La pollution maritime

Projet tableau de bord

Clients :
Madame Wahiba Bahsoun
Monsieur Riad MOKADEM

Fournisseurs :
Enzo MARTINEAU (Chef de projet)
Cyril GAILLARD (Database administrator)
William AZZOUZA (Responsable de gestion de configuration)
Fanny ROLLET (Responsable Assurance et Contrôle qualité)

UNIVERSITÉ TOULOUSE III PAUL SABATIER

SID

BIG DATA
FORMATION - RECHERCHE - INDUSTRIE

Remerciements

Nous tenons à remercier en premier lieu la formation Statistique et Informatique Décisionnelle (SID) qui nous a permis de réaliser ce projet. Nous remercions principalement nos professeurs référents, Madame Wahiba Bahsoun et Monsieur Riad MOKADEM , qui nous ont accompagné et conseillé durant toute la durée du projet.

Nous remercions également l'ensemble des professeurs de notre formation qui nous ont apporté les connaissances nécessaires pour mener à bien ce projet.

Historique du document

Libellé de la modification	Participants	Dates de mise à jour
MAJ partie : Gestion de configuration	William Azzouza	25/01/2019
MAJ partie : Démarche de développement (SADT)	Enzo Martineau	28/01/2019
MAJ partie : Démarche de développement (Collecte de données & Préparation des données)	Cyril Gaillard et William Azzouza	08/02/2019
MAJ partie : Assurance et contrôle qualité (Charte de codage)	Fanny Rollet	18/02/2019
Restructuration du rapport	Cyril Gaillard, Enzo Martineau, Fanny Rollet, William Azzouza	04/03/2019
MAJ parties : Objectifs du document, Organisation de développement, Démarche de développement	Cyril Gaillard, Enzo Martineau, Fanny Rollet, William Azzouza	05/03/2019
MAJ parties : Objectif/But du document, Démarche de développement	Cyril GAILLARD	06/03/2019
MAJ parties : Organisation de développement, Démarche de développement	Enzo Martineau, William Azzouza, Cyril GAILLARD	10/03/2019
MAJ parties : Démarche de développement, Assurance qualité	Enzo Martineau, Cyril Gaillard, Fanny Rollet, William Azzouza	15/03/2019
MAJ parties : Démarche de développement, Bilan de projet	Cyril Gaillard, William Azzouza	19/03/2019

Sommaire

Remerciements 1

Historique du document 2

I. Objectif et but du document 5

A) Objectif du document 5

B) But du document 5

II) Documents de référence et documents applicables 6

A) Documents applicables 6

B) Documents de référence 6

C) Problématique 7

III. Terminologie 8

IV. Organisation de développement 9

A) Ressources humaines 9

B) Planification 10

C) Ressources matérielles 14

V. Démarche de développement 15

A) Recueil des données 16

1 - Sélection des sources cibles 16

2 - Extraction des données 17

i) Récupération des liens 17

ii) Recherche des métadonnées des articles pour chaque site 18

iii) Création des fichiers bruts 18

3 - Préparation des données 19

i) Nettoyage des contenus, titres et descriptions de chaque article 19

ii) Construire un vocabulaire par article 19

iii) Créer les fichiers json finaux 20

B) Traitement des données 21

1 - Création du schéma relationnel 22

2 - Création de la base de données 23

3 - Peuplement de la base de données 24

C) Administration des données 25

1 - Création des requêtes 25

2 - Exécution des requêtes 27

3 - Extraction des informations 27

D) Diffusion des données 28

1 - Sélection des informations et traitement statistique 28
2 - Visualisation des résultats 29

VI) Gestion de configuration 40

VII) Assurance et contrôle qualité 41
21 Janvier 2019 41
28 Janvier 2019 41
08 Février 2019 42
18 Février 2019 42
25 Février 2019 43
04 Mars 2019 43
15 Mars 2019 43
18 Mars 2019 44
22 Mars 2019 44

VIII) Bilans de projet 45
Bilans personnels 45
1 - William 45
2 - Cyril 45
3 - Enzo 46
4 - Fanny 46
B) Bilan global 46

Annexe 47
Annexe 1 : Travail préparatoire 47
Annexe 2 : Charte de codage 49
A) Convention logiciel 49
1 - Langage Python 49
2 - Langage SQL 49
B) Convention de nommage des variables, fonctions 49
C) Commentaires 50
D) Lisibilité du code 50
1 - Langage Python 50
2 - Langage SQL 51

I. Objectif et but du document

A) Objectif du document

L'objectif de ce projet, réalisé dans le cadre de la formation SID au niveau Master 1, est de faire travailler des groupes de quatre étudiants ensemble sur un sujet d'actualité qu'ils ont choisi. C'est un projet qui s'inscrit dans un contexte client-fournisseur puisqu'un client (professeurs de la formation SID) demande à un fournisseur (groupe de quatre étudiants de cette même formation) de leur fournir une analyse détaillée concernant un sujet d'actualité. Pour mener à bien ce projet, les étudiants doivent développer un système d'aide à la décision élaboré par des analyses détaillées sur un certains nombre de points d'intérêts et à partir de la visualisation de données textuelles issues de bases de données en ligne.

L'objectif est de récupérer des données de médias numérique en les scrappant, de mettre en forme ces données sous la forme d'une base de données, puis de les valoriser sous forme de visualisation. Tout ça dans un but de diffusion de connaissance.

