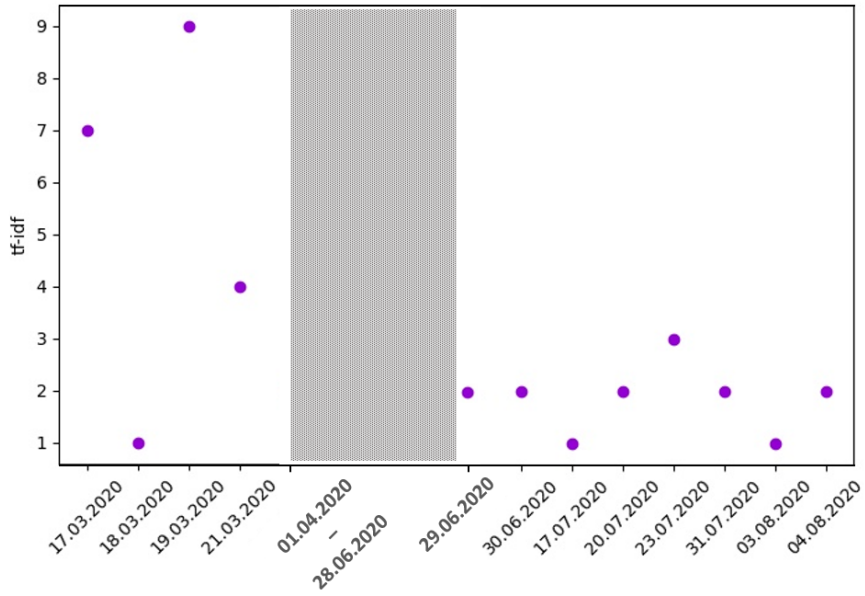
Figure 20 « Confinement » sur www.tdg.ch



Concernant le site de la Tribune de Genève, un pic d'importance élevée du terme « Confinement » a eu lieu entre la mi et fin juillet, contrairement aux deux sites genevois précédents dont l'importance a augmenté entre fin juillet et début août. Une valeur maximale du TF-IDF de 4.00 a été calculée pour ce terme au début du troisième tiers du mois de juillet. Une hausse d'importance de ce terme est visible à partir des premiers jours du mois d'août.

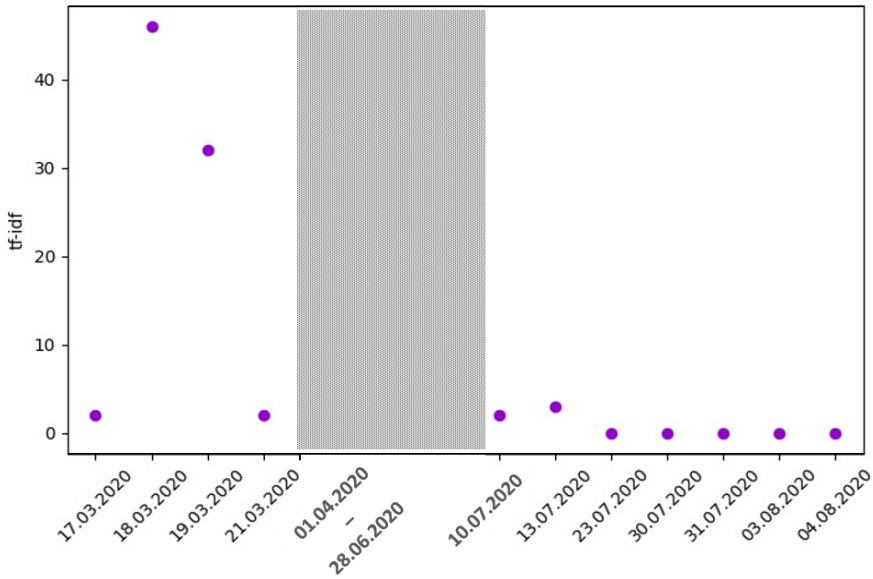
Les graphiques ci-dessus représentent donc l'évolution d'importance du terme « Confinement » sur trois sites médiatiquement importants dans la région genevoise. Les deux graphiques ci-dessous représentent les résultats pour deux sites médiatiques français.

Figure 21 « Confinement » sur www.lemonde.fr



Contrairement aux trois sites genevois, nous avons quelques résultats datant du mois de mars pour le site du journal français « Le Monde ». Durant ce troisième mois de l'année, une valeur de 9.00 de TF-IDF a été calculée pour le terme « Confinement » sur ce site. Puis, durant le troisième tiers du mois de juillet, un nouveau pic, pour une valeur de TF-IDF de 3.00, a été calculé. Une hausse d'importance de ce terme est visible à partir des premiers jours du mois d'août.

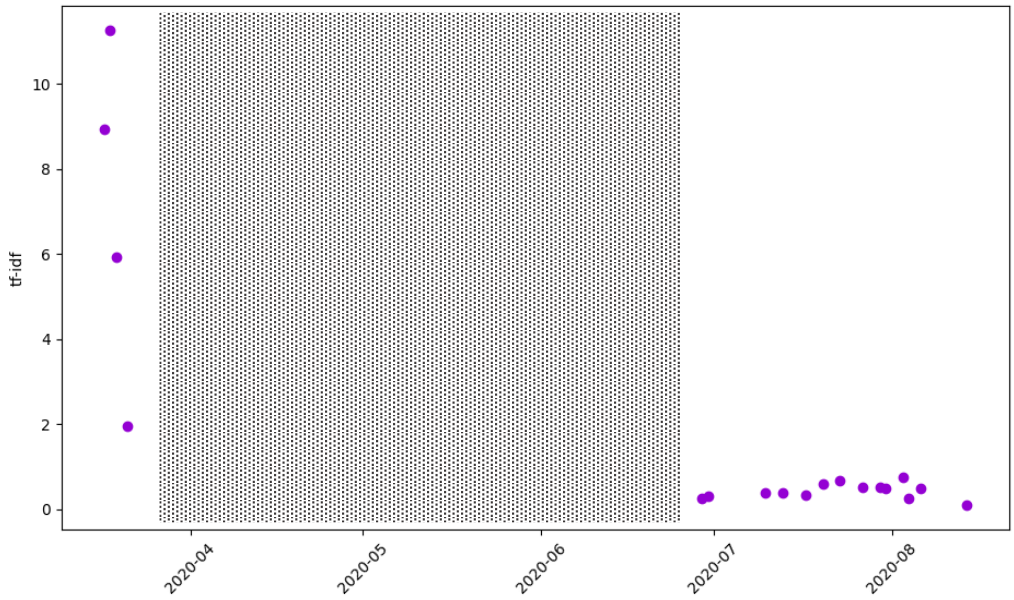
Figure 22 « Confinement » www.leparisien.fr



Une forte importance du terme « Confinement » a été calculée sur le site du journal français « Le Parisien » durant le mois de mars, avec une valeur de TF-IDF proche de

50. Puis, durant la première quinzaine du mois de juillet, une hausse d'importance, frôlant la valeur de 5.00 mais retombant à 0 jusqu'à début août.

Figure 23 : Moyenne du terme « Confinement »



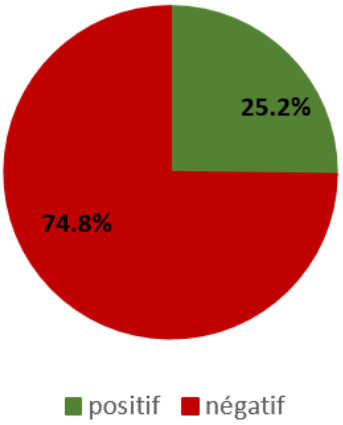
Le graphique ci-dessus montre une importance moyenne, calculée sur l'ensemble des sites présents dans la base de données, supérieure à 10.00 lors du mois de mars, puis variant entre 1.00 et 2.00 entre juillet et août.

Notre outil permet donc d'observer la « vie médiatique » du terme « Confinement ».

