Projet tableau de bord

"L'impact des réseaux sociaux sur notre vie et notre comportement"

Janvier-Mars 2021

M1 SID

Client:

Wahiba BASHOUN

Riad MOKADEM

Fournisseurs :

Thierno Mamadou DIALLO

Modou Bamba DIOUF

Antoine GODIN

Yacine SEBA

Historique du document

Date	Numéro de version	Chapitre concerné	Cause de la modification
10.02.2021	V1.0	Tous	Création du document
14.02.2021	V1.1	Objectif du document Organisation de développement	Ecriture du rapport
21.02.2021	V1.2	Document de références et terminologie	Ecriture du rapport
28.02.2021	V2.0	Tous	Modification du plan du rapport
01.03.2021	V2.1	Documents de références et terminologie Démarche de développement	Ecriture du rapport
07.03.2021	V2.2	Assurance Qualité Démarche de développement	Ecriture du rapport
12.03.2021	V3.0	Tous	Modification du plan du rapport
14.03.2021	V3.1	Gestion de configuration Démarche de développement	Mise à jour et écriture du rapport
17.03.2021	V3.2	Assurance qualité Démarche de développement Bilan de projet	Ecriture du rapport Mise en forme et mise en page
20.03.2021	V3.3	Tous	Clôture du rapport

Sommaire

I. Objectifs du document

I.1 Objet

Ce document constitue un rapport visant à détailler toutes les étapes que nous avons réalisé pour mener à bien notre projet. Ce projet, intitulé « Projet Tableau de bord », s'inscrit dans le cadre de notre formation de Master 1 en Statistiques et Informatique Décisionnelle (M1 SID). L'objectif de ce projet est de développer un système de visualisation des données afin de répondre à une problématique traitant d'un sujet de notre choix.

Le document présent regroupe l'ensemble des méthodes et étapes que nous avons mis en place au cours de ces deux mois de travail. De la présentation du projet à la démarche de développement, en passant par la composition de notre équipe et l'organisation du développement ou encore par la gestion de configuration et l'assurance qualité mise en place, nous détaillerons ici toutes les difficultés rencontrées ainsi que les solutions qui ont été apporté.

I.2 But

Conformément aux consignes évoquées lors de la présentation du projet, nous avons toujours gardé à l'esprit que ce travail s'inscrivait dans un contexte de relation Client-Fournisseur dans lequel nos enseignants Mme. Bashoun et M. Mokadem représentent les clients. De l'autre côté, nous, quatre étudiants de la formation SID, représentons le fournisseur.

I.3 Domaine d'application

Nous avons évoqué plus haut la liberté qui fût la nôtre de choisir librement le sujet et la problématique qui nous convenait. Après quelques jours de réflexion nous sommes tombés d'accord sur une problématique qui nous a paru intéressante et suffisamment d'actualité pour nous permettre de rassembler un nombre d'articles conséquent : « Quels sont les impacts des réseaux sociaux sur notre vie et sur notre comportement ? ».

Etant tous issus de la génération précédant immédiatement le passage à l'an 2000, nous avons pu constater « en direct » l'apparition et le développement des réseaux sociaux au cours des années 2000 et 2010. A titre personnel, nous avons remarqué que nous avions découvert les réseaux sociaux au début de notre adolescence, et que cela a fortement impacté le comportement, et la manière de communiquer de chacun.

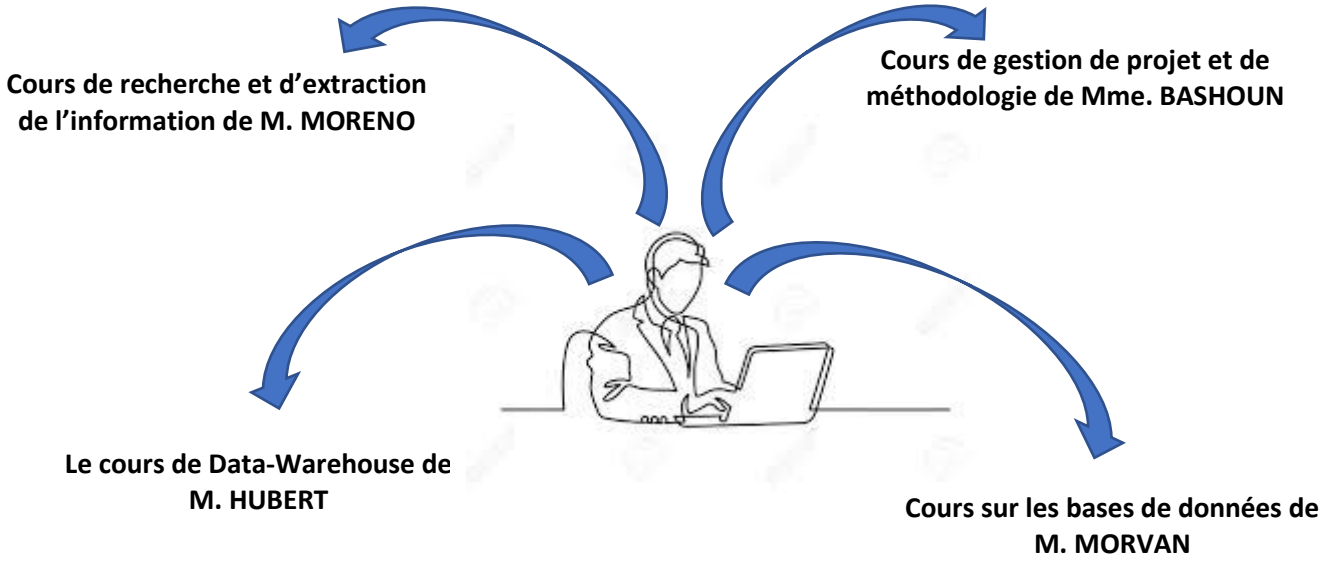
De plus, nous avons chacun pu observer dans notre entourage la différence entre notre expérience et celles menée par des membres plus jeunes de notre famille qui découvrent les réseaux sociaux de plus en plus tôt, dès l'enfance. Cela impacte le comportement différemment selon les générations, mais nous pensons qu'il y a autant d'impacts négatifs que positifs, ces derniers ayant notamment été mis en avant depuis la crise sanitaire.

Notre objectif à travers ce projet est donc de vérifier nos suppositions et d'observer les tendances générales issues de l'étude d'un grand nombre de données.

II. Documents de références et terminologie

II.1 Documents de références

Afin de mener à bien ce projet nous avons pu nous appuyer sur différents documents de références. Ces derniers nous ont apporté la connaissance et la maîtrise de techniques indispensables à la réalisation de ce projet :



Mais aussi :

- L'ensemble des documents résultant du groupe Scraping du projet Interpromo

- Documentation sur l'API (Elsevier - https://dev.elsevier.com/api_docs.html)

- Documentation sur le Framework Streamlit (https://docs.streamlit.io/en/stable/api.html#display-charts)

II.2 Documents d'application

Un document d'application est un document qui regroupe les consignes que nous avons appliqué telles quelles. Dans le cadre de ce projet, notre document d'application de référence est le cahier des charges fourni par nos clients (*une copie de ce cahier des charges est disponible en annexe*).

II.3 Terminologie

Tout au long de ce projet et de ce rapport, nous avons travaillé en mentionnant des désignations et des notions propres à un domaine relativement technique. Nous avons donc jugé nécessaire d'établir une terminologie afin d'expliciter certains de ces termes.

. MCD (= Modèle Conceptuel des Données)

Un MCD a pour but d'écrire de façon formelle les données qui seront utilisées par le système d'information. Il s'agit donc d'une représentation des données, facilement compréhensible, permettant de décrire le système d'information à l'aide d'entités.

. MLD (= Modélisation logique des données)

Le MLD est une étape de la conception qui consiste à décrire la structure des données utilisées sans faire référence à un langage de programmation. Il s'agit de préciser le type de données utilisées lors des traitements.

. Scraping

Le scraping est une méthode d'extraction de l'information depuis des sites internet. Cette méthode permet notamment de collecter des informations diverses et variées en grande quantité. Une fois scrappées, ces informations sont enregistrées dans des tableaux ou des bases de données.

. API (= Application Programming Interface)

Une API est une interface de programmation qui permet d'établir des connexions entre plusieurs logiciels dans le but d'échanger des données. Elle se compose d'un ensemble de fonctions qui vont permettre à un développeur d'utiliser une application dans son programme.

. SQL (= Structured Query Language)

Le SQL est un langage informatique de requêtes et d'interrogation de données. On l'utilise notamment lorsque que l'on veut exploiter une ou des bases de données.

. SADT (= Structured Analysis and Design Technique)

La méthode SADT, ou analyse fonctionnelle descendante en français est une technique d'analyse et de modélisation très utile dans une gestion de projet.

. Diagramme de GANTT

Le diagramme de GRANTT est un outil de gestion de projet qui permet de visualiser les différentes échéances prévisionnelles d'un projet.

III. Organisation de développement

III.1 Ressources humaines

Notre équipe de projet est constituée de quatre étudiants suivant un Master 1 SID-Big Data à l'université Toulouse III Paul Sabatier. Ayant l'habitude de travailler ensemble à distance depuis le début de la crise sanitaire nous connaissions bien les qualités et domaines de prédilections de chacun d'entre nous. Cet avantage a facilité la répartition des rôles, qui s'est faîte comme suit :

- Thierno Mamadou DIALLO : Chef de projet
  - Organisation et coordination de l'équipe
  - Pilotage du projet
  - Planification et suivi de l'avancement
  - Responsable de la base de données SQL

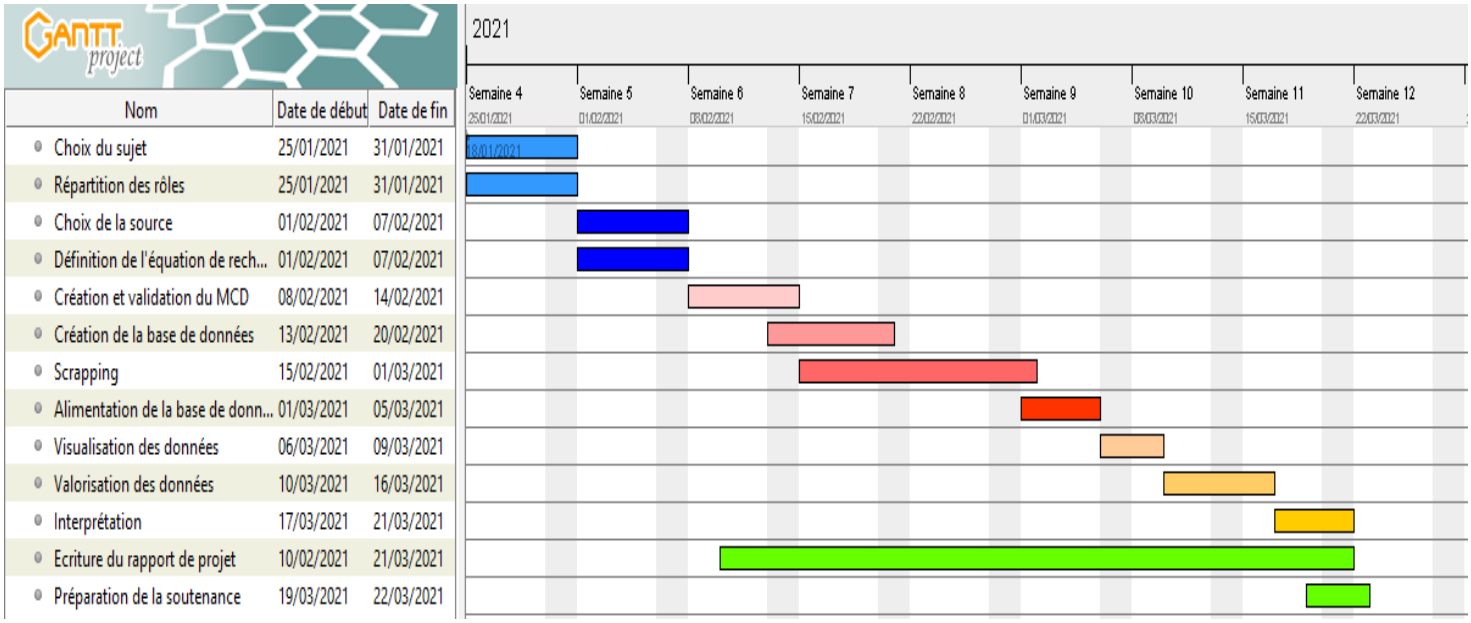
- Modou Bamba DIOUF : Responsable gestion de configuration
  - Chargé de la traçabilité des versions
  - Codage du scraping
  - Elaboration du MCD
  - Nettoyage des données

- Antoine GODIN : Responsable du contrôle qualité et rédaction du rapport
  - Organisation des revues
  - Rédaction du rapport
  - Elaboration du MCD
  - Interprétation des résultats

- Yacine SEBA : Responsable du code
  - Codage du scraping
  - Elaboration de l'équation de recherche
  - Alimentation de la base de données
  - Implémentation de l'interface Streamlit

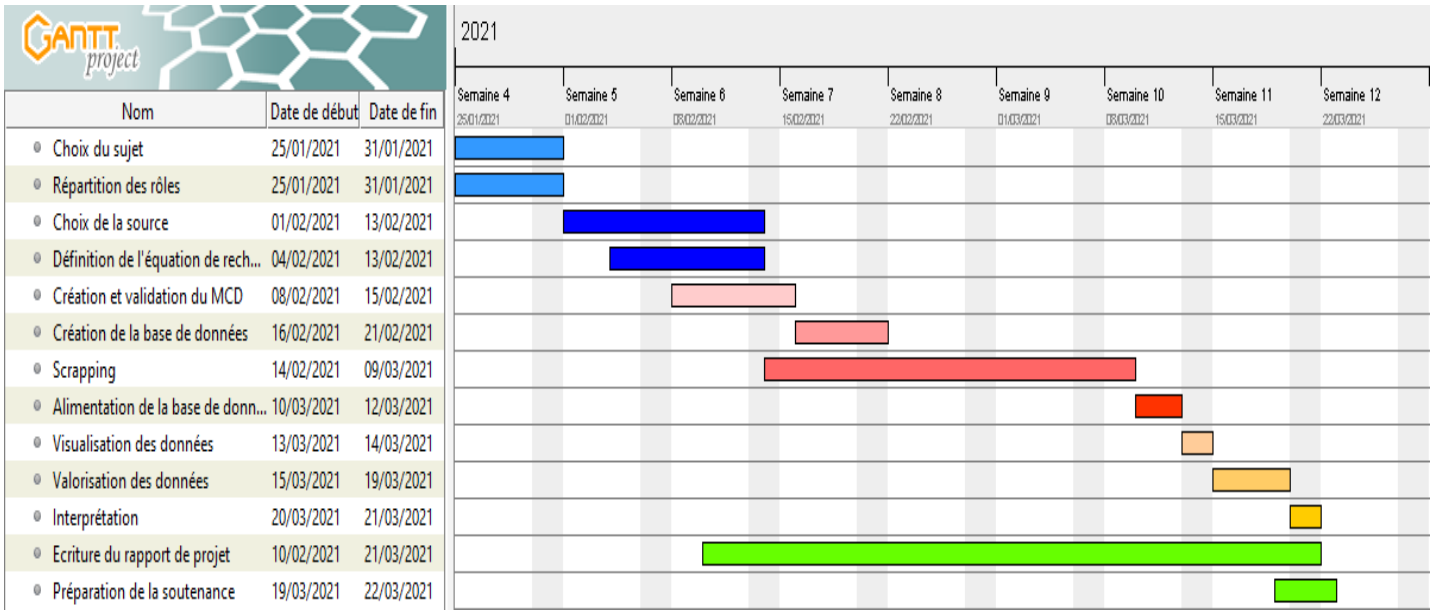
III.2 Gestion de projet et diagramme de GANTT

. Diagramme de GANTT : Calendrier prévisionnel



Nom	Date de début	Date de fin
Choix du sujet	25/01/2021	31/01/2021
Répartition des rôles	25/01/2021	31/01/2021
Choix de la source	01/02/2021	07/02/2021
Définition de l'équation de rech...	01/02/2021	07/02/2021
Création et validation du MCD	08/02/2021	14/02/2021
Création de la base de données	13/02/2021	20/02/2021
Scrapping	15/02/2021	01/03/2021
Alimentation de la base de donn...	01/03/2021	05/03/2021
Visualisation des données	06/03/2021	09/03/2021
Valorisation des données	10/03/2021	16/03/2021
Interprétation	17/03/2021	21/03/2021
Ecriture du rapport de projet	10/02/2021	21/03/2021
Préparation de la soutenance	19/03/2021	22/03/2021

. Diagramme de GANTT : Calendrier réel



Nom	Date de début	Date de fin
Choix du sujet	25/01/2021	31/01/2021
Répartition des rôles	25/01/2021	31/01/2021
Choix de la source	01/02/2021	13/02/2021
Définition de l'équation de rech...	04/02/2021	13/02/2021
Création et validation du MCD	08/02/2021	15/02/2021
Création de la base de données	16/02/2021	21/02/2021
Scrapping	14/02/2021	09/03/2021
Alimentation de la base de donn...	10/03/2021	12/03/2021
Visualisation des données	13/03/2021	14/03/2021
Valorisation des données	15/03/2021	19/03/2021
Interprétation	20/03/2021	21/03/2021
Ecriture du rapport de projet	10/02/2021	21/03/2021
Préparation de la soutenance	19/03/2021	22/03/2021

. En observant les deux diagrammes de GANTT ci-dessus on peut remarquer de légères différences entre notre calendrier prévisionnel et la réalité des faits. Il faut garder à l'esprit

qu'il y a toujours des imprévus, notamment sur un projet qui s'étale sur plusieurs semaines, voire plusieurs mois. Ainsi on remarque qu'une des premières étapes, le choix de la source, nous a pris un peu plus de temps que prévu. En effet nous avons hésité longuement entre deux sources à scrapper : PubMed et ScienceDirect. Chacune présentant des avantages et des inconvénients (facilité de scraping, nombre d'articles, cohérence des résultats vis-à-vis de l'équation de recherche etc…).

In fine, nous avons choisi de scrapper les articles via ScienceDirect dans un souci de cohérence, bien que ce dernier nous était personnellement plus compliqué à scrapper.

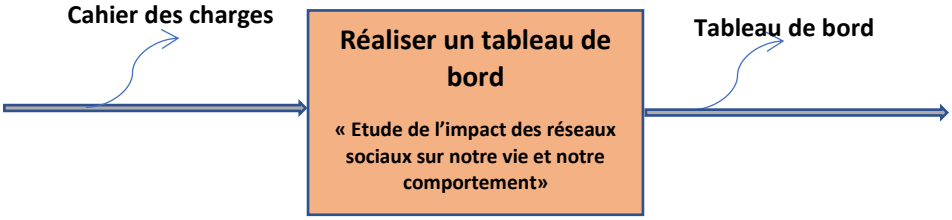
Ainsi, en plus de la semaine de retard prise dès le début du projet, le scraping fut plus complexe et retord que prévu, ajoutant quelques jours de retard supplémentaires par rapport à notre planning prévisionnel, et ce, malgré avoir accéléré la cadence.

Heureusement, une meilleure répartition des tâches dans la deuxième partie du projet nous a permis de rattraper ce retard sans toutefois négliger la partie « valorisation des données », étape clé de la finalité de notre travail.

IV. Démarche de développement

IV.1 Définition du processus (SADT)

Dans le cadre de ce projet nous avons utilisé une méthode de gestion de projet appelé SADT (définie dans la terminologie). Cette méthode nous permet d'analyser toutes les étapes du projet à différents niveaux de conception. Cela donne la possibilité d'avoir aussi bien une vision globale qu'une vision détaillée de chacune des étapes constituant le projet.



Modèle A-0

Le modèle A-0 ci-dessus est la vision la plus globale du projet. On dispose d'un cahier des charges fourni par notre client, et on doit réaliser un tableau de bord sur un thème choisi librement. In fine, nous devons livrer un tableau de bord au client.

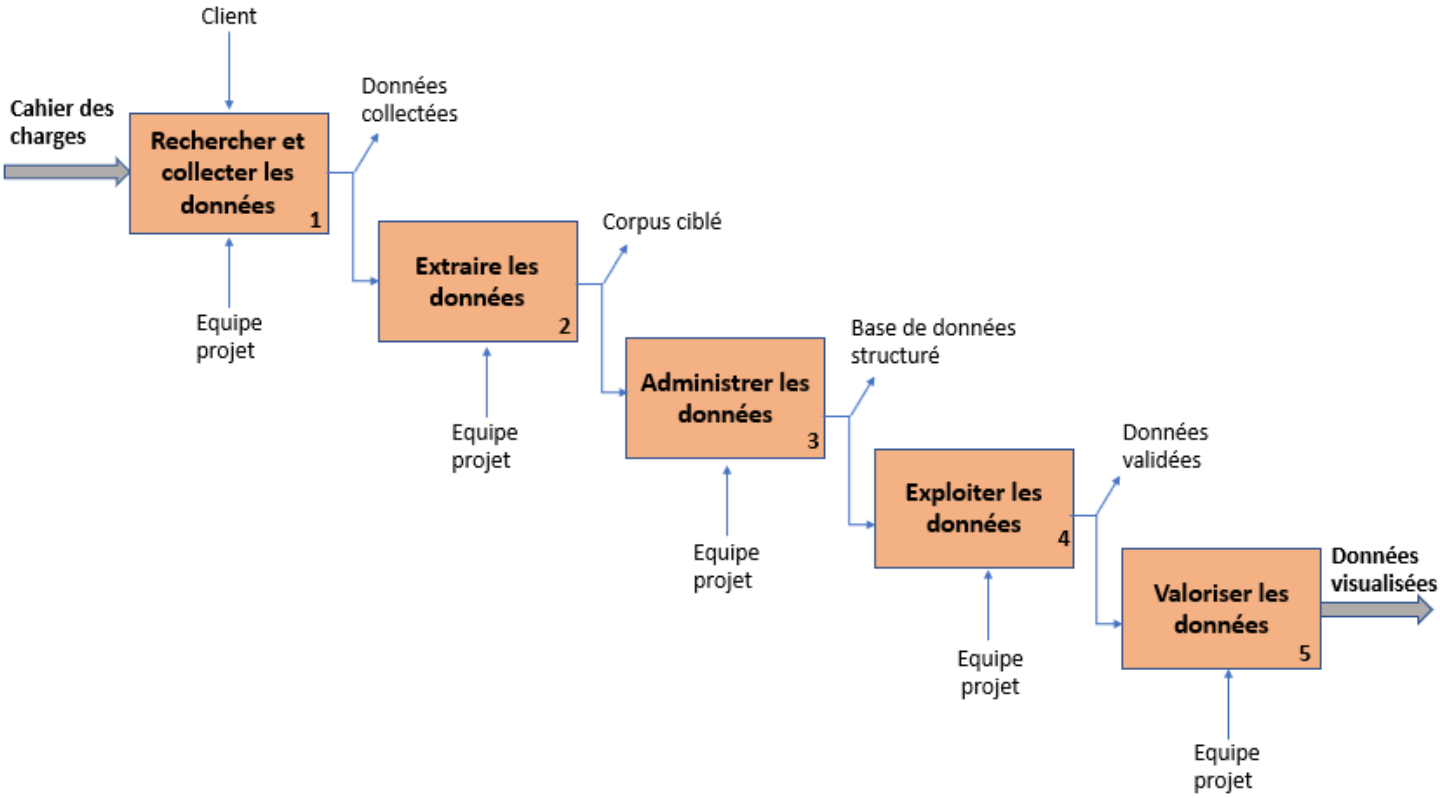


Diagramme A0 : Réaliser un tableau de bord

Le diagramme A0 visible ci-dessus détaille de manière globale les étapes nécessaires à la réalisation de ce tableau de bord. On a retenu cinq grandes étapes et nous allons étudier plus en détail le contenu de certaines d'entre elles que nous avons jugées intéressantes.