4.2 Mesures « Pandémie »

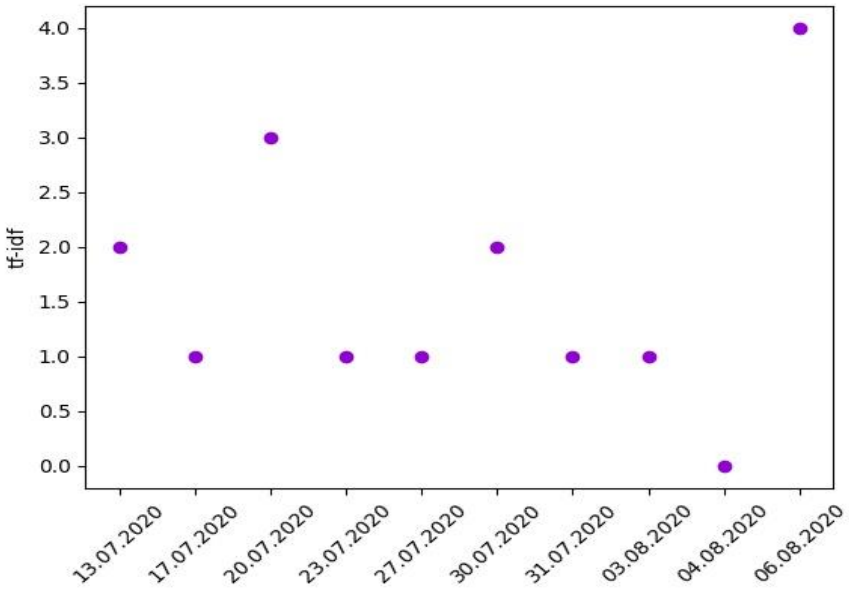
Afin de garder une certaine cohérence, les résultats présentés dans cette section correspondent aux mêmes sites mais pour le terme « Pandémie ».

Figure 24 Pourcentage de sites parlant de « Pandémie »



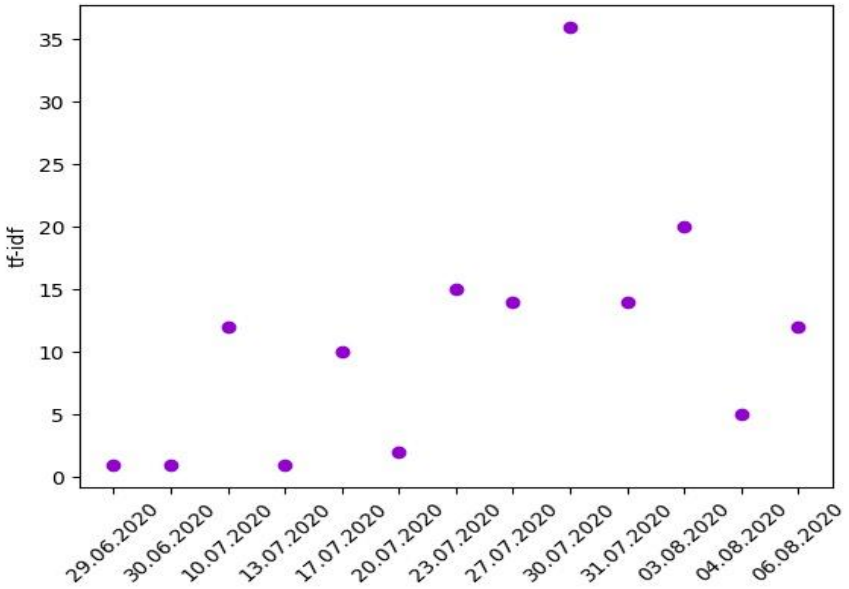
Sur les 119 sites différents présents dans la base de données, seuls 30 d'entre eux mentionnent le terme « Pandémie », ce qui est légèrement plus élevé que pour le terme « Confinement ».

Figure 25 « Pandémie » sur www.ge.ch



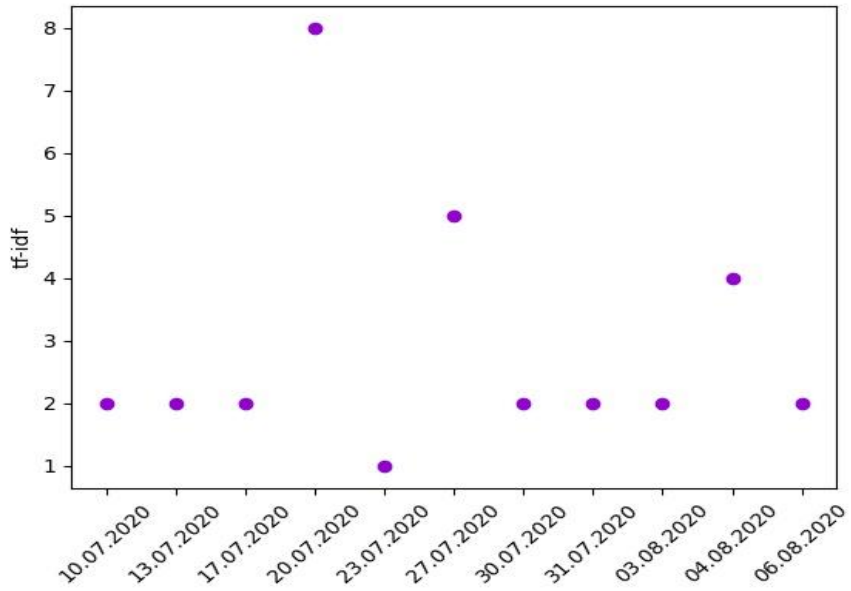
L'importance du terme « Pandémie » n'est pas tombée à 0.00 sur le site de l'Etat de Genève, hormis le 4 août précisément, pour ensuite remonter à une valeur de 4.00, valeur qu'elle n'a pas atteint le reste de l'été.

Figure 26 « Pandémie » sur www.rts.ch



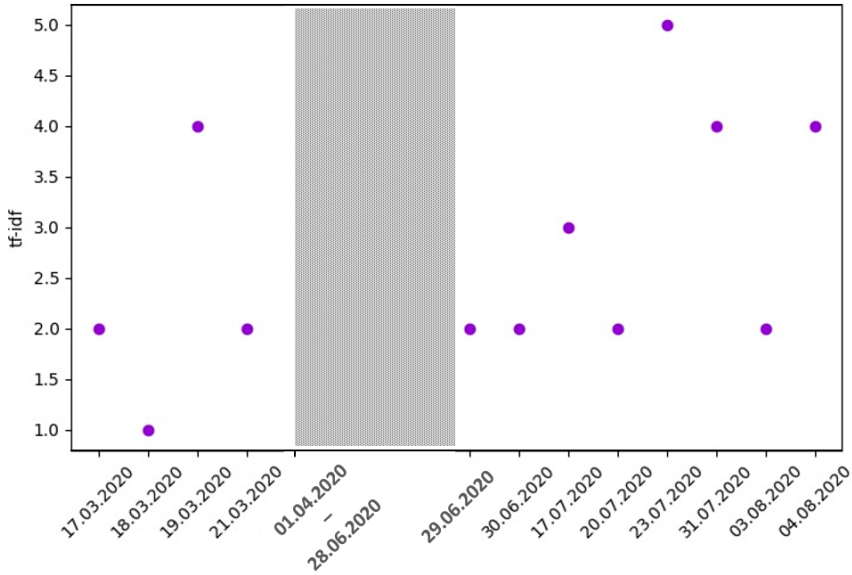
Sur le site de la Radio Télévision Suisse, ce terme a atteint une importance d'une valeur supérieure à 35.00 à la fin juillet. L'importance de ce terme sur le site de la RTS a été largement supérieure à celle du site de l'Etat de Genève durant tout l'été.

Figure 27 « Pandémie » sur www.tdg.ch



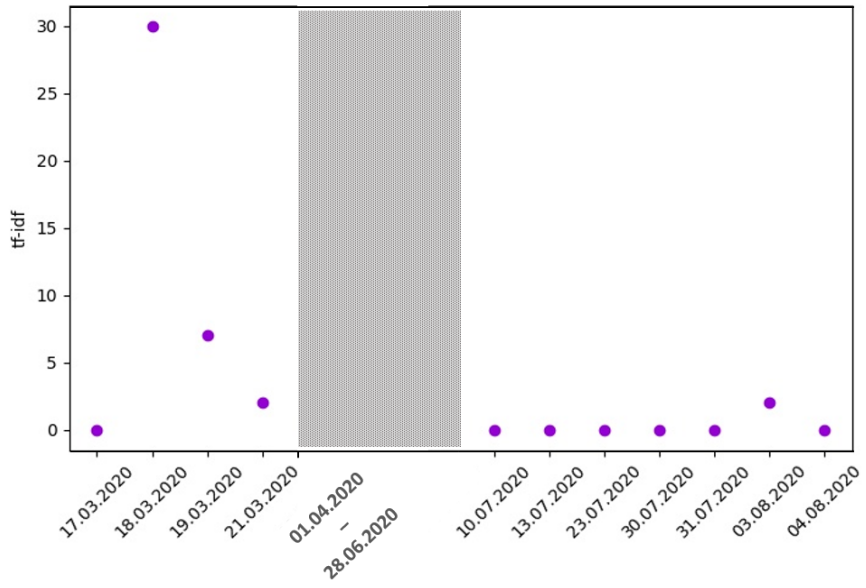
En ce qui concerne le site de la Tribune de Genève, un pic d'importance a été calculé au début du troisième tiers du mois de juillet avec une valeur de TF-IDF à 8.00.