IV.2 Collecte des données

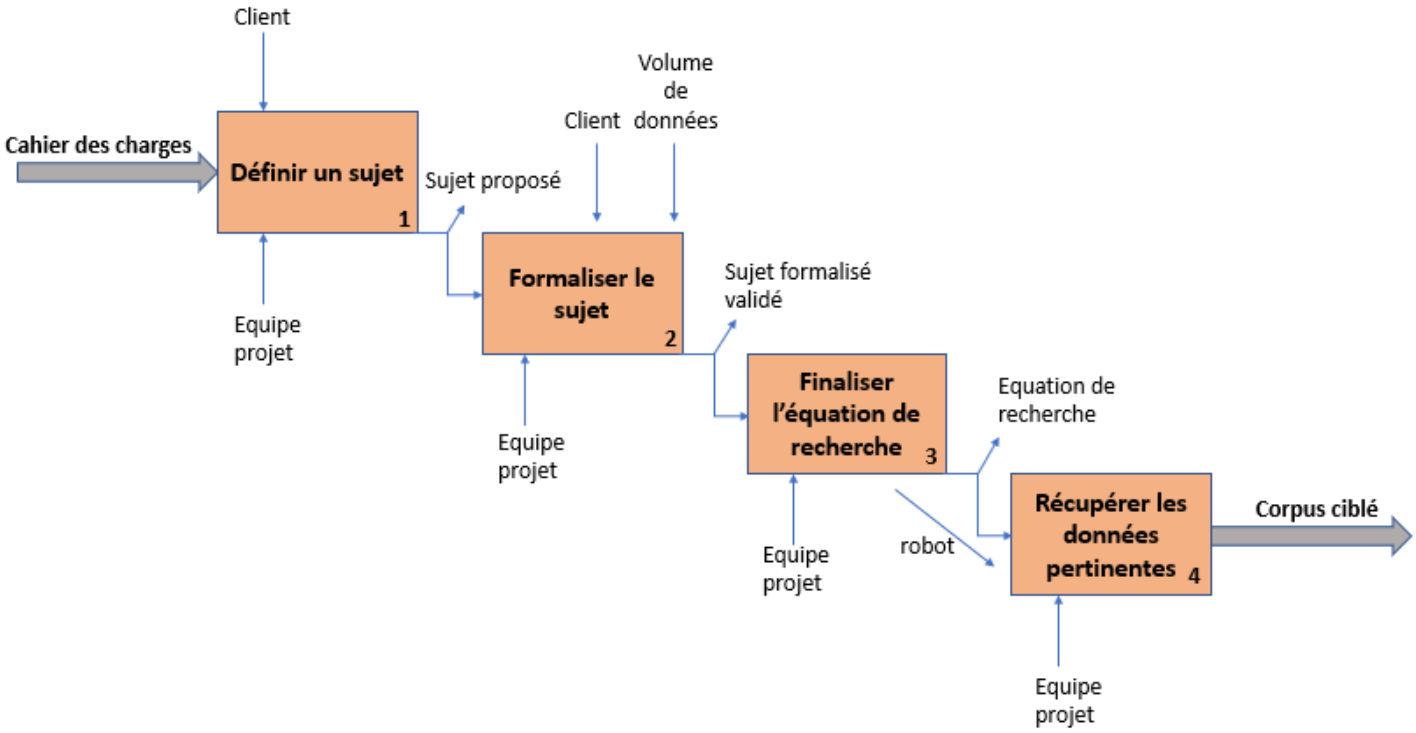


Diagramme A1 : Rechercher et collecter les données

Le diagramme A1 recense les quatre étapes nécessaires à la recherche et à la collecte des données. Tout d'abord, la définition et la formalisation du sujet. Ce dernier doit être en adéquation avec les demandes spécifiées dans le cahier des charges. Ces deux étapes sont faîtes sous le contrôle des clients qui doivent valider le sujet une fois ce dernier formalisé. Une contrainte dans ce projet était l'obligation de réunir un grand nombre de données (d'articles) afin d'avoir des résultats fiables à la fin.

Afin de réunir un nombre d'articles suffisants (au moins 5 000), la définition de l'équation de recherche fût une étape clé de notre projet. Après plusieurs propositions nous avons retenu l'équation suivante : « (social network AND Social life) OR (social media AND social fabric risks) OR (social web AND social life impact) ».

Cette équation nous a permis de retenir un nombre d'articles que nous avons jugé suffisant (quasiment 7 000), tout en permettant de conserver une certaine qualité dans le contenu de nos articles. Trouver le juste milieu entre une équation de recherche trop stricte, retournant trop peu d'articles, et une équation trop générale retournant un grand nombre d'articles mais assez peu cohérents avec notre sujet, fut le défi de cette étape.

IV.3 Préparation et extraction des données

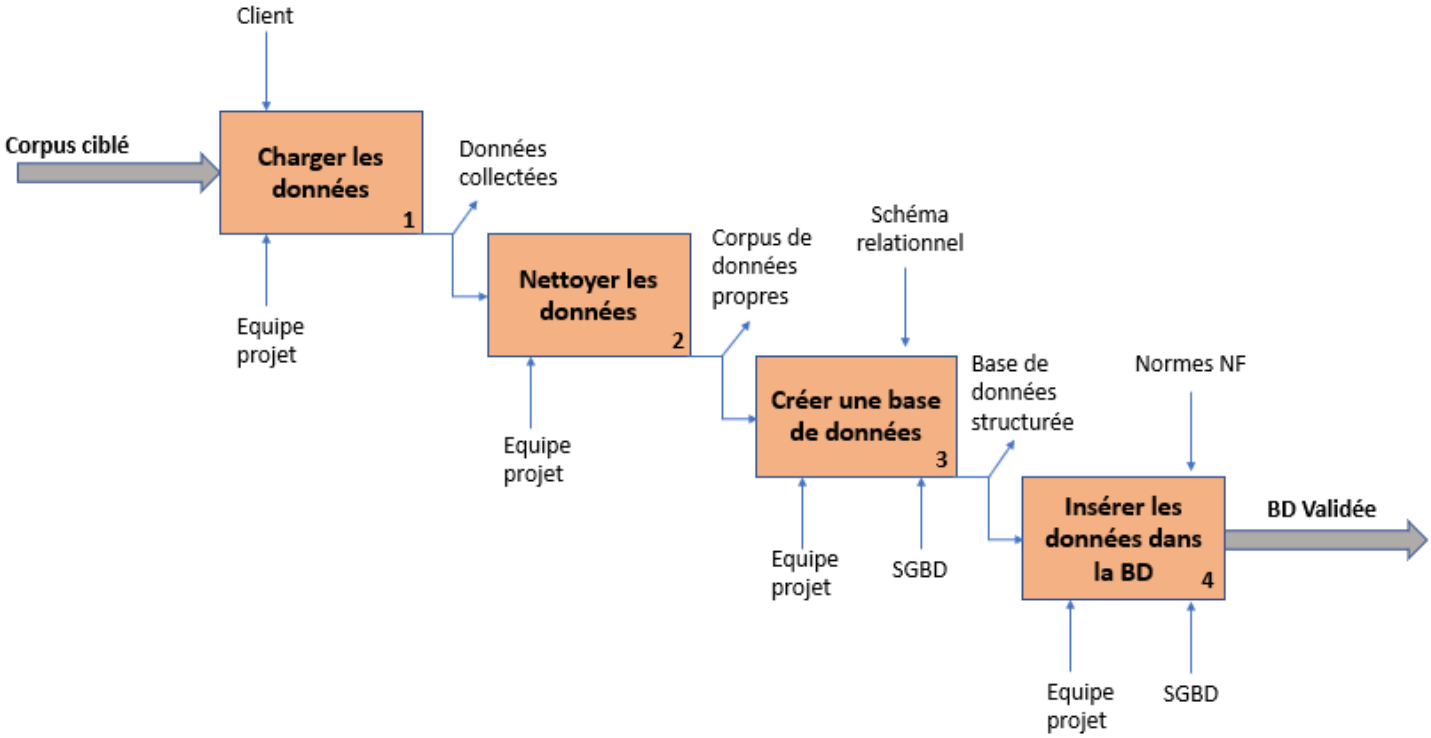


Diagramme A2 : Extraire les données

Le diagramme A2 ci-dessus nous permet de retrouver les étapes réalisées afin d'extraire les données. Après le chargement de ces dernières sous le contrôle de nos clients, une grande étape de ce processus fut le nettoyage des données. Nous avons procédé comme suit, avec entre autres :

- Nettoyage de la nationalité associée à chaque auteur. Passage de la chaîne de caractère en minuscule, vérification de la correspondance des dénominations, notamment pour les abréviations comme « UK » ou « USA » que nous avons convertis à la main. De plus, pour certains auteurs nous ne disposions que de la ville ou de l'état (notamment dans le cas des U.S) et nous avons également assigné à la main le pays correspondant.
- Nettoyage des mots clés des articles. Passage des chaînes de caractères en minuscule. Suppression des mots vides (stop-words). Découpage des mots-clés composés (si contenant 3 mots ou plus). Nous avons également supprimé les caractères spéciaux.
- Nettoyage des titres des articles. Passages des chaînes de caractère en minuscule. Suppression des mots vides. Nous avons ici décidé de supprimer la ponctuation.

Une fois ces points réalisés, nous disposions enfin de données « propres ». Intervient ici une deuxième partie de l'extraction des données qui englobe également l'administration de celle-ci. Nous avons conçu un MCD, validé par nos clients et disponible en annexe, sur lequel nous nous sommes appuyés pour créer une base de données structurée. Afin de ne pas perdre de

temps, deux d'entre nous se sont concentrés sur le scraping tandis que deux autres étaient chargé de la création de la base de données en parallèle. Ces deux étapes terminées, nous avons rapidement pu insérer nos données nettoyées dans la base de données, en respectant les normes NF. Ainsi nous avons obtenu une base de données complète et validée.

IV.4 Exploitation des données

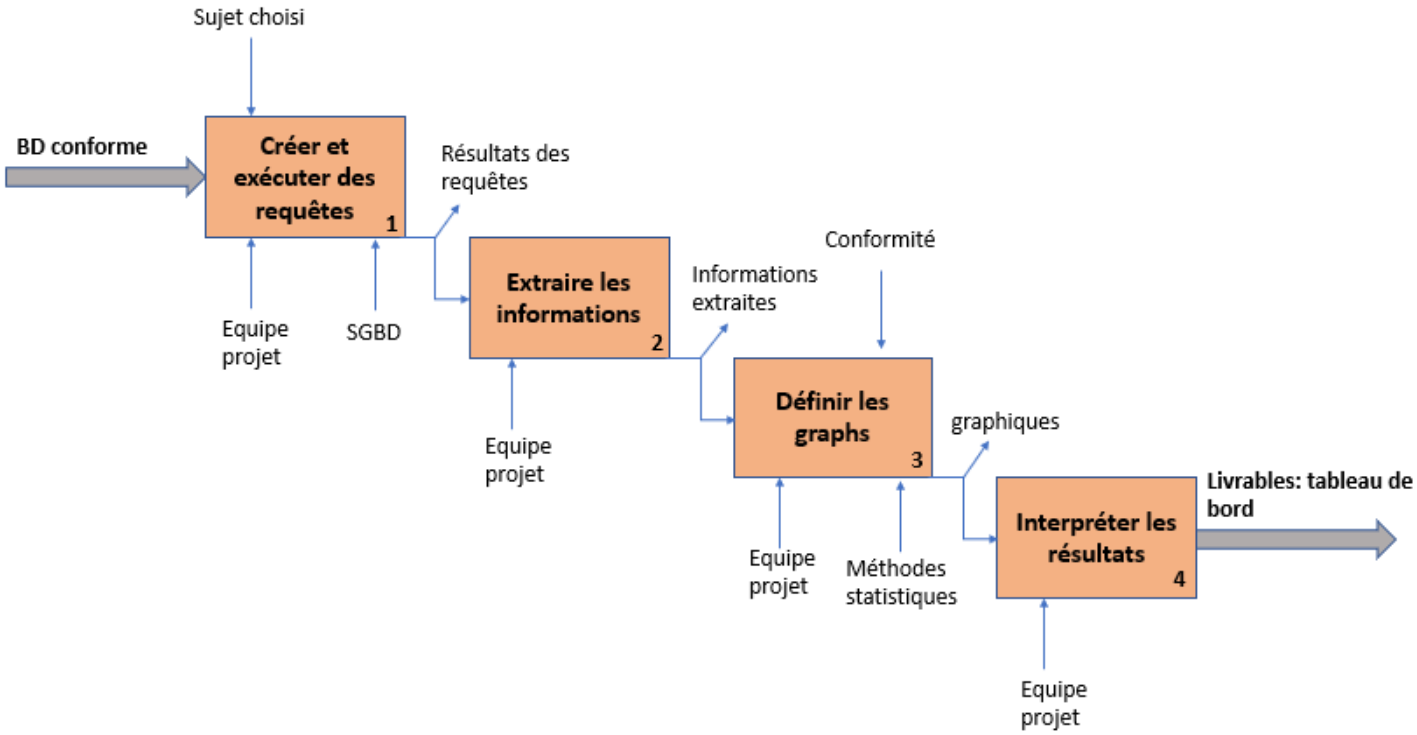


Diagramme A4 : Exploiter les données

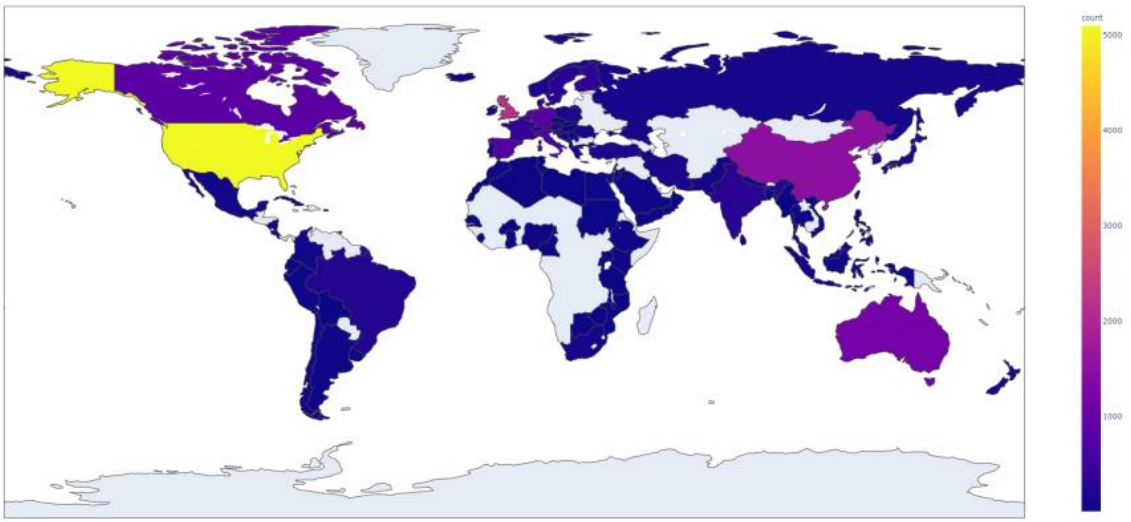
La dernière grande partie de notre projet se résume à travers le diagramme A4 ci-dessus. La phase d'exploitation des données est la finalité de notre projet. On élabore des requêtes en corrélation avec notre sujet afin d'apporter des réponses à nos interrogations et on extrait les informations qui nous semblent pertinentes. Une fois les informations extraites, nous avons mis en place des méthodes statistiques dans le but de construire des graphiques permettant d'interpréter les résultats.