Figure 28 « Pandémie » sur www.lemonde.fr



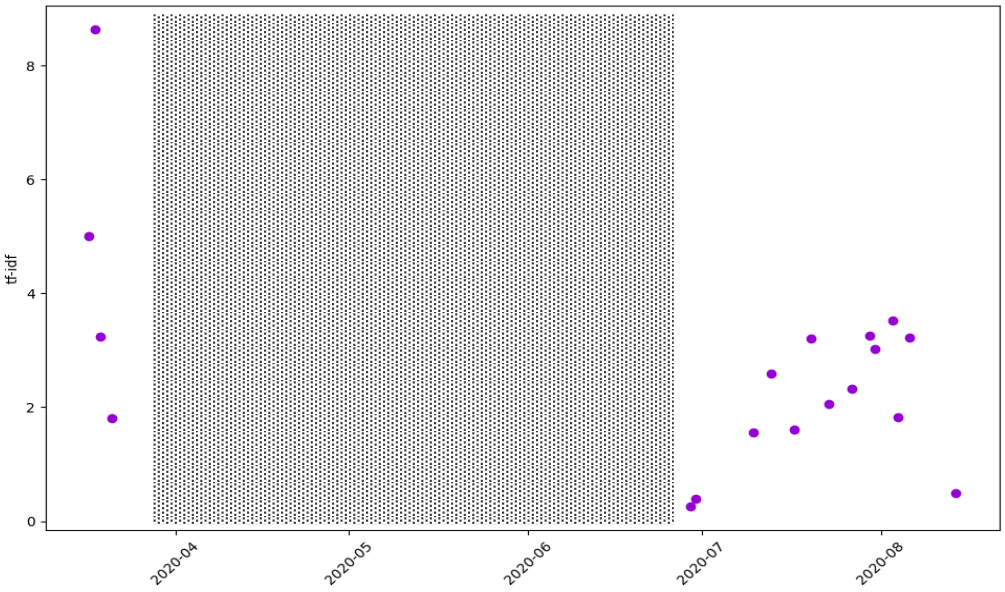
Pour le site du journal français « Le Monde », on remarque un pic important entre la mi et fin mars, pour ensuite conserver une valeur entre 2.00 et 5.00 durant l'été.

Figure 29 « Pandémie » sur www.leparisien.fr



Le terme « Pandémie » a eu une importance élevée à la valeur de 30.00 aux alentours de mi-mars, pour ensuite conserver une valeur à 0 entre juillet et août.

Figure 30 Moyenne du terme « Pandémie »



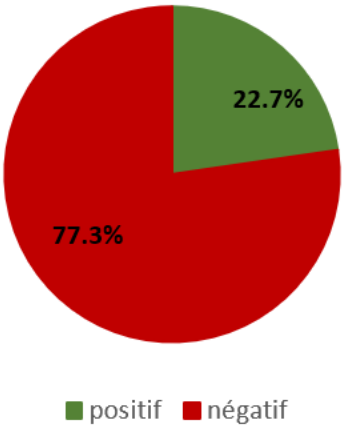
En moyenne, ce terme a eu une importance supérieure à 8.00 durant le mois de mars, puis entre 2.00 et 4.00 entre juillet et août, ce qui est plus élevé que la moyenne du terme « Confinement » durant la même période estivale.

Notre outil permet donc d'observer la « vie médiatique » du terme « Pandémie ».

4.3 Mesures « Epidémie »

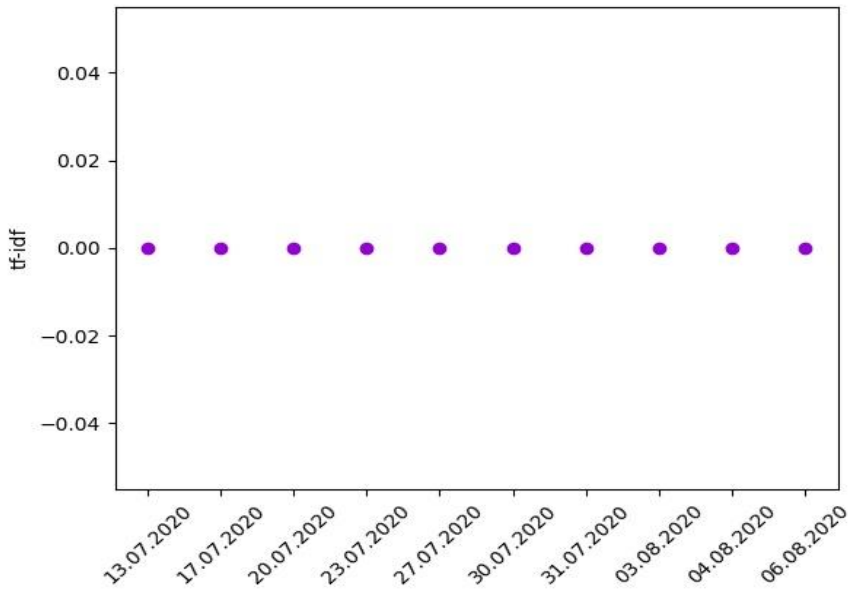
Afin de faire un lien avec le terme précédent, cette section présente les résultats du terme « Epidémie » toujours sur les mêmes sites que les résultats présentés précédemment.

Figure 31 Pourcentage de sites parlant de « Epidémie »



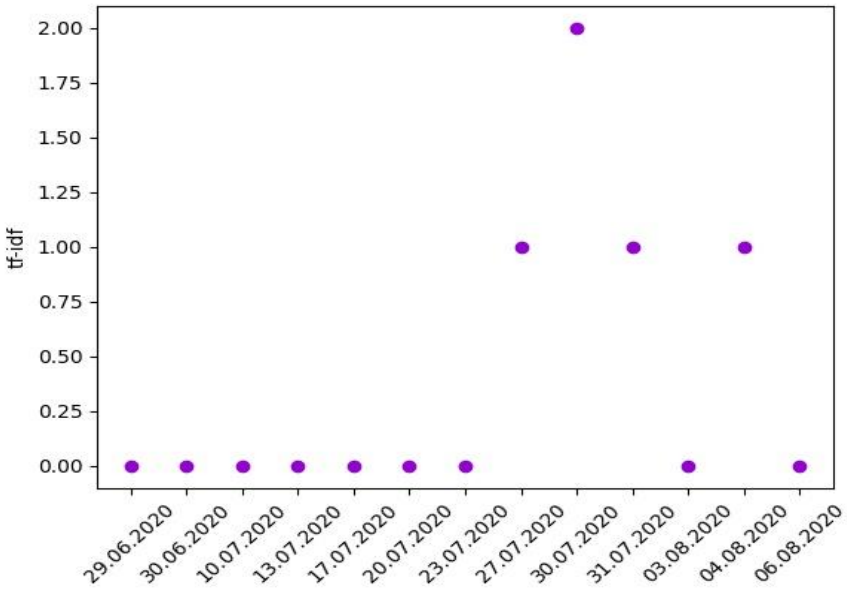
Le terme « Epidémie » a été mentionné par 27 sites sur les 119 présents dans la base de données, ce qui reste dans le même ordre de grandeur que les deux termes précédents.

Figure 32 : « Epidémie » sur www.ge.ch



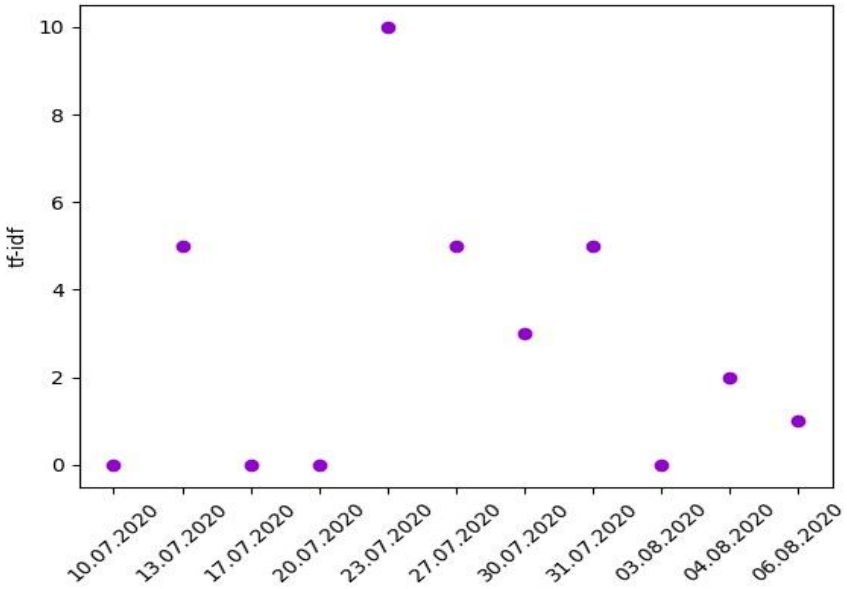
Il est intéressant de constater que ce terme, pourtant fortement lié au terme « Pandémie » étymologiquement, n'est pas apparu sur le site de l'Etat de Genève durant l'été. Un poids constant de 0.00 a été calculé.