IV.5 Valorisation et diffusion des données

Dans cette partie nous allons visualiser, détailler et interpréter l'ensemble des graphiques que nous avons construit. L'objectif de cette visualisation étant de dégager des tendances et de faire des observations dans l'optique de répondre à notre problématique.

Notons que nous avons construit d'autres graphique sur Streamlit mais nous ne détaillons ici qu'une partie d'entre eux.

. Analyse de l'origine des auteurs :



Répartition des auteurs par nationalité

L'analyse de la carte ci-dessus se révèle intéressante pour plusieurs raisons. Tout d'abord on peut observer l'ultra dominance en termes de représentation des auteurs issus des Etats-Unis. Cela s'explique surement par le fait que les réseaux sociaux ont pris forme de ce côté de l'Atlantique avec notamment :

- Facebook : société américaine
- Instagram : société américaine (racheté par Facebook)
- Twitter : société américaine
- Snapchat : société américaine

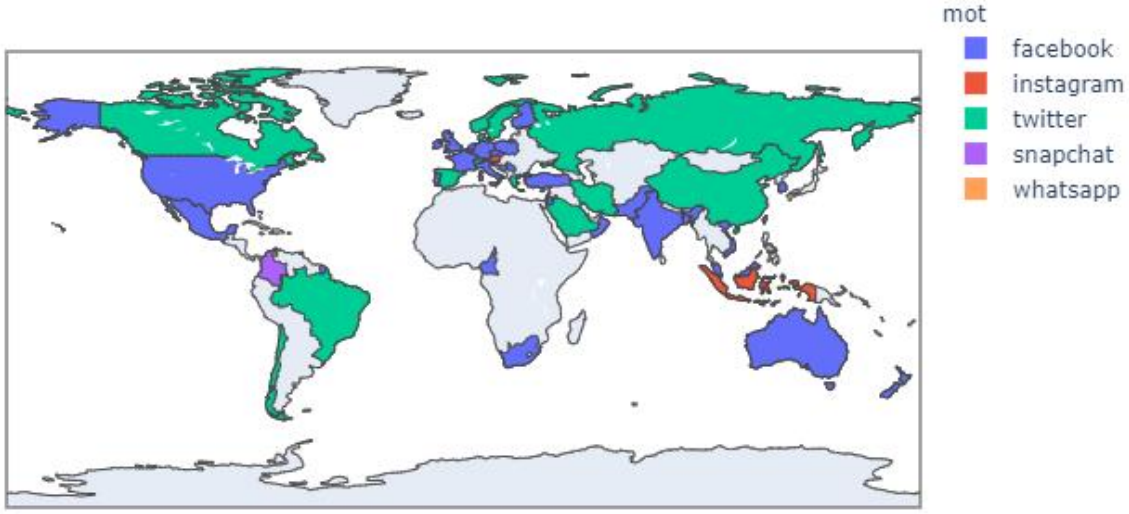
Les quatre réseaux sociaux que nous avons cités ci-dessus sont les plus utilisé dans le monde à ce jour, et les quatre ont été fondé, et sont toujours développés, aux Etats-Unis. Il apparaît donc logique que la grande majorité des auteurs d'articles traitant des réseaux sociaux soient originaire de ce pays.

En outre, quelques autres pays sortent du lot derrière les USA. C'est notamment le cas de la Chine, à laquelle on a envie d'associer la création de « TikTok », réseau social chinois extrêmement en vogue ces derniers temps dans le monde entier. Ce dernier est actuellement le seul pouvant bousculer les quatre premiers réseaux sociaux que nous avons cité plus haut. L'attrait des auteurs chinois pour ce sujet semble donc tout à fait logique.

Enfin on observe d'autre pays bien représentés comme le Royaume-Uni ou l'Australie, des pays anglophones (le corpus étudié est exclusivement en anglais) et à la pointe de la technologie

. Analyse des mots-clés

➔ Quels réseaux sociaux ?



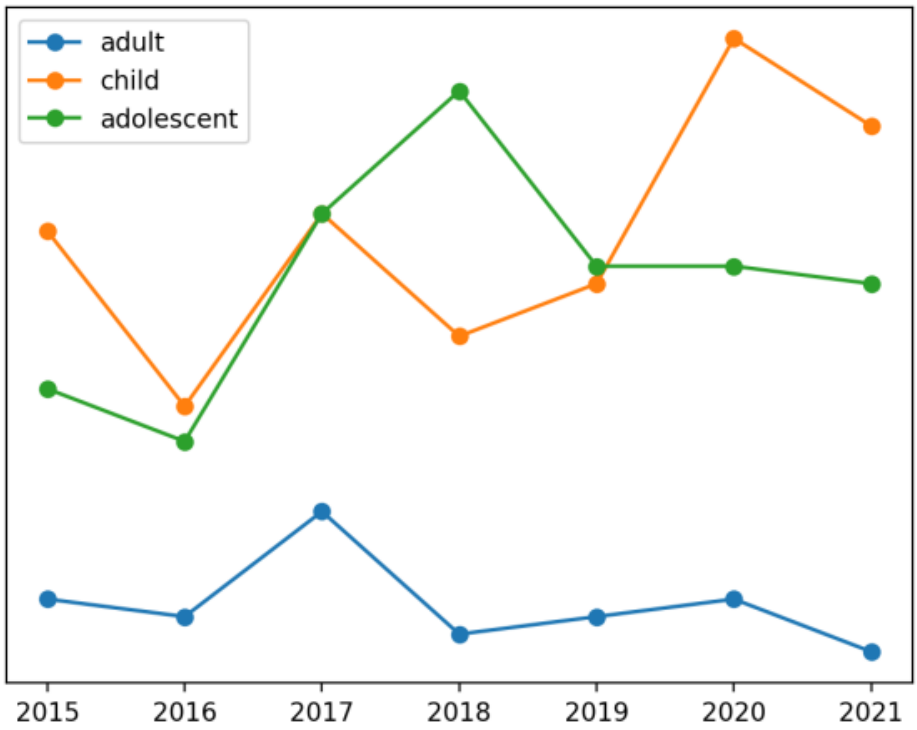
Réseaux sociaux les plus mentionnés par pays

Avant de s'intéresser aux éventuels impacts des réseaux sociaux sur nos comportements, nous devons savoir de quels réseaux nous parlons. Pour cela nous avons élaboré la carte ci-dessus mettant en évidence, pour chaque pays, le réseau social le plus mentionné dans les mots-clés.

Sans surprise, le monde semble se découper en deux catégories : Les pays préférant l'utilisation de Facebook, et ceux celle de de Twitter. Ces deux derniers s'étant affirmé comme les leaders mondiaux, en termes de nombre d'utilisateurs, depuis plus de quinze ans.

Ainsi, hormis quelques exceptions comme l'Indonésie (avec Instagram) ou la Colombie (Snapchat), nous savons désormais ce qui se cache le plus souvent derrière le terme « social media ».

➔ Quel public ?



Evolution du nombre d'occurrences des mots-clés « adult », « child » et « adolescent »

Notre objectif est d'étudier l'impact des réseaux sociaux sur le comportement humain. Cependant, nous nous doutions que ces impacts ne touchent pas toutes les catégories de personne de la même manière. Les réseaux sociaux étant apparu au milieu des années 2000, et s'étant fortement développés et imposés ces dernières années (d'où le choix de notre période d'étude 2015-2021), il était fort probable qu'on les associe davantage à la jeune génération.

C'est ce que nous avons voulu vérifier à travers le graphique ci-dessus. Nous avons étudié l'évolution du nombre d'occurrence des mots « child », « adult » et « adolescent », qui représente trois étapes bien distinctes de la vie d'un être humain. Premièrement, on observe le faible nombre d'apparition du mot-clé adulte, et ce de manière constante. A l'inverse, et sans surprise, les mots-clés identifiant les enfants et les adolescents suivent une augmentation notable, et sont bien plus nombreux que ceux identifiant les adultes.

Toutefois, nous ne nous attendions pas à ce que les articles de notre corpus réfèrent davantage aux enfants qu'aux adolescents. Nous imaginions que l'utilisation des réseaux sociaux serait plus associée aux adolescents, la plupart des réseaux imposant une limite d'âge minimum correspondant à l'adolescence (13 ans pour Facebook par exemple).

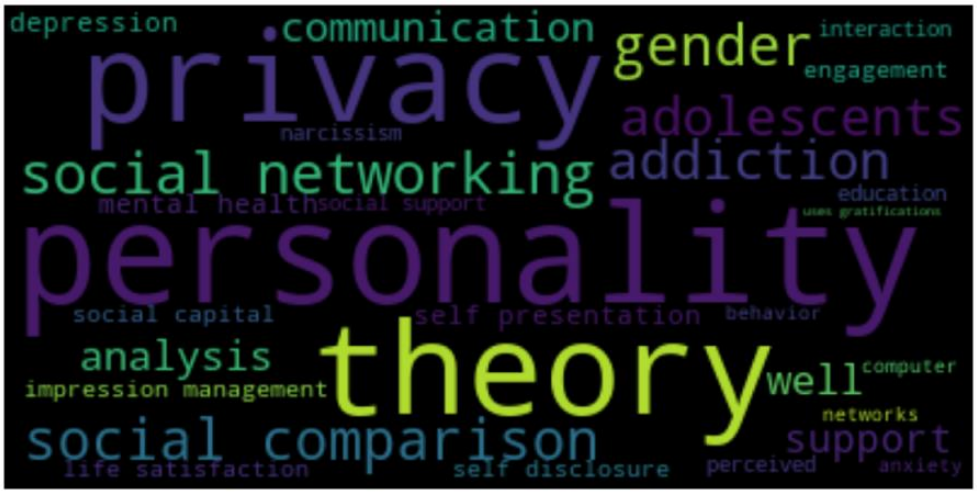
Ce « détail » pourrait s'avérer d'autant plus inquiétant en fonction des résultats de l'analyse des impacts ci-après.

➔ Quels impacts ?

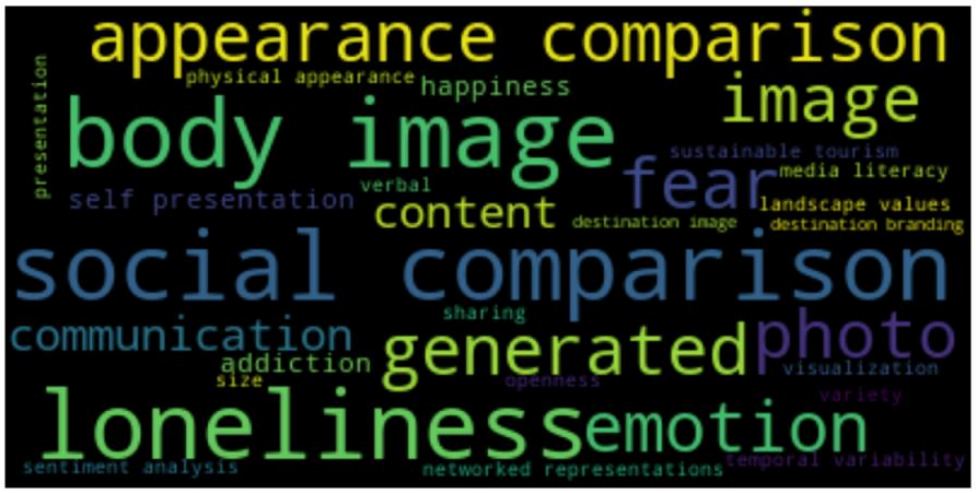
Nous savons désormais de quelles parties du globe proviennent les articles de notre corpus. Nous savons également quels sont les réseaux sociaux les plus mentionnés (Facebook & Twitter), et quelle tranche d'âge est la plus souvent cité (enfants et adolescent).

Cependant l'objectif de notre travail, est avant tout de découvrir quels sont les impacts des réseaux sociaux concernés sur le public identifié.

Dans cette optique, nous avons décidé d'étudier les nuages de mots suivants. Ils ont été construits à partir des mots-clés revenant le plus souvent lorsque Facebook est également mentionné. Même chose pour Instagram.



Nuage des mots-clés lorsque Facebook est mentionné



Nuage des mots-clés lorsque Instagram est mentionné

En observant le premier nuage de mots-clés, lié à Facebook, on remarque plusieurs choses intéressantes. D'abord, les mots qui reviennent le plus souvent sont « personality », « privacy » ou « theory ». Si l'apparition des deux premiers n'est pas vraiment surprenante, la

question du respect de la vie privée étant un sujet récurrent quand on parle de réseaux sociaux, le mot clé « theory » peut s'avérer plus surprenant. Sa présence semble en effet témoigner de l'intérêt grandissant porté aux théories du complots, un sujet malheureusement fortement développé et crédité sur Facebook notamment.

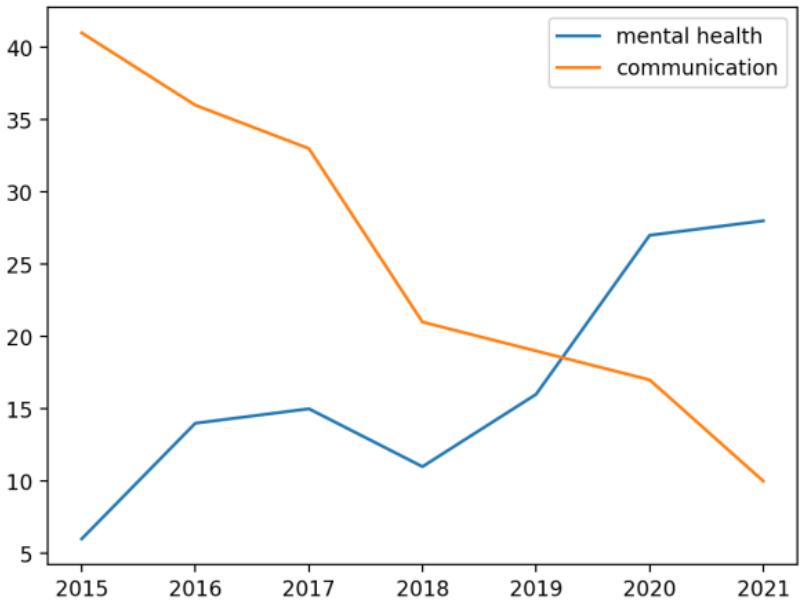
Dans un second temps, on dénote la présence de nombreux mots-clés plus ou moins inquiétant, comme « depression », « addiction » ou « mental health ». Ces derniers tiennent la même place que des mots beaucoup plus à même de ressortir quand on parle de réseaux sociaux, comme « communication » ou « interaction » par exemple.

En parallèle nous avons construit un nuage de mot similaire, avec les mots-clés liés à Instagram cette fois, un réseau social qui contrairement à Facebook, est davantage basé sur le partage de photos que de messages. Cette simple nuance s'identifie immédiatement dans le nuage de mot ci-dessus. En effet, le vocabulaire de la comparaison est extrêmement présent, les mots-clés les plus récurrents étant « social comparison », « body image » ou encore « appearance comparison ».

De plus, les thèmes venant juste derrière ne sont pas plus positifs, avec à l'instar de Facebook, de nombreux mots-clés relatifs à la santé mentale comme la peur (fear), ou l'isolement (loneliness).

L'étude de ces nuages de mot n'est pas rassurante. Les différents impacts des réseaux sociaux sur notre comportement semblent en effet être négatif dans l'ensemble, au point même qu'ils sont souvent associés à des impacts sur notre santé mentale. Notre corpus regroupe un nombre égal d'articles par année sur une période relativement récente (2015-2021). Il nous donne ainsi la possibilité d'étudier certaines tendances.

➔ Quelles tendances ?



Occurrences d'apparition des mots-clés « mental health » et « communication »

Pour cette analyse nous avons sélectionné deux mots-clés prépondérants dans les nuages de mots. Toutefois, nous n'avons pas choisis ces deux mots au hasard pour réaliser cette analyse comparative.

L'un, « communication », représente l'essence même de l'utilité des réseaux sociaux au départ. L'idée de communiquer à distance via différents moyens (photo, messages, appels, visioconférences etc…).

L'autre, « mental health », représente à l'opposé l'un des dégâts collatéraux les plus souvent reproché aux réseaux sociaux : L'impact que ces derniers peuvent avoir sur la santé mentale des utilisateurs.

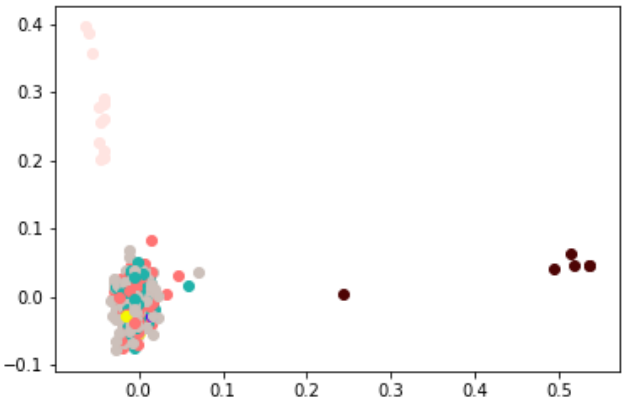
Nous pouvons nous permettre d'analyser l'évolution des occurrences de ces mots-clés au cours de la période 2015-2021 car nous avons sélectionné aléatoirement un nombre égal d'articles pour chaque année.

Ce que met en évidence le graphique ci-dessus c'est tout d'abord la chute libre de l'association des réseaux sociaux avec le principe de la communication. En tout juste six ans, le nombre d'occurrence a été divisé par quatre. Comment expliquer cela ? Alors que la communication est le but même des réseaux sociaux.

La réponse à la question précédente se trouve dans l'étude de la seconde courbe. En effet, si l'on s'intéresse moins à un sujet, c'est qu'une autre problématique a pris le pas sur ce dernier. Ainsi on dénote une explosion du nombre d'occurrences du mot-clé « mental health ». Ce nombre a même été multiplié par six sur la période étudiée, si bien que depuis 2019, on parle davantage des effets des réseaux sociaux sur notre santé mentale, que de leur utilité à la communication.

Nous tenons ici une première réponse à notre problématique. L'impact des réseaux sociaux sur notre comportement se fait entre autres ressentir par une influence, très probablement négative, sur notre santé mentale.

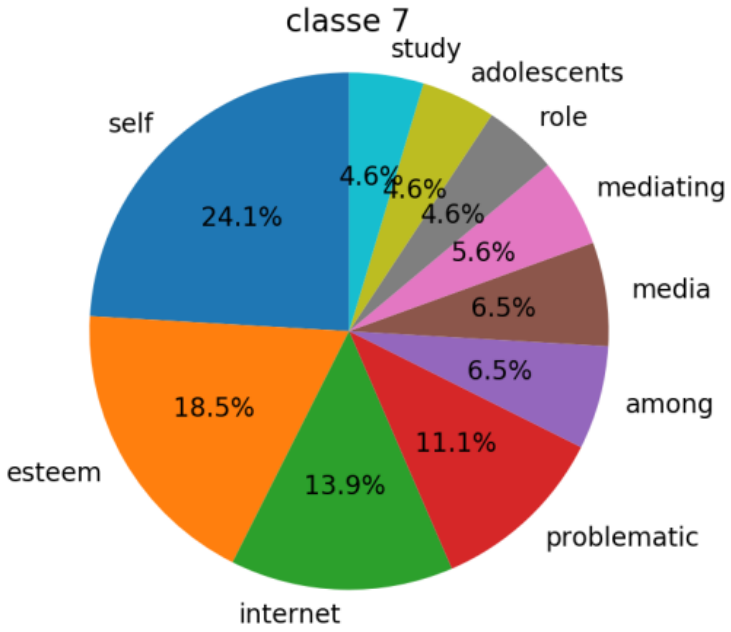
. Analyse des titres :



Clusters identifiés par la méthode des K-means (après ACP)

Jusqu'ici nous analysions uniquement les mots-clés associés aux articles, or nous avons également la possibilité de tirer des informations de leurs titres. Pour ce faire nous avons tout d'abord jugé intéressant d'appliquer l'algorithme des K-means afin d'identifier les sujets

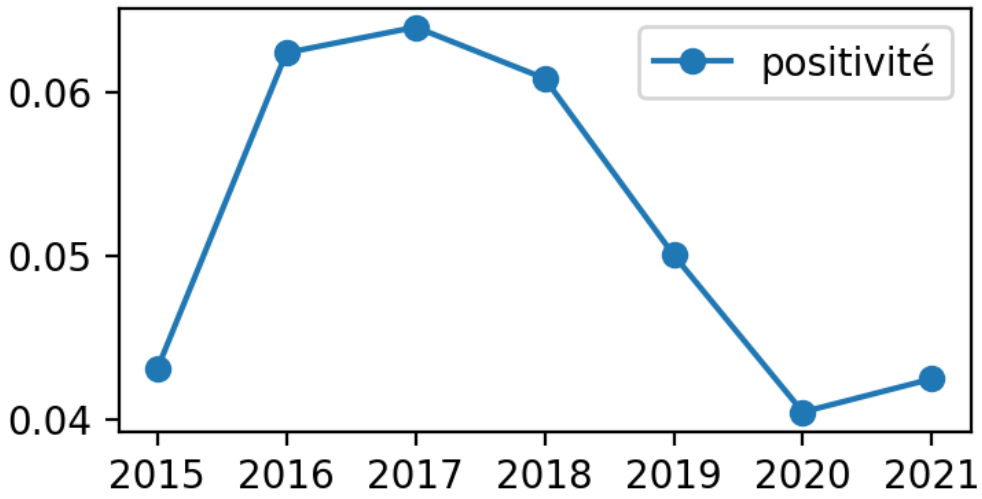
abordés par les articles de notre corpus. Sur le graphique ci-dessus on observe une répartition en huit classes (recommandée par l'algorithme). Une grande majorité de nos articles semblent extrêmement proches en termes de sujet abordés, ce qui est tout à fait normal étant donné que le but de notre équation de recherche était d'isoler des articles traitant d'un même sujet. On observe toutefois une minorité d'articles « à l'écart » des autres, mettant en évidence que sur la totalité des articles sélectionnés, certains ne traitent sans doute pas du sujet voulu. Etant donné leur faible nombre, ces articles ont peu de chance d'influer sur nos observations.



Thèmes abordés par les articles du cluster N°7

Ainsi, nous nous sommes intéressés au contenu d'un des clusters « pertinent », où se retrouve la majorité de nos articles. Les termes qui reviennent le plus dans les titres des articles associés à ce cluster semblent une nouvelle fois confirmer ce que nous avons pu observer lors de l'étude des mots-clés.

En effet, les mots « self » et « esteem » sont largement les plus représentés, ces derniers sont d'ailleurs souvent associés, formant « self-esteem », terme faisant référence à l'estime de soi, un sujet que l'on peut relier aux soucis de comparaison identifié plus tôt. D'autres mots intéressants sont récurrents dans ce cluster comme « adolescent », également déjà vu plus haut, ou encore « problematic », soulignant que l'impact des réseaux sociaux n'est pas forcément positif, au contraire.



Analyse de la positivité des titres

L'analyse des mots-clés nous a montré que les réseaux sociaux étaient de plus en plus associés à des impacts sur notre santé, voire notre santé mentale. Nous avons présumé qu'il s'agissait d'impacts négatifs, mais nous voulions approfondir cette intuition. Ainsi nous avons décidé de réaliser une étude de la positivité des titres composant notre corpus.