Figure 33 « Epidémie » sur www.rts.ch



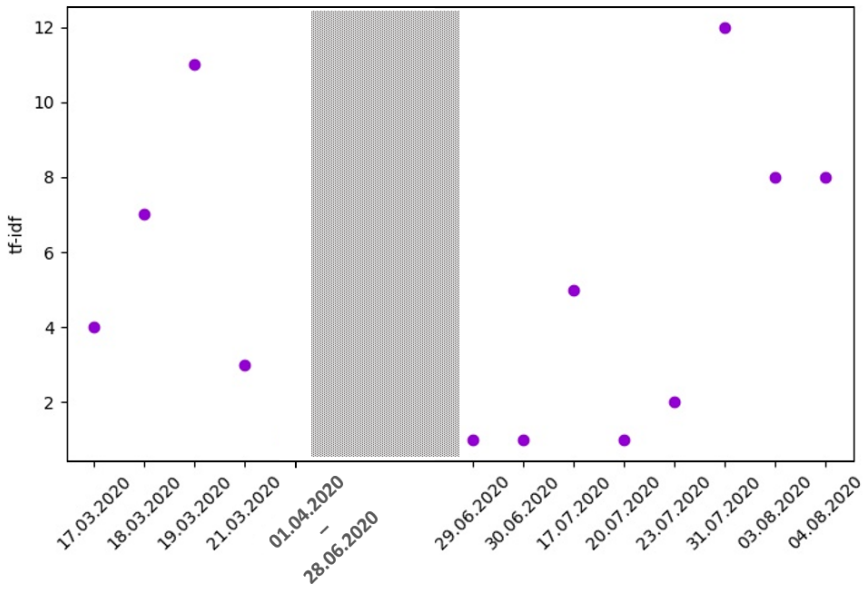
Même constat que pour le site de l'Etat de Genève, l'importance du terme « Epidémie » sur le site de la Radio Télévision Suisse est presque constant à 0.00 durant l'été, hormis un pic d'une valeur de 2.00 fin juillet et une valeur de 1.00 début août.

Figure 34 « Epidémie » sur www.tdg.ch



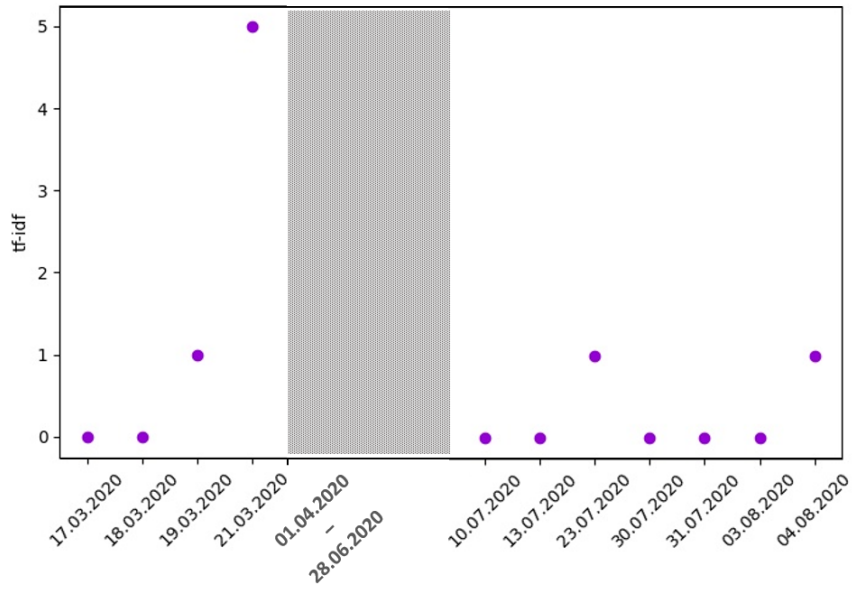
L’importance du terme « Epidémie » varie cependant plus sur le site de la Tribune de Genève, par rapport à ses deux précédents compatriotes, avec un poids maximal à 10.00 durant le troisième tiers du mois de juillet, pour ensuite descendre à 6.00 fin juillet et enfin 2.00 début août.

Figure 35 « Epidémie » sur www.lemonde.fr



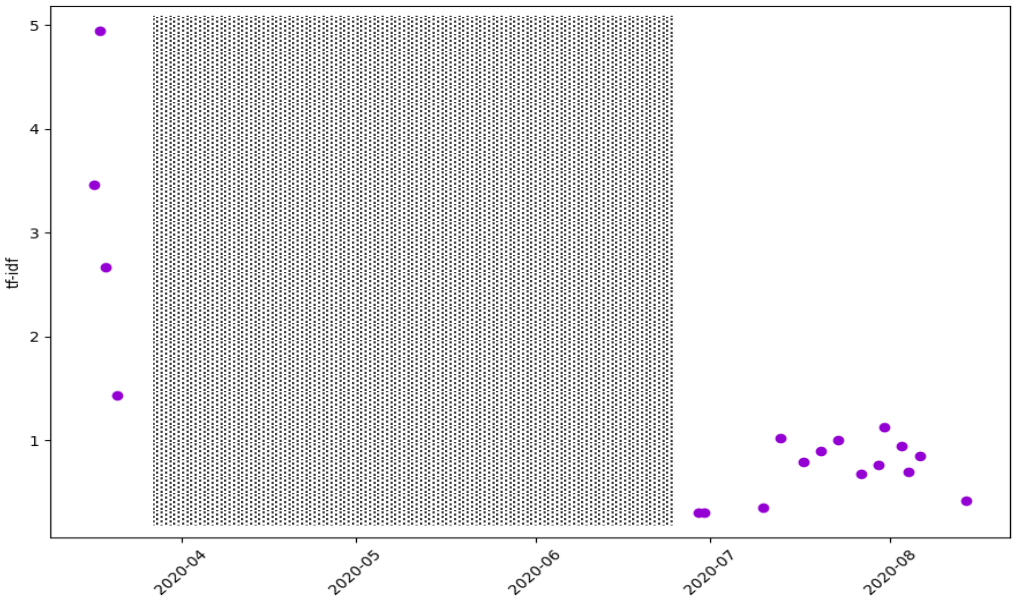
Avec un poids proche de 12.00 à la mi-août, le terme « Epidémie » perd en importance durant le début de l’été, avec tout de même un poids de 5.00 à la mi-juillet, pour ensuite atteindre une valeur de 12.00 entre fin juillet et début août.

Figure 36 « Epidémie » sur www.leparisien.fr



Un poids de 5.00 a été calculé à la mi-mars pour le site français « Le Parisien », puis de 1.00 à la mi/fin juillet, enregistrant ensuite une hausse à partir de début août.

Figure 37 Moyenne du terme « Epidémie »



Suivant la courbe des deux moyennes précédentes, l'importance du terme « Epidémie » s'élève à 5.00 durant le mois de mars, puis varie entre 1.00 et 2.00 entre juillet et août.