Pour ce faire nous avons utilisé la librairie TextBlob qui permet d'estimer la positivité d'une phrase. En appliquant cette fonctionnalité sur les titres de nos sept-milles articles, nous avons pu faire deux observations majeures.

Tout d'abord, comme constaté lors de l'analyse des mots-clés, les réseaux sociaux sont de moins en moins associés à des aspects positifs. On observe en effet sur la courbe ci-dessus une forte baisse de la valeur de positivité retourné par TextBlob depuis 2017 notamment.

De plus, comme attendu, cette valeur, bien que plus élevé en 2017 qu'aujourd'hui, a toujours été extrêmement faible. Une phrase positive aura en effet une valeur proche de 1. D'après notre graphique, cette valeur oscille ici entre 0.04 et 0.06 sur une période de sept ans. Notre intuition semble donc bien se confirmer : Les réseaux sociaux ont une image de plus en plus négative dans la société.

Nuage des mots venant après les termes « impact » ou « risk »

Nous allons conclure notre interprétation avec le nuage de mot ci-dessus, puis répondre à notre problématique. Nous avons construit ce nuage de mot différemment des deux précédents. En effet ce dernier n'est pas lié à un réseau social en particulier, mais s'intéresse davantage à la question de l'impact. Ainsi, les mots qui ressortent le plus sur ce graphique sont ceux qui viennent directement après les termes « impact », « risk » ou « change » dans les titres des articles de notre corpus.

Que remarque-t-on ? D'abord, le mot « health », comme attendu suite à nos précédentes observations, est très bien représenté. Nous pouvons lui associer d'autre mots comme « life » (vie) ou « policy » (habitudes), qui semble également être fortement impacté par les réseaux sociaux. Une autre dimension est représentée par le vocabulaire de l'apprentissage et du milieu scolaire avec notamment « assessment » (évaluation), « education », « research », « learning », ou encore « development » qui sont encore une fois tous plus représentés que des termes comme communication.

Qu'interpréter de tout ça ? Bien entendu ces impacts sur notre vie, nos habitudes, notre santé mentale ou sur notre capacité d'apprentissage, pourraient tout à fait être des impacts positifs.

Cependant, si l'on repense au résultat l'analyse de positivité menée sur les titres, cela a peu de chances d'être le cas. De plus, nous avons également mis en évidence que des termes comme la dépression, l'isolement, ou l'addiction était récurrents. Ces derniers rejoignent l'analyse faisant part d'une mauvaise influence des réseaux sociaux sur notre santé mentale.

Ainsi il apparaît clair que les réseaux sociaux ont un fort impact sur notre vie de tous les jours et sur notre comportement. Malheureusement, il faut garder à l'esprit que ces déviances touchent davantage les enfants et les jeunes adolescents comme nous avons pu l'observer. Premièrement, on ressent donc probablement un impact sur leur santé mentale, de plus en plus de jeunes sont amenés à se comparer entre eux et aux « influenceurs » qui véhicule une image de « vie parfaite ». De ce fait, nombre d'entre eux souffrent du climat de comparaison permanent qui règne sur les réseaux sociaux, et se sentent mal dans leur peau.

Cet impact peut ensuite donc amener à des comportements soient addictifs, soient dépressifs, conduisant à une sensation d'isolement. Ces impacts négatifs semblent également, selon le

dernier graphique, se répercuter sur le niveau scolaire puisque l'on a identifié des impacts négatifs sur le vocabulaire scolaire (éducation, évaluation, apprentissage etc.).

Les réseaux sociaux semblent donc bel et bien avoir un impact sur nos vies et nos comportements. Cet impact étant très probablement négatif dans la plupart des cas comme nous avons pu le voir.

Toutefois plusieurs choses sont à nuancer. Ces impacts ne touchent pas tout le monde, mais plus particulièrement une tranche d'âge, la jeune génération plutôt que les adultes. Cela peut conduire à une fracture entre les générations, celles qui ne connaissent pas vraiment les réseaux sociaux (adultes, parents, grands-parents), peuvent ainsi avoir du mal à comprendre le mal-être des plus jeunes. De plus, nous avons vu que les réseaux sociaux ne sont pas un sujet qui passionnent le monde entier, comme le prouve l'origine des auteurs. Ces derniers étant majoritairement issus des pays les plus avancés (USA, Chine, Royaume-Uni etc..), on peut supposer que l'impact est moindre dans d'autre parties du globe ou la technologie, est à l'heure actuelle, plus difficile d'accès.

Enfin il faut garder à l'esprit que les chercheurs et journalistes ont davantage intérêt à écrire un article mettant en avant les impacts négatifs des réseaux sociaux (afin de prévenir, et de chercher des solutions par la suite), plutôt que vanter leurs aspects évidemment positifs (communication, interaction, partage etc..) car bien connu de tous, d'où le succès des réseaux sociaux à travers le monde.

V. Gestion de configuration

V.1 Environnement de travail

Au cours de ce projet nous avons évolué avec un environnement de travail adapté aux attentes de nos clients. Nous détaillons ci-dessous les principaux outils qui ont composé cet environnement :

Nos ordinateurs portables personnels :

L'ensemble de ce projet s'étant déroulé à distance, nous avons développé la totalité de notre travail depuis nos propres ordinateurs personnels. Ces derniers nous ont également permis de communiquer entre nous et d'utiliser les logiciels que nous allons présenter ci-après.

Jupyter Notebook :

Jupyter Notebook est une application Web de programmation. Dans ce projet nous l'avons utilisé pour coder, en langage python, sur des notebooks les fonctions utiles à notre travail (scraping, visualisation des graphs etc…). Ce code sera disponible en annexe.

SQL Developper :

SQL Developper est un environnement de développement intégré permettant de créer, d'administrer et d'interroger des bases de données. Dans le cadre de ce projet SQL Developper nous a permis d'administrer les informations extraites à la suite de l'exécution du scraping. Nous y avons également nettoyé les données afin de pouvoir les exploiter et les visualiser.

Zoom :

Zoom est une application permettant d'organiser des visioconférences avec un grand nombre de participants. Nous l'avons principalement utilisé pour communiquer avec nos clients et organiser les revues hebdomadaires avec ces derniers.

Discord :

Discord est un logiciel de messagerie permettant également d'organiser des réunions vocales et des visioconférences. Nous l'avons utilisé pour travailler en groupe, faute d'avoir pu se réunir en présentiel. Nous avons également pu y partager nos documents et organiser des « revues intermédiaires » afin de faire des bilans d'avancement avant les revues avec nos clients.

GanttProject :

GanttProject est un logiciel de gestion de projet permettant notamment de réaliser des diagrammes de GANTT. Ici nous l'avons utilisé pour élaborer deux diagrammes de GANTT. Un premier diagramme « prévisionnel » au début de notre projet puis un diagramme « réel » à la fin de notre projet. Ceci ayant eu pour but d'analyser à postériori les différences entre nos prévisions d'avancement et la réalité.

Pack Office :

Le Pack Office est une suite bureautique développée par Microsoft. Nous avons notamment utilisé Word pour la réalisation de ce rapport de projet mais aussi PowerPoint pour certains schémas ainsi que pour la préparation de la soutenance de projet.

V.2 Corpus

Au moment du scraping nous avons dû choisir quelles données extraire des articles résultants de notre équation de recherche.

Voici les informations contenues dans chaque article et que nous avons pensé à scraper afin de former notre corpus :

- Le titre de l'article
- Les dates des articles
- L'abstract (le résumé) de l'article
- Le nom des auteurs
- Les mots-clés associés à chaque article

Nous avions élaboré une première version de notre MCD contenant toutes ces informations. Ceci formant une v0 de notre corpus d'informations à extraire.

Cependant, après une revue avec nos clients, il nous a été fortement conseillé de scraper également la nationalité (l'affiliation) des auteurs. Nous avons ainsi modifié notre MCD et donc constitué une v1 des données à scraper. Toutefois le résultat de notre premier scraping ne fût pas complétement satisfaisant. En effet, scrapé tel que nous l'avions fait, les dates des articles était fournies sur trop de formats différents pour pouvoir en tirer quelque chose. Nous avons par la suite réalisé une version v2 du scraping en utilisant l'API pour récupérer toutes les dates au même format. Nous avons conservé cette v2.

En effet, nous avons jugé cette version v2 de notre corpus suffisamment fournie. Cette dernière nous permet tout de même de réaliser des analyses sur les auteurs, leurs origines, le ton de l'article via l'étude de son titre ainsi que les thèmes et les termes prépondérants de chaque article via l'étude des mots-clés. Notons toutefois, que bien que nous ayons scrapé les abstracts des articles, nous ne les avons pas envoyés dans la base de données. En effet, leur analyse a été jugé peu pertinente par nos clients lors d'une revue.

VI. Assurance qualité

VI.1 Revue

Tout au long du projet nous avons hebdomadairement réalisé des revues afin de contrôler l'avancement de notre travail. Ces revues ont pour la plupart été organisé en présence d'un ou de nos deux clients afin de s'assurer que tout se déroulait correctement et que nous étions « dans les temps ». En outre, nous avons également programmé des revues intermédiaires, avec les membres de l'équipe uniquement, quand cela était nécessaire.

Les tableaux ci-dessous résument l'ensemble des revues réalisées en présence de nos clients. Ceux-ci nous permettent de conserver une traçabilité sur les réunions de projet. Nous y retrouvons la date de la revue, les participants, ainsi que le bilan de la revue (= ce qui a été fait) et les objectifs à venir (= ce qu'il reste à faire).

Revue N°1 : Présentation du projet – 25/01/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
W. Bashoun R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Demande de réalisation d'un projet « Dashboard » -Choix du sujet libre -Par équipe de 4 -Présentation du cahier des charges -Spécification des attentes du client	-Analyser le cahier des charges -Répartir les rôles au sein du groupe -S'organiser sur les moyens de communications -Choisir un sujet -Réfléchir à la source adéquate

Revue N°2 : Choix et validation du sujet – 01/02/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
W. Bashoun R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Les rôles de chacun ont bien été défini -Un groupe de discussion ainsi que des réunions sur Discord ont été misent en place	-Choisir la source à scraper -Définir une équation de recherche adaptée -S'assurer de la cohérence des résultats

		-Elaboration d'un diagramme de GANTT -Le sujet est bien validé -Attention, le choix de la source n'est pas encore décidé ! Ne pas tarder.	-Réfléchir à la conception du MCD

Revue N°3 : Définition de l'équation de recherche – 08/02/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Choix de la source à scraper validé -Equation de recherche à affiner -MCD à revoir, certaines informations sont manquantes	- Améliorer l'équation de recherche -Revoir le MCD afin de le faire valider -Mieux se répartir les tâches entre nous afin de rattraper le retard

Revue N°4 : Validation du MCD – 15/02/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Equation de recherche validée -MCD également validé (suppression de l'abstract) -Attention ne pas oublier l'élaboration du rapport de projet, du retard a été pris sur cette partie là	-Créer la base de données -Commencer à scraper -Davantage se répartir les tâches, assigner un membre à la charge de la rédaction du rapport -Faire des réunions entre nous plus régulièrement -Planifier des séances de travail pendant la semaine de congé

Revue N°5 : Suivi du Scraping – 03/03/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Base de données prête -Scraping bien avancé mais pas encore terminé. Se dépêcher -Attention toujours un peu de retard sur la rédaction du rapport	-Accélérer sur le scraping et absolument l'avoir terminé d'ici la semaine prochaine afin de commencer l'alimentation de la base de données -Avancer le rapport

Revue N°6 : Alimentation de la base de données – 10/03/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
W. Bashoun R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Scraping tout juste terminé -Alimentation de la base commencé ce jour -Ecriture du rapport quasiment à jour	-Finir rapidement le peuplement de la base de données -Commencer, voire finir la visualisation des données d'ici la semaine prochaine -Choisir les moyens de visualisations les plus adaptés -Eventuellement commencer la phase de valorisation des données -Garder le rapport à jour

Revue N°7 : Visualisation des données – 15/03/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
W. Bashoun R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Peuplement de la base de données terminée -Visualisation des données commencées mais pas terminée -Valorisation des données pas commencée -Rapport bien à jour	-Finir rapidement la visualisation des données afin de démarrer au plus vite la phase de valorisation des données. Cette phase est essentielle, c'est la finalité de notre travail, il ne faut donc pas la négliger et essayer de lui consacrer au moins une semaine -Continuer la rédaction du rapport et commencer à préparer la soutenance

Revue N°8 : Revue de fin de projet – 17/03/2021

Participants		Bilan	Objectifs
Clients	Fournisseurs		
W. Bashoun R. Mokadem	T.M. Diallo M.B. Diouf A. Godin Y. Seba	-Visualisation des données toujours en cours mais bientôt terminée. Se dépêcher -Rapport à jour -Préparation de la soutenance pas encore commencée. Attention à ne pas négliger cette partie	-Finir la visualisation des données -Consacrer la dernière semaine à la valorisation et à l'interprétation des résultats -Finaliser la rapport (à rendre lundi avec le code en annexe) -Préparer la soutenance de fin de projet (PowerPoint, présentation de 20 minutes, 10 minutes de question)

VII. Bilan du projet et conclusion

VII.1 Aspects négatifs

Arrivé au terme de ce projet nous ne retenons que très peu d'éléments négatifs. Nous avons simplement envie de mentionner la situation sanitaire et les mesures qui en découlent. Ainsi, nous n'avons pas pu travailler en présentiel ni même nous retrouver pour collaborer ou échanger avec les clients.