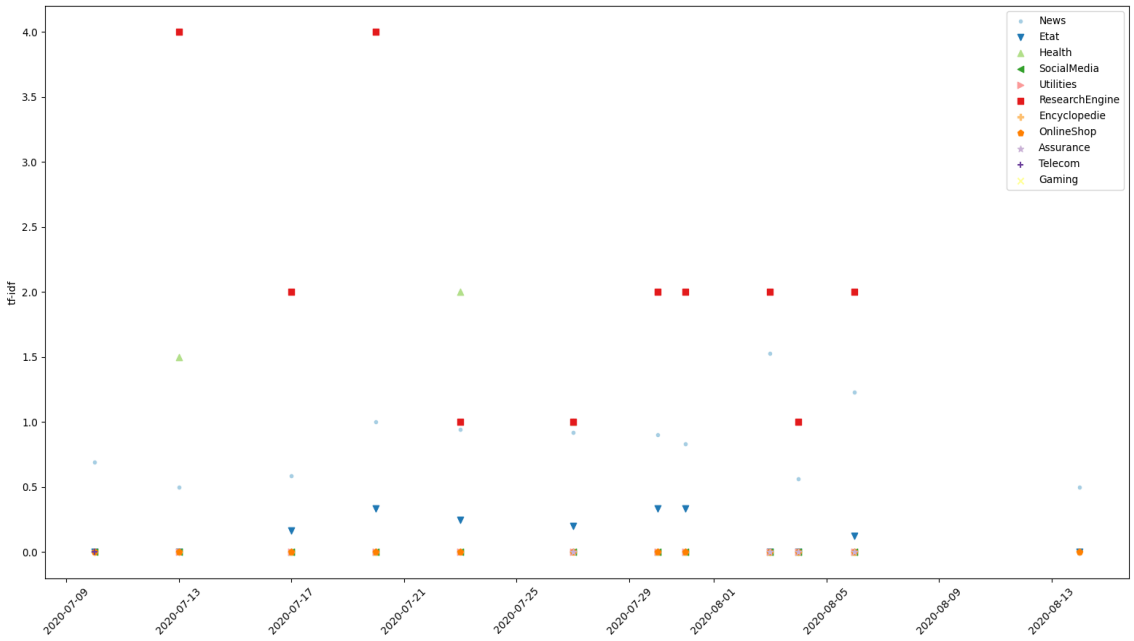
Notre outil permet donc d'observer la « vie médiatique » du terme « Confinement ».

4.4 Mesures globales

Les résultats présentés ci-dessous illustrent, en moyenne, pour chacun de ces trois mots l'évolution de leur importance par catégorie de site (cf 3.4,Regroupement des données).

4.4.1 Confinement

Figure 38 : Catégorisation de « Confinement »



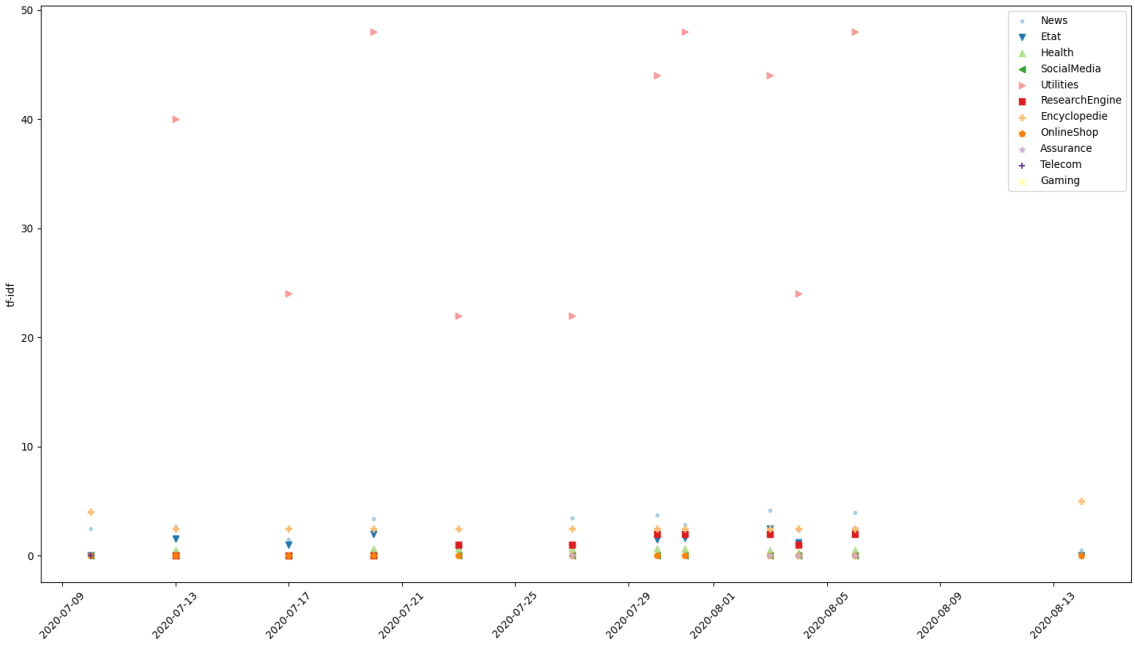
Sur les 11 catégories de sites créées, seules quatre d'entre-elles ont pu être utilisées pour le terme « Confinement ». Ces quatre catégories sont :

- News
- Health
- Etat
- Research Engine

Le poids moyen calculé pour la catégorie « Research Engine » monte jusqu'à 4.00 durant mi-juillet, s'élève à 2.00 pour la catégorie « Health » au début du dernier tiers de juillet, varie entre 0.5 et 1.5 pour les « News » entre mi-juillet et début août et entre 0 et un peu moins de 0.5 pour les sites étatiques durant la même période.

4.4.2 Pandémie

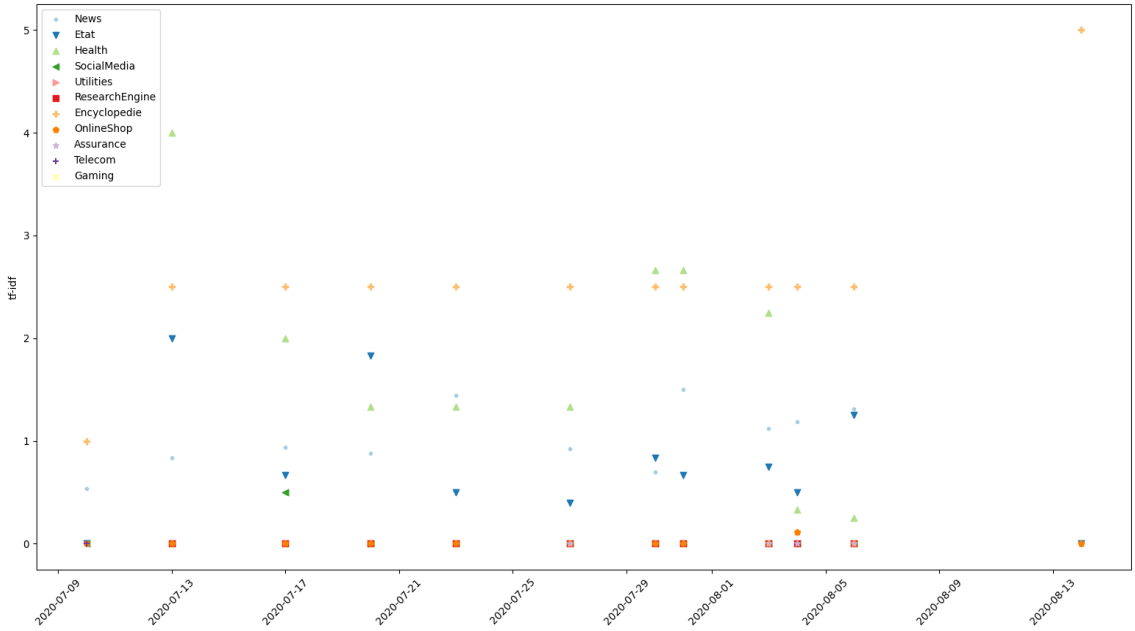
Figure 39 : Catégorisation de « Pandémie »



Pour six catégories sur 11, un poids moyen pour le terme « Pandémie » a pu être calculé. Un poids moyen variant entre 20.0 et environ 50.0 a été calculé pour les sites de la catégorie « Research Engine », tandis que pour les sites des cinq catégories restantes, le poids se situe entre 1.0 et 5.0 durant l'été, avec une hausse concernant le poids moyen calculé pour les sites de la catégorie « Encyclopédie ».

4.4.3 Epidémie

Figure 40 : Catégorisation de « Epidémie »



Idem que pour le terme « Pandémie », un poids moyen pour le terme « Epidémie » a pu être calculé pour les sites appartenant à six catégories sur 11. Deux de ces six catégories ont un poids moyen calculé ponctuellement, à savoir 0.5 à mi-juillet pour les « Social Media » et approximativement 0.2 début août pour les « Online Shop ».

L'importance du terme « Epidémie » pour les sites étatiques et les « News » varie globalement entre 0.5 et 2 durant l'été, tandis que pour les sites ayant pour thématique la santé (« Health ») elle varie entre 1.5 et 3 durant la deuxième moitié du mois de juillet, puis tend à redescendre en dessous de 1 à partir de début août. Les sites encyclopédiques voient une constance d'importance du terme « Epidémie » à 2.5 entre mi-juillet et début août, pour ensuite accroître à 5.00 jusqu'à mi-août.