VII.2 Aspects positifs

A l'inverse nous retiendrons davantage d'éléments positifs. Tout d'abord la chance d'avoir pu travailler avec les camarades de notre choix, permettant de former un groupe soudé et habitué à travailler ensemble. Ainsi, même sans se rencontrer nous avons toujours pu communiquer efficacement entre nous, que ce soit via Discord, Zoom ou Messenger.

De plus, les attentes du projet nous ont semblé en adéquation avec les cours dont nous avons bénéficié par le passé. Précisons toutefois que nous avons eu la chance d'avoir deux membres de notre équipe assignés au groupe « Scraping » lors du projet Interpromo de janvier. Cet élément a été, selon nous, un bénéfice non-négligeable car il a apporté à certain d'entre nous des connaissances complémentaires à nos cours, et utile pour ce projet.

VII.3 Conclusion

Pour conclure, nous sommes ravis d'avoir mené ce projet à son terme. C'est un sujet que nous avons pris au sérieux et nous nous y sommes investis complétement. Il nous a également permis de développer de nouvelles compétences qui nous seront utiles par la suite.

Enfin nous tenons à remercier nos professeurs le temps qu'ils ont également consacré à ce projet, ainsi que pour l'ensemble des conseils qu'ils nous ont prodigué.

TABLE DES ANNEXES

Annexe 1 : Cahier des charges

Université Paul Sabatier **Année Universitaire 2020 - 2021**
Formation SID

Projet 'Tableau de Bord'
SID Master 1
Cahier des charges
V 5

HISTORIQUE DES MODIFICATIONS

Référence	SID Master 1 /Tableau de bord

Version	Objet de la modification	Date	Auteur(s)	Relecteur
V 0	Création du document	6/12/2016	W.Bahsoun, R Mokadem	-
V 1	Mise à jour du cahier des charges	11/01/2017	W.Bahsoun, R Mokadem	
V2	Mise à jour du cahier des charges	17/01/2018	W.Bahsoun, R Mokadem	-
V3	Mise à jour du cahier des charges	16/01/2019	W.Bahsoun, R Mokadem	-
V4	Mise à jour du cahier des charges	15/01/2020	W.Bahsoun, R Mokadem	-
V5	Mise à jour du cahier des charges	22/01/2021	W.Bahsoun, R Mokadem	-

SOMMAIRE

1. Introduction

1.1 Avant Propos

Le présent document constitue le Cahier des Charges de l'application à développer dans le cadre du projet 'Tableau de bord' en SID Master 1ère année. Il contribue à la création d'un cadre formel de type Client-Fournisseur où les enseignants jouent le rôle du client et les étudiants, celui du fournisseur.

Ce document précise l'organisation de l'équipe de développement et le fonctionnement tout au long du projet.

Le projet proposé dans le cadre de ce bureau d'études couvre les principaux thèmes suivants :

- Les entrepôts de données (data waherouse) [EDD],
- Les bases de données [BDD],
- L'Analyse statistique,
- La visualisation de données
- Le Génie logiciel [GL],
- La Gestion de projet [GPRO],
- La Gestion de Configuration
- …et la programmation (Java, Perl, Apache, Python…).

1.2 Objectifs

L'objectif de ce projet est de développer un système d'aide à la décision élaboré par des analyses détaillées sur des points d'intérêts à partir de la visualisation de données textuelles issues des BDD en ligne.

Ce système va recouvrir plusieurs étapes du processus décisionnel telles que :

1. Recherche et Collecte de données,
2. Préparation de données,
3. Valorisation de données.
4. Diffusion des connaissances.

Nous regroupons ces étapes en trois grandes parties à savoir :

- Recherche et préparation de données.
- Valorisation des données : Data Warehouse,
- Visualisation des données

Ces activités seront supportées par les méthodes de génie logiciel notamment :

- Processus de développement,
- Gestion de projet,
- Ainsi que les règles d'assurance de qualité.

Partie Recherche d'informations et Préparation de données

Recherche et Collecte des données. Cette étape permet de rechercher l'information et d'identifier celle qui est utile dans les sources sélectionnées pour alimenter l'analyse. Nous définirons un domaine d'analyse pour chaque groupe.

La collecte de données se fera sur un périmètre de sujet bien défini à l'avance entre le fournisseur et le client. Cela est raffiné suivant une équation de recherche (pré-traitement de données). Enfin, le client validera par la suite la pertinence des données collectées (extraction de données pertinentes).

Partie Valorisation de données

Stockage de données,

Dans cette partie, un filtrage est appliqué sur la base de données afin d'obtenir un sous ensemble qui répond aux besoins des clients. Les sous ensembles constituent une vue restreinte de la base. Ces sous ensembles sont utilisés comme point d'entrée pour chaque groupe afin que les étudiants puissent appliquer leurs requêtes.

A cause de la crise sanitaire et du contexte d'enseignement en distanciel, le système de gestion de bases de données utilisé sera Oracle. Les étudiants pourront implémenter leur base de données à distance via SQL Developer. Oracle constitue un SGBD très répandu et très puissant dans le milieu socio-économique.

Interrogation de données

Cette partie consiste à générer des structures de données pour chaque groupe. Cela est possible en appliquant des requêtes sur les sous ensembles de la base de données produite lors de l'étape précédente. En fonction des données manipulées par chaque groupe, une méthodologie sera proposée par l'enseignant pour exporter les données afin de peupler l'entrepôt de données. Par la suite, les requêtes doivent être écrites. Ces requêtes concernent les données présentes dans les différentes tables obtenues. Suivant le type des requêtes, les sorties (résultats) sont classifiées en quatre (4) types (dimensions):

- Une variable.
- Un vecteur.
- Une matrice.
- Un cube.

Partie Visualisation de Données

Les fonctions relatives à la visualisation de données sont essentielles pour réussir la présentation d'un travail de veille et pour convaincre les décideurs à travers une présentation lisible, pertinente et concise.

La visualisation de données consiste à représenter graphiquement (icônes, carte, arbres…) une information (données, processus, relations, concepts) souvent abstraite et/ou à très grande volumétrie.

Pour chaque groupe, nous définissons différents types de sorties adaptés à chaque type de requête tels que les histogrammes d'évolution 2D et 3D, les cartes géographiques, les nuages de mots,…etc.

2. Documents applicables et de références

2.1 Documents applicables

[CONT] Contrat de développement.

2.2 Exemples de documents de référence

[GL] Cours « Génie Logiciel»
Wahiba Bahsoun

[DAWA] Cours « Data Waherouse»
Gilles Hubert

[CONC] Cours « Concepts Fond. BD »
Franck Morvan

[GPRO1] Cours et TD « Gestion de Projet »
Wahiba Bahsoun

2.3 Terminologie

Tous les termes utilisées dans le contexte de ce projet doivent etres expliquées par l'équipe de projet.

3. Exigences sur le processus de développement

3.1 Organisation

[ORG-00] La réalisation intégrale du projet est à la charge du Fournisseur constitué d'un groupe de 4 étudiants de Master M1 de la formation SID. La promotion actuelle est d'un effectif de 36 étudiants. Cela conduit alors à la formation de 9 groupes.

[ORG-01] Le Fournisseur doit assigner, pour toute la durée du projet :
- un chef de projet,
- un responsable de gestion de configuration,
- un responsable Assurance et Contrôle Qualité.

Ces fonctions doivent être assurées par des ressources distinctes.

[ORG-03] L'organisation de l'équipe de développement mise en place par le Fournisseur sera présentée au Client.

[ORG-04] Chaque réunion d'avancement se fera en présence d'au moins un représentant du Client, du chef de projet, du responsable de gestion de configuration et du responsable Assurance et Contrôle Qualité.

[ORG-05] Le Fournisseur doit être en mesure de présenter au Client l'état d'avancement du projet pendant toute sa durée.

3.2 Livrables

[LIV-00] Le Fournisseur doit livrer au Client :
- Un rapport du projet,
- Le cahier de Recettes,
- le code source et les données associées.

Les exigences relatives à ces éléments sont décrites au chapitre suivant.

[LIV-01] Les livrables mentionnés en [LIV-00] seront fournis par voie électronique (wahiba.bahsoun@irit.fr et riad.mokadem@irit.fr).

[LIV-02] Les livrables mentionnés en [LIV-00] seront remis au Client lors de la soutenance du projet (date prévisionelle : lundi 22 mars 2021).

3.3 Planification (Application Scrum)

[PLAN-00] Le projet sera réalisé dans le cadre de séances de cours et TP d'une durée de 2 heures ou de 4 heures. Les dates et heures précises sont communiquées dans l'emploi du temps.

[PLAN-02] Les réunions d'avancement seront hebdomadaires.

[[PLAN-04] La soutenance aura lieu vers la fin du mois de mars 2021 (date prévisionnelle : mercredi 24 mars 2021), ceci correspond à la présentation des résultats aux clients.

4. Exigences sur le produit logiciel

4.1 Plan de développement

[PD-Q-01] [GPRO] est applicable au plan de développement.

[PD-Q-02] Le rapport du Projet doit présenter les dispositifs d'assurance Qualité mise en place par le fournisseur, et plus précisément :

- En Introduction, la description du sujet d'analyse
- Les méthodes, techniques et outils utilisés,
- Les règles de développement appliquées sur toutes les phases de développement. en conception, en codage,
- Les dispositifs de suivi du projet
- Et une interprétation rigoureuse du sujet d'analyse.

[PD-Q-03] Le plan de développement doit présenter les dispositifs de gestion de configuration mise place pendant toute la durée du projet.

5. Exemples de Projets sur les sujets suivants :

Exemples de sujets traités 2015-2016:

Bitcoin, COP21, Harcèlement sexuel, Euro 2016, Les addictions

Exemples de sujets traités en 2016-2017:

Internet of things (maisons connectées) , La thérapie par réalité virtuelle, Impression 3D des organes humains, Inégalité homme / femme, Impact des écrans sur les enfants, Données abérantes.

Exemples de sujets traités en 2017-2018:

Blockchains, L'impact des attentats, Les catastrophes naturelles, Les cyber attaques, L'évolution du e-sport, e-commerce, SmartGrid, Légalisation du cannabis, Ted Talks.

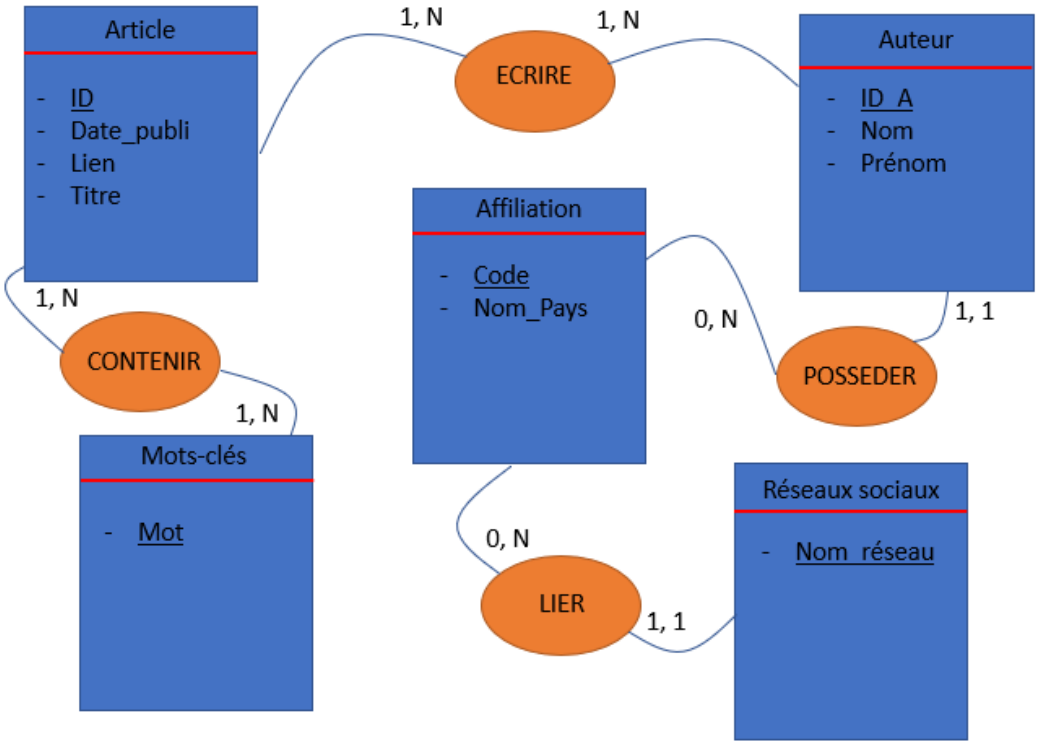
Exemples de sujets traités en 2018-2019:

Pollution due au plastique, Véhicules autonomes, Catastrophes naturelles, Brexit, La déforestation, Le flux migratoire, La coupe du monde au Qatar (Corruption dans le sport), Gilets jaunes.