4.4.4 Résultats

Globalement, l'importance du terme confinement a eu une tendance plus élevée à partir de début juillet mais tend à diminuer sur la fin de l'été.

En ce qui concerne le terme « Pandémie », son importance est plutôt groupée durant toute la période estivale. Il peut être conclu que les différentes catégories de média s'accordent sur une importance similaire concernant le terme « Pandémie ».

En revanche, pour le terme « Confinement », son importance entre les différentes catégories de média durant l'été est plus dispersée.

Selon les résultats obtenus, notre outil permet de suivre, dans une certaine mesure, l'évolution de la « vie numérique » d'un terme au sein d'internet. Le chapitre suivant, *5 Limites et perspectives*, présente de façon plus détaillée les limites et les perspectives de ce projet. Ce qu'on peut retenir pour le moment, c'est que chaque terme à une « vie numérique » propre à chaque média, bien que celle-ci varie en fonction du média, du terme ainsi que de l'instant précis.

5. Limites et perspectives

Ce chapitre présente les limites de ce projet, ainsi que les perspectives à lui donner afin d'enrichir les résultats obtenus.

5.1 Limites

A mon sens, les données et résultats obtenus au terme de ce travail ne permettent pas de définir une quelconque marque d'autorité de la part d'un site par rapport à un autre. Le principal problème provient du fait que les données récoltées soient réparties dans le temps et soient récoltées manuellement. Cette façon de faire laisse place à des inégalités temporelles, comme constaté lors de la présentation des résultats. Certains écarts temporels étant importants, c'est-à-dire une période sans avoir de données à analyser dû à ma mobilisation à la Protection Civile genevoise entre avril et juin, il est très difficile de pouvoir tirer des conclusions quant à l'émergence d'un nouveau terme du domaine.

L'idée serait de concevoir un outil effectuant cette récolte tous les jours automatiquement. Un tel outil aurait pallié ce manque de régularité de récolte manuelle et aurait eu l'avantage de pouvoir présenter des résultats plus pertinents et plus homogènes.

Pour continuer sur le thème de la récolte des données, les premières données récoltées datent du mois du mars, soit du début de la pandémie du Covid-19, puis à partir du mois de juillet. Cela signifie qu'un certain retard a été pris lors de la récolte des données sur le thème du Covid-19, étant donné qu'au mois de juillet cette pandémie était déjà au centre des discussions sur le web. Les données de ce printemps auraient été plus intéressantes à analyser compte tenu de la nouveauté du sujet sur Internet.

Concernant les limites physiques (liées à mon ordinateur personnel) du projet, l'extraction du texte de chaque lien URL et l'insertion des données ont pris beaucoup de temps. Le script a tourné durant trois heures afin de parcourir les 3'000 URLs présentes sur les fichiers HTML récupérés. De plus, le temps de calcul du poids TF-IDF pour un terme sur tous les sites varie entre 40 et 50 minutes. L'accès à distance à une machine plus performante que mon ordinateur personnel aurait aidé à réduire ces différents temps de calculs.

Une limite certes imposée mais non négligeable fut celle du temps consacré à ce projet. Etant limité sur une période de trois mois, un certain nombre de concepts n'ont pas pu être mis en place.

5.2 Perspectives

Les sites étant catégorisés, il peut paraître intéressant d'affiner cette classification, surtout concernant la catégorie « Social Media » qui correspond principalement aux réseaux sociaux. En effet, l'analyse des termes s'est faite uniquement sur leurs nombres d'occurrences sur un site web. Etudier la « vie » d'un terme au sein des réseaux sociaux peut être un axe d'analyse intéressant étant donné que ces médias ont une place importante dans nos vies de tous les jours.

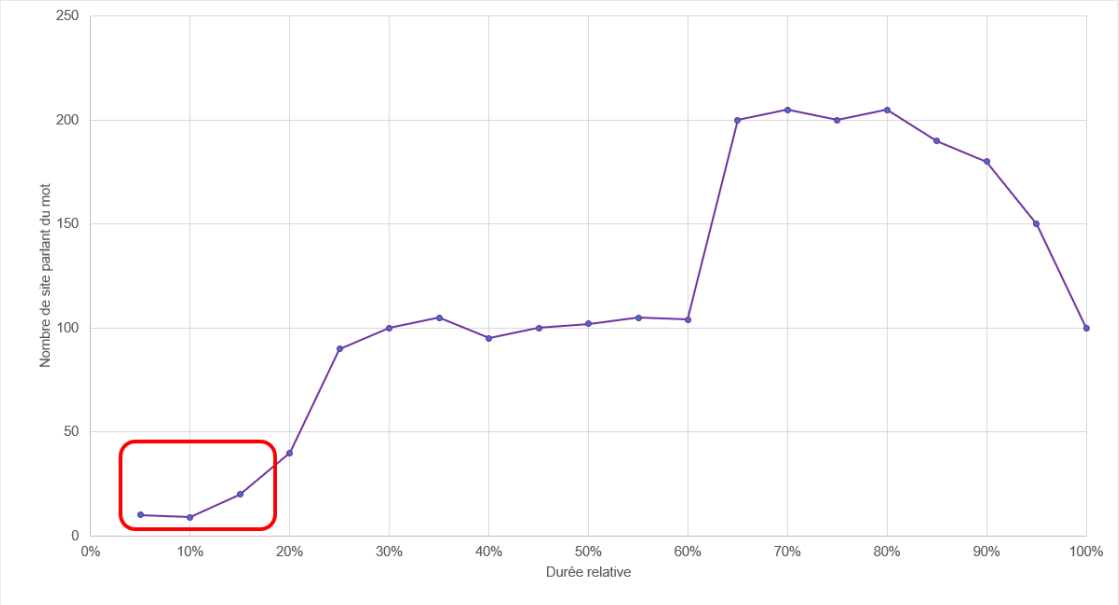
Les tendances démontrées dans la partie résultat sont basées sur un jeu de données relativement petit. Il serait intéressant d'augmenter la taille du jeu de données afin de confirmer ou non ces tendances.

En termes de récolte des données, le développement (ou l'utilisation d'un outil existant) d'un outil permettant une récolte de données automatisée à intervalle de temps régulier permettrait d'être plus précis dans l'analyse de celles-ci. Il faudrait également être capable de récolter des données rétroactivement afin d'obtenir une vision de la « vie » du terme à analyser.

Concernant la prédiction de la « vie numérique » future d'un terme, il faudrait pouvoir mettre en place un réseau de neurones artificiels, utilisant du *machine learning*, pour pouvoir être capable de prédire l'évolution de la « vie » de ce terme dans le futur. Un pré-traitement qui a été pensé mais qui n'a pas pu être mis en place par manque de temps, est le fait d'attribuer un poids aux sites parlant d'un mot à un certain intervalle de temps puis de répéter l'opération sur des termes différents afin de définir quel site attribue le plus d'importance à ces termes durant cet intervalle de temps. Idéalement, cet intervalle devrait s'arrêter au moment où de plus en plus de sites commencent à donner de l'importance à ce terme.

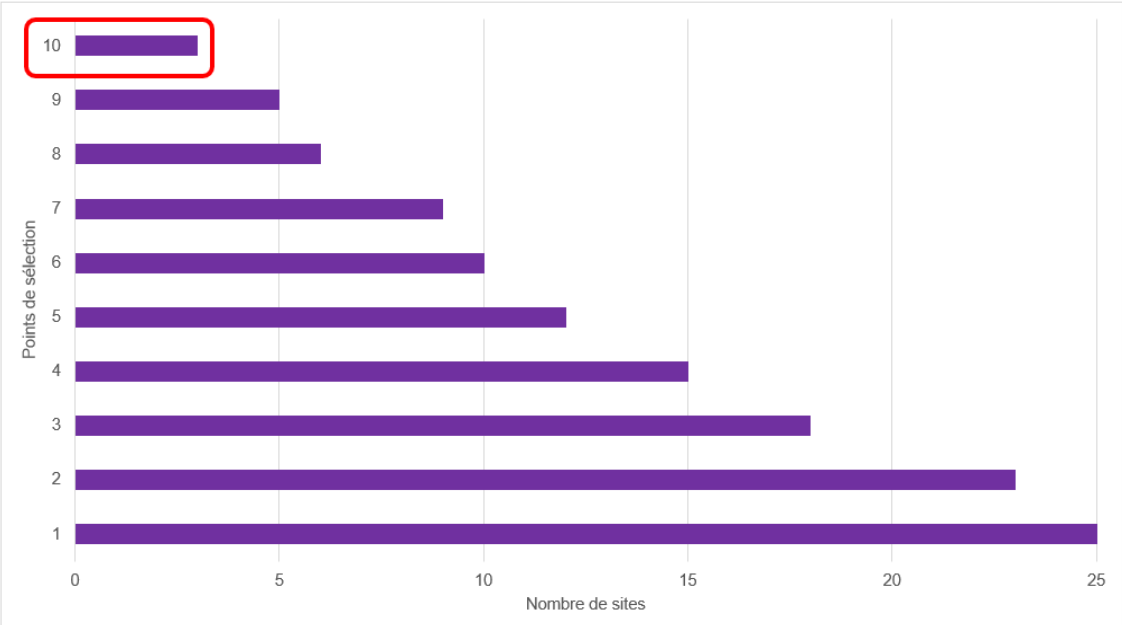
Les deux figures ci-dessous illustrent des exemples de graphiques pouvant être mis en place avec les explications données ci-dessus :

Figure 41 : Exemple d'évolution du nombre de sites parlant d'un terme



Sur le graphique ci-dessus, la zone encadrée en rouge correspond à la période estimée prédictive. Tous les sites parlant de ce mot durant cette période se voient attribué un point, les autres zéro point. Cette opération se répète sur plusieurs mots en cumulant les points obtenus pour chaque site pour pouvoir faire ressortir les sites les plus prédictifs selon un ensemble de mots.

Figure 42 : Représentation des sites ayant obtenu le plus de points



La figure ci-dessus présente le nombre de sites ayant obtenus *x* points. L'idée serait de conserver les sites ayant obtenus un maximum de points (encadré en rouge sur la figure) et de se baser sur ces sites, considérés comme prédictifs, pour tenter de prédire le comportement du terme dans le futur.

Pour entraîner un réseau de neurones artificiels à faire ce travail, il faudrait également avoir une base d'événements passés et donner au réseau les sites ayant été définis comme prédictifs peu de temps avant la survenance de l'événement.

Un axe d'analyse intéressant à développer, est celui de la signification du texte présent sur une page web. Dans le cadre de ce travail, nous n'avons analysé que la présence des mots sur la page. Un mot peut être présent dans deux phrases différentes ayant un sens complètement opposé. Il pourrait donc être intéressant d'utiliser des outils de traitement du langage naturel afin d'analyser les émotions et les sentiments présents sur le contenu du site.

6. Conclusion

Le travail fourni dans le cadre de ce projet avait pour objectif de répondre aux problématiques mentionnées dans la section *1.2 Problématique* en développant un outil capable de récupérer des données depuis le web et d'analyser l'évolution de la « vie numérique » d'un terme.

Le développement d'un *web crawler* permet de créer un jeu de données basé sur des données récupérées sur le web. Le jeu de données se base sur le thème de la pandémie du Covid-19 présente dans le monde entier depuis la fin de l'hiver 2020. Présentant un contenu médiatique différent jour après jour, ce thème paraissait pertinent pour une analyse temporelle.

Lorsque les données ont été récoltées après plusieurs jours, l'utilisation de la mesure « *Term Frequency-Inverse Document Frequency* » a permis d'évaluer l'importance de certains termes faisant partie du domaine du Covid-19 à travers le temps. La « vie numérique » d'un terme a ainsi pu être évaluée selon le laps de temps des données récoltées.

La mise en place d'une base de données NoSQL orientée graphe, à savoir Neo4J, a permis de modéliser temporellement la « vie médiatique » d'un terme propre à un domaine donné.

Dans l'état actuel, l'outil développé ne permet pas de construire une prédiction de comportement de cette entité à court ou moyen terme. Pour ce faire, il faudrait explorer les différentes pistes exposées dans la section *5.2 Perspectives*. Les données récoltées ayant un intervalle temporel irrégulier et ne remontant pas assez dans le temps, la prédiction d'un comportement futur est compromise.

Au vu des résultats obtenus et décrits dans la section *4 Résultats* et des pistes d'amélioration exposées dans la section *5.2 Perspectives*, je reste confiant sur le fait qu'une suite donnée à ce projet puisse aboutir à de meilleures précisions de la modélisation de la « vie numérique » d'une entité ainsi qu'à une prédiction à court, moyen voire long terme du comportement de cette entité.

La récolte de données, leur nettoyage ainsi que leur analyse jouant un rôle primordial dans le milieu des Sciences de l'Information, ce travail m'a permis de mettre en pratique certaines de mes compétences en tant qu'informaticien de gestion tout en créant un lien avec les Sciences de l'Information.

Bibliographie

AMITAY, Einat, CARMEL, David, HERSCOVICI, Michael, LEMPEL, Ronny et SOFFER, Aya, 2004. Trend detection through temporal link analysis. In : *Journal of the American Society for Information Science and Technology*. décembre 2004. Vol. 55, n° 14, p. 1270-1281. DOI 10.1002/asi.20082.

BJÖRNEBORN, Lennart et INGWERSEN, Peter, 2004. Toward a basic framework for webometrics. In : *Journal of the American Society for Information Science and Technology*. 2004. Vol. 55, n° 14, p. 1216-1227. DOI 10.1002/asi.20077.

CHRISTIAN, Hans, AGUS, Mikhael Pramodana et SUHARTONO, Derwin, 2016. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). In : *ComTech: Computer, Mathematics and Engineering Applications*. 31 décembre 2016. Vol. 7, n° 4, p. 285. DOI 10.21512/comtech.v7i4.3746.

COADIC, Le et F, Yves, 2005. Mathématique et statistique en science de l'information et en science de la communication: infométrie mathématique et infométrie statistique des revues scientifiques. In : *Ciência da Informação*. décembre 2005. Vol. 34, n° 3, p. 15-22. DOI 10.1590/S0100-19652005000300002.

COOLEY, R., MOBASHER, B. et SRIVASTAVA, J., 1997. Web mining: information and pattern discovery on the World Wide Web. In : *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*. S.l. : s.n. novembre 1997. p. 558-567.

HEILBRON, Johan, 2002. La bibliométrie, genèse et usages. In : *Actes de la recherche en sciences sociales*. 2002. Vol. n° 141-142, n° 1, p. 78-79.

IBEKWE-SANJUAN, FIDELIA, 2007. *Fouille de textes : méthodes, outils et applications*. Hermès Science. Paris : s.n. ISBN 978-2-7462-1609-9.

LIBEN-NOWELL, David et KLEINBERG, Jon, 2007. The link-prediction problem for social networks. In : *Journal of the American Society for Information Science and Technology*. mai 2007. Vol. 58, n° 7, p. 1019-1031. DOI 10.1002/asi.20591.

LOPER, Edward et BIRD, Steven, 2002. NLTK: The Natural Language Toolkit. In : *arXiv:cs/0205028* [en ligne]. 17 mai 2002. [Consulté le 1 octobre 2020]. Disponible à l'adresse : http://arxiv.org/abs/cs/0205028.

NEO4J, 2020a. Cypher Query Language. In : *Neo4j Graph Database Platform* [en ligne]. 20 septembre 2020. [Consulté le 27 septembre 2020]. Disponible à l'adresse : https://neo4j.com/developer/cypher/.

NEO4J, 2020b. What is a Graph Database? In : *Neo4j Graph Database Platform* [en ligne]. 20 septembre 2020. [Consulté le 20 septembre 2020]. Disponible à l'adresse : https://neo4j.com/developer/graph-database/.

PHILIPPE LACOMME, CHRISTIAN PRINS et MARC SEVAUX, 2003. *Algorithmes de graphes*. Eyrolles. Paris : s.n. ISBN 2-212-11385-4.

SAMUEL, Philip, MALL, Rajib et KANTH, Pratyush, 2007. Automatic test case generation from UML communication diagrams. In : *Information and Software Technology*. février 2007. Vol. 49, n° 2, p. 158-171. DOI 10.1016/j.infsof.2006.04.001.

SEN, B K, 2004. - CYBERMETRICS MEANING, DEFINITION, SCOPE AN~CONSTITUENTS. In : . 2004. Vol. 51, n° 3, p. 5.

SURAUD, Marie-Gabrielle, 1996. La scientométrie : une méthode d'évaluation de la recherche ? In : *Communication et organisation* [en ligne]. 1 novembre 1996. n° 10. [Consulté le 20 septembre 2020]. DOI 10.4000/communicationorganisation.1881. Disponible à l'adresse : http://journals.openedition.org/communicationorganisation/1881.

THELWALL, Mike, 2001. A web crawler design for data mining. In : *Journal of Information Science*. 1 octobre 2001. Vol. 27, n° 5, p. 319-325. DOI 10.1177/016555150102700503.