Exemples de sujets traités en 2019-2020:

Dopage dans le sport, feux de foret, fuite des cerveaux, impact des écrans, impact des OGM sur la santé, impact du plastique sur l'environement, impact de l'organisation des jeux olympique, PMA, thérapie sonore, le tourisme spatial

Annexe 2 : Modèle conceptuel des données (MCD) / Modélisation logique des données (MLD)



MCD

Article(ID , Date_publi , lien,titre)

Auteur(Id_A , Nom , Prenom , #NomPays)

Affiliation(Code , pays)

Mots-clés(Mot)

Reseaux_sociaux(Nom_reseau , #NomPays)

Contenir(#IdArticle , #Mot)

Ecrire(#ID , #id_A)

MLD

Annexe 3 : Extraits de code

L'ensemble de notre code est disponible dans le fichier .zip, mais en voici quelques extraits relatifs à des étapes clés du projet.

```
## all the libraries we used for data enhancement
import streamlit as st #Create an online dashboard
import numpy as np
import pandas as pd ##data table manipulation
import matplotlib as mpl #plot the data
import matplotlib.pyplot as plt #plot the data
from wordcloud import WordCloud, ImageColorGenerator #Generate
word cloud of data
from PIL import Image #load graph images
from sqlalchemy import create_engine #establish connection
with database
import cx_Oracle
import json #allow to transferring data
import plotly.express as px #plot
import SessionState #optimisaiton of the dashboard

#all the data stored at the start of the dashboard in order to
optimize resources
session_state = SessionState.get(imageClusters=0,figPays =
0,dfAuteurPaysCount=0,createview =0,dicoPaysCode=0,start
=1,a=0, b=0,c=0)
```

Ensemble des librairies utilisées

```
 engine =
create_engine('oracle://dlt1940a:Bint@Cherif96@telline.univ-
tlse3.fr:1521/etupre')
    connection = engine.raw_connection() #establish connection
with database
    # Check if account exists
    try:
        cursor = connection.cursor()
        cursor.execute("select level n from dual connect by
level < 10") #try a random request
        for row in cursor:
            print(row)

        cursor.close()
        connection.commit()
    finally:
        connection.close() #close connection
```

Connexion à la base de données

```
    def create_view(requete,colonnes=[]) :

        """ Documentation
      This function returns a Data Frame

      Parameters :
      requete : the requete SQL

      colonnes : The columns of data frame
      """
        connection = engine.raw_connection()
        cursor = connection.cursor()
        if len(colonnes) ==0 :
            return
pd.DataFrame.from_records(cursor.execute(requete))
        else:
            return
pd.DataFrame.from_records(cursor.execute(requete),columns=colo
nnes)
```

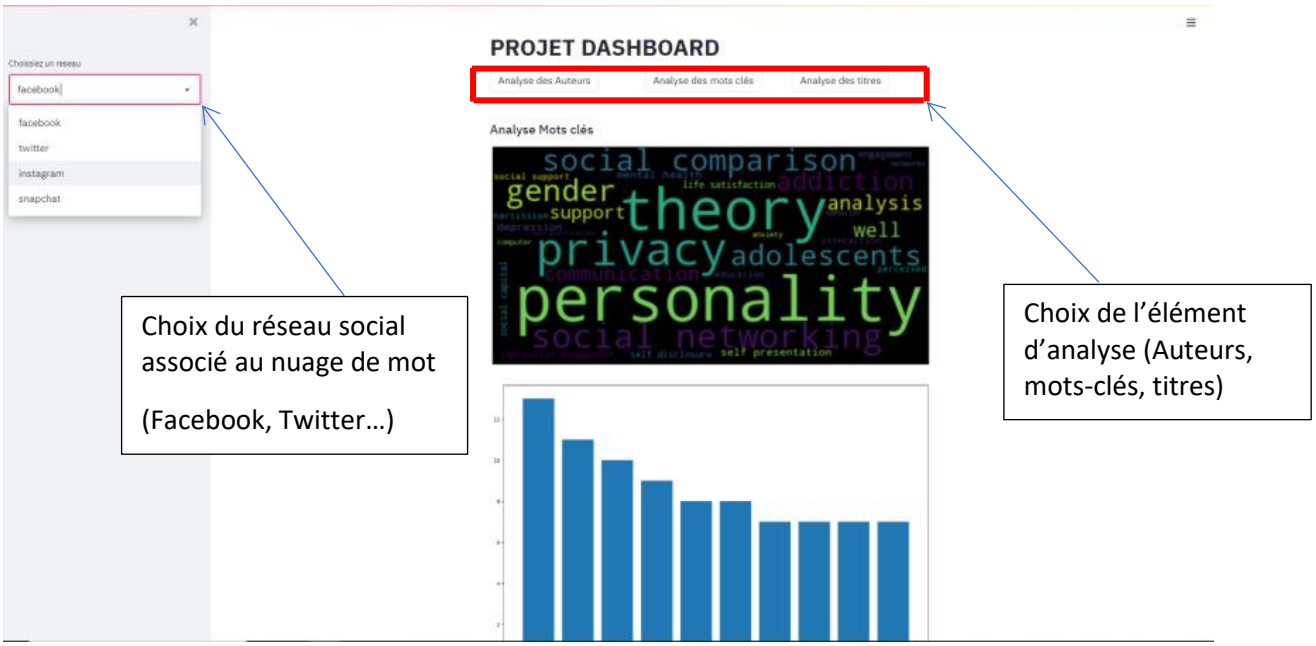
Connexion à la base de données (suite)

```
#----------------------------------------PLOT AUTHOR
DISTRIBUTION-----------------------------------------
    requete = 'select Nom_pays,count(*) from Auteur GROUP BY
Nom_pays'
    dfAuteurPaysCount = session_state.createview(requete,
['pays','count'])
    dfAuteurPaysCount['code_pays'] = [dictPaysCode[i] if i in
list(dictPaysCode.keys()) else ''  for i in
dfAuteurPaysCount['pays']]
    session_state.dfAuteurPaysCount = dfAuteurPaysCount

    df = px.data.gapminder().query("year==2007") #get list of
countries
    session_state.figPays =
px.choropleth(session_state.dfAuteurPaysCount[session_state.df
AuteurPaysCount['code_pays'].isna()==False],
locations="code_pays",
                    color="count",
                    hover_name="pays",

color_continuous_scale=px.colors.sequential.Plasma)
#----------------------------------------------------------------
-------------------------------------------------------
```

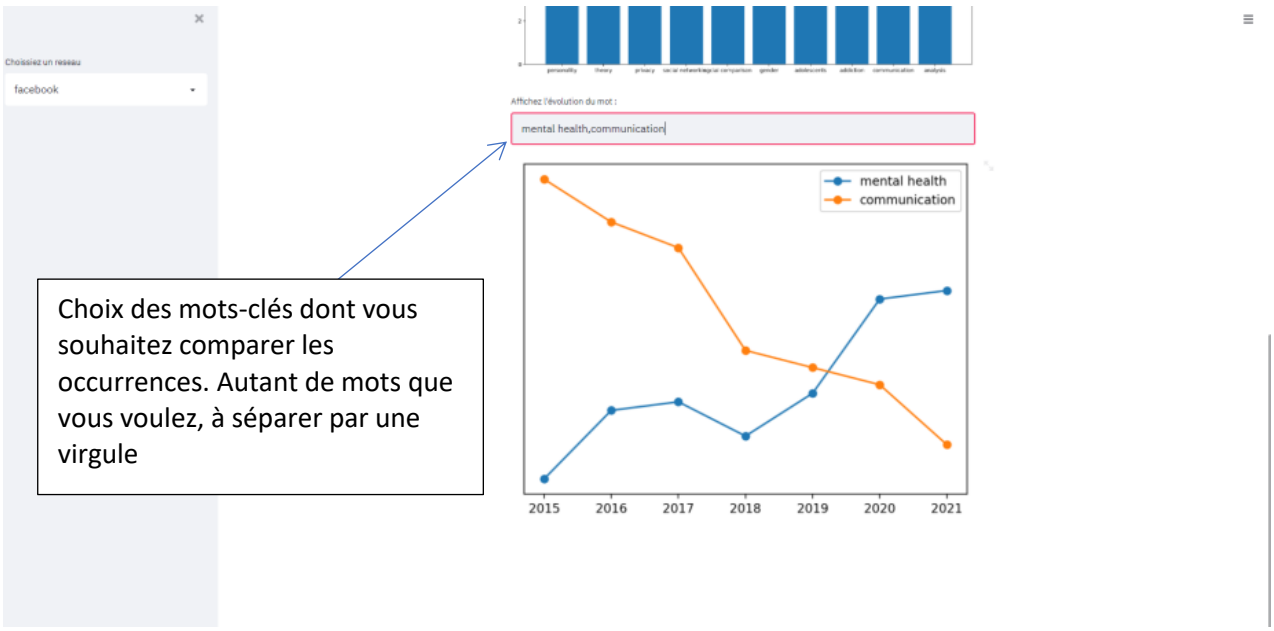
Affichage de la carte en fonction du nombre d'auteur par pays

Annexe 4 : Visualisation sur Streamlit

L'accès à la visualisation de nos graphiques via Streamlit est disponible depuis le code (+ un fichier texte explicatif). En voici toutefois quelques extraits.



Capture d'écran N°1 de l'interface Streamlit



Capture d'écran N°2 de l'interface Streamlit

INSTRUCTIONS POUR LANCER LE DASHBOARD AVEC STREAMLIT :

-ouvrir un invité de commande

-exécuter : pip install streamlit

-se rendre dans le dossier streamlit (contenu dans le zip envoyé) depuis l'invité de commande

-Exécuter : streamlit run streamlit3.py

-Vous pouvez indiquer votre mail ou laissez vide

-Copier le lien localhost qui s'affiche sur le cmd et le collé dans un navigateur

Annexe 5 : Quelques requêtes SQL

Exemple de requêtes utilisées

Récupérer le nombre d'auteur par pays:

```
requete = 'select Nom_pays,count(*) from Auteur GROUP BY Nom_pays'
```

Récupérer le classement des réseaux sociaux pour chaque pays:

```
requete = "select count(*),auteur.nom_pays,motscle from ecrire,auteur,article,conte
nir where ecrire.id_a = article.id_a and ecrire.id_auteur=auteur.id_auteur and cont
enir.id_a=article.id_a and contenir.motscle in ('facebook','snapchat','twitter','in
stagram','reddit','whatsapp','telegram','tiktok','') group by (auteur.nom_pays,mots
cle) order by auteur.nom_pays,count(*) desc"
```

Récupérer les key-words associés a un réseau social :

```
requeteMotsAssociesResaux = "select  count(*),motscle from contenir where id_a in
(select id_a from contenir where motscle='"+mot+"') and motscle not in ('internet',
'site','use','pinterest','online','sites','network','whatsapp','social networks','t
witter','snapchat','social newtork','networking','social network','social media','s
ocial','media','instagram','facebook','"+mot+"')  group by (motscle) order by (coun
t(*)) desc"
```

Récupérer l'évolution de l'occurrence d'un mot clé en fonction des années :

```
requete = "select MotsCle, count(MotsCle) from Contenir,Article WHERE  MotsCle LIK
E '" + str(word)+ "%' and Date_publi LIKE '"+ str(year) + "%' and Contenir.ID_A = A
rticle.ID_A GROUP BY (MotsCle)"
```

Récupérer les titres d'articles et leur date de parution :

```
requete = "SELECT Date_publi, Titre_art FROM Article"
```