WANG, Peng, XU, Baowen, WU, Yurong et ZHOU, Xiaoyu, 2014. Link Prediction in Social Networks: the State-of-the-Art. In : *arXiv:1411.5118 [physics]* [en ligne]. 8 décembre 2014. [Consulté le 21 juillet 2020]. Disponible à l'adresse : http://arxiv.org/abs/1411.5118.

7. Annexes

7.1 Nettoyage des URLs

Figure 43 : Requête SQL retournant les liens pertinents à l'analyse

```
-- Requête retournant les URLs ne contenant pas le mot clé "google"
-- excluant ainsi les liens redirigeant vers des pages de paramètres ou de compte Google, par exemple
select t1.id, t1.url, t1.date_consultation from Site as t1
    left join Site as t2 on t1.url = t2.url and lower(t2.url) like '%.google.%'
    where t2.id is null;
```

Figure 44 : Résultat de la requête SQL

id	url	date_consultation
44	https://www.20minutes.fr/monde/2741819-20200317-coronavirus-direct-france-entre-confinement-general-lundi-midi	17.03.2020
45	https://www.lemonde.fr/planete/article/2020/03/17/coronavirus-le-point-sur-les-interdictions-et-les-autorisations-liees-au-confinement-en-france_6033337_32...	17.03.2020
46	https://www.rtbf.be/info/societe/detail_coronavirus-faut-il-imposer-en-belgique-un-lockdown-total-comme-en-france?id=10459854	17.03.2020
47	http://www.leparisien.fr/societe/coronavirus-j-ai-probablement-trop-rassure-les-francais-le-mea-culpa-de-michel-cymes-17-03-2020-8281755.php	17.03.2020
48	https://www.laprovence.com/actu/en-direct/5936036/coronavirus-a-marseille-les-resultats-prometteurs-des-essais-cliniques-a-lhydroxychloroquine.html	17.03.2020
49	https://www.francetvinfo.fr/sante/maladie/coronavirus/coronavirus-voici-l-attestation-de-deplacement-obligatoire_3871031.html	17.03.2020
50	https://www.cnews.fr/monde/2020-03-17/coronavirus-une-carte-en-temps-reel-pour-suivre-levolution-de-lepidemie-920567	17.03.2020
51	https://www.bfmtv.com/societe/coronavirus-voici-l-attestation-que-vous-devez-telecharger-pour-vous-deplacer-durant-le-confinement-1876207.html	17.03.2020
52	https://www.lefigaro.fr/sciences/2020/03/17/01008-20200317LIVWWW00001-en-direct-coronavirus-confinement-mesures-ecoles-hopitaux-Macron-annonces-...	17.03.2020
53	https://www.santemagazine.fr/sante/maladies/maladies-infectieuses/maladies-virales/tout-savoir-sur-les-infections-respiratoires-a-coronavirus-431783	17.03.2020
55	https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019/advice-for-public	17.03.2020
56	https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019/advice-for-public	17.03.2020
57	https://www.who.int/fr/health-topics/coronavirus/coronavirus	17.03.2020
58	https://www.who.int/fr/health-topics/coronavirus/coronavirus	17.03.2020
59	https://www.who.int/fr/news-room/q-a-detail/q-a-coronaviruses	17.03.2020
60	https://www.who.int/fr/news-room/q-a-detail/q-a-coronaviruses	17.03.2020
62	https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd	17.03.2020
63	https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd	17.03.2020
64	https://www.vd.ch/toutes-les-actualites/hotline-et-informations-sur-le-coronavirus/informations-destinees-aux-parents-et-aux-professionnels-de-la-formation-...	17.03.2020
65	https://www.bag.admin.ch/bag/fr/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov.html	17.03.2020
67	https://webcache.googleusercontent.com/search?q=cache:2H_34A9oi48J:https://www.bag.admin.ch/bag/fr/home/krankheiten/ausbrueche-epidemien-pande...	17.03.2020
68	https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019/advice-for-public	17.03.2020
70	https://webcache.googleusercontent.com/search?q=cache:PrAEq6gH80MJ:https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019/advice-for-p...	17.03.2020
71	https://www.cdc.gov/coronavirus/2019-ncov/about/index.html	17.03.2020
73	https://webcache.googleusercontent.com/search?q=cache:96NCPUWRCrkJ:https://www.cdc.gov/coronavirus/2019-ncov/about/index.html+&cd=17&hl=fr&ct...	17.03.2020
75	https://www.cdc.gov/coronavirus/2019-ncov/index.html	17.03.2020
77	https://webcache.googleusercontent.com/search?q=cache:B_FPQc5umtEJ:https://www.cdc.gov/coronavirus/2019-ncov/index.html+&cd=18&hl=fr&ct=clnk&g...	17.03.2020

7.2 Récupération des liens redirigeant sur d'autres pages Google

```
-- Requête retournant les URLs redirigeant vers une autre page de résultat Google
select * from Site
    where lower(url) like "%start%";
```

id	url	date_consultation
91	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=10&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBAx	17.03.2020
92	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=20&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBAz	17.03.2020
93	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=30&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA1	17.03.2020
94	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=40&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA3	17.03.2020
95	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=50&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA5	17.03.2020
96	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=60&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA7	17.03.2020
97	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=70&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA9	17.03.2020
98	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=80&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBA_	17.03.2020
99	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=90&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBBB	17.03.2020
100	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk02Y0ZhRvEA9XSv6CtJE12teQhdiDA:1584438896474&ei=cJ5wXvTPHLaS1fAP9MuFkA8&start=100&sa=N&ved=2ahUKEwj0x97snqHoAhU2SRUIHRIAfQ8tMDegQIFBBD	17.03.2020
196	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk00EFxVhqKWREl5blmomB28FVb4tA:1584565939563&ei=s45yXuBIo6imwW1zbjoCw&start=10&sa=N&ved=2ahUKEwiop82P-KToAhUO0aYKHbUmDr0Q8tMDegQIFBAw	18.03.2020
197	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk00EFxVhqKWREl5blmomB28FVb4tA:1584565939563&ei=s45yXuBIo6imwW1zbjoCw&start=20&sa=N&ved=2ahUKEwiop82P-KToAhUO0aYKHbUmDr0Q8tMDegQIFBAy	18.03.2020
198	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk00EFxVhqKWREl5blmomB28FVb4tA:1584565939563&ei=s45yXuBIo6imwW1zbjoCw&start=30&sa=N&ved=2ahUKEwiop82P-KToAhUO0aYKHbUmDr0Q8tMDegQIFBA0	18.03.2020
199	https://www.google.com/search?q=coronavirus&lr=1C1ASVA_enCH829CH829&sxsrf=ALeKk00EFxVhqKWREl5blmomB28FVb4tA:1584565939563&ei=s45yXuBIo6imwW1zbjoCw&start=40&sa=N&ved=2ahUKEwiop82P-KToAhUO0aYKHbUmDr0Q8tMDegQIFBA2	18.03.2020

7.3 Processus de nettoyage des mots

Figure 45 : Fonction de nettoyage du dataset de mot utilisant la librairie NLTK

```
def nltk_process(text):
    sentence_tokens = [x.replace('\n','') for x in sent_tokenize(text, language="french")]
    stop_words = stopwords.words('french')
    filtered_text = [] # tableau final contenant les données nettoyées
    sent_tab = [] # tableau indiquant l'indice de la phrase correspondant à chaque mot
    sent_id = 0
    for s in sentence_tokens:
        for w in word_tokenize(s.lower()):
            if w not in stop_words:
                if w.isalpha():
                    filtered_text.append(w)
                    sent_tab.append(sent_id)
        sent_id += 1
    return filtered_text, sent_tab # len(filtered_text) = len(sent_tab)
```

7.4 Liste de mots à analyser

Tableau 4 : Liste de mots

Pandémie	Épidémie	Virus	Mondial	Transmission
Protection	Masque	Mortel	Létalité	Soin
Décès	Contamination	Quarantaine	Prévention	Pays
Détection	Armée	Hôpitaux	Médecin	Infirmière
Docteur	Respirateur	Dépistage	Clinique	Vaccin
Remède	Nourrisson	Risque	Âgées	Analyse
Symptôme	Rhume	Odorat	Fièvre	Hydroalcoolique
Désinfectant	Traitement	Gouvernement	Confinement	Déconfinement
Sanitaire	Isolement	Auto-isolement	Cas	Covid
Sras	Coronavirus	Main	Surface	Frontière
Incubation	Infection	Contact	Reconfinement	Reconfiner
Taux				