La relation client-fournisseur est un concept qui met en évidence que pour satisfaire les deux parties il faut s'entendre sur un échange (contrat formel ou informel) le plus clair possible.

B) But du document

Ce document est donc destiné au client (deux professeurs de la formation SID). Il a pour but de recenser l'ensemble des étapes réalisées par le fournisseur (quatre étudiants SID) afin de mener à bien la mission confiée. Il devra donc présenter en détail l'ensemble des phases réalisées par les étudiants, les difficultés rencontrées, les solutions trouvées et surtout répondre à la problématique donnée: proposer une analyse détaillée d'un sujet d'actualité.

II) Documents de référence et documents applicables

A) Documents applicables

Pour bien commencer un projet, il est indispensable de rédiger un cahier des charges, c'est un document qui doit être pensé lors de la conception d'un projet.

Cet outil de pilotage est primordial pour définir les besoins et les spécifications d'un projet et, par dessus tout, les rendre compréhensibles par tous.

B) Documents de référence

Au cours de notre formation nous avons suivi des cours indispensables à la réalisation de ce projet, les voici :

Cours de Génie logiciel par Mme Bahsoun :
- Processus de développement
- Qualité du logiciel

Cours d'extraction d'information dans les document textes par Mr Moreno
- Extraction d'informations à partir de textes
- Evaluation : Méthodologie et métrique

Cours de concepts fondamentaux des bases de données par Mr Morvan
- Introduction aux bases de données
- Elaboration du schéma conceptuel
- Conception du schéma logique

Cours de langage de requête par Mme Lechani
- Le langage d'interrogation SQL
- Le langage de contrôle SQL
- Le langage PL/SQL

Cours de développement logiciel par Mme Bahsoun
- Concepts fondamentaux des modèles de développements
- Système Qualité
- Modèle itératif, Notions de livrables

C) Problématique

Nous avons choisi un sujet d'actualité qui nous concerne tous : les problèmes environnementaux. Nous avons choisi de travailler sur la pollution maritime et plus particulièrement sur le 7ème continent également appelé vortex de déchets.

Ce vortex de déchets est situé dans un zone du gyre subtropical du Pacifique nord. La masse de plastiques présente au sein de tous les océans est estimée à plus de sept millions de tonnes, dont environ 269 000 tonnes de déchets plastiques flottants selon une vaste étude internationale parue dans la revue PLOS ONE. Cependant, les chercheurs à l'origine de cette étude ont tenu à souligner qu'il s'agissait plus d'un ordre de grandeur.

Des expéditions "7ème continent" ont déjà été effectuées afin d'étudier et de comprendre les raisons de ce regroupement de déchets plastiques à cet endroit précis du globe.

Nous étudierons donc l'ampleur de la pollution maritime, puis nous nous demanderons quels en sont les acteurs et comprendre l'impact mondial du phénomène.

III. Terminologie

SCRAPPER : Récupérer les données sur un site web en s'appuyant sur le code html. Dans notre cas, il s'agissait de s'appuyer sur la librairie python bs4 (soup) qui nous a permis de collecter nos données.

PUSH : Action sur GIT : celle d'envoyer des fichiers

PULL : Action sur GIT : celle de récupérer des fichiers

COMMIT : Action sur GIT : celle de valider le travail accompli

MERGE : Lorsque deux individus sur un github ont fait des modifications sur des fichiers différents, mais qu'un seul des deux a envoyé ces travaux. L'autre individu (après commit) devra effectuer un merge pour récupérer les travaux (modifié) du premier individu.

CONFLICT (GIT) : Lorsque deux individus ont effectué des modifications sur le même fichier, et qu'une version antérieure sur le git existe. Un conflit apparaîtra si l'une des deux parties essaie de push son travail. Il devra alors choisir quels sont les éléments à conserver afin de garantir une version stable de son application sur le git.

MCD : Le Modèle Conceptuel de Données est une représentation graphique de haut niveau qui permet facilement et simplement de comprendre comment les différents éléments sont liés entre eux à l'aide de diagrammes codifiés

SADT : Structured Analysis and Design Technics, permet de spécifier les fonctions de notre projet, en se basant sur le cahier des charges délivré par le client

SGBD : Le Système de Gestion de Base de Données permet de contrôler toutes opérations sur une base de données ainsi que de l'administrer.

LIBRAIRIE (PYTHON) : Ensemble de fonctions permettant de faciliter le travail du programmeur qui n'aura qu'à lire la documentation de ces librairies afin d'en comprendre le fonctionnement. Elles réalisent une tâche précise et feront gagner du temps au programmeur.

TF : Term Frequency, le nombre de fois qu'un mot va apparaître dans un document texte. Il permet de savoir si le mot est pertinent dans un contexte d'analyse de corpus de texte.

IV. Organisation de développement

A) Ressources humaines

Notre équipe de projet se compose de quatre membres de M1 de la formation Statistique et informatique décisionnelle (SID). Chacun des étudiant a un rôle défini. Enzo MARTINEAU est le Chef de projet, William AZZOUZA le responsable de la gestion de configuration, Fanny ROLLET est quand à elle responsable Assurance et Contrôle qualité, et pour finir Cyril GAILLARD est Database administrator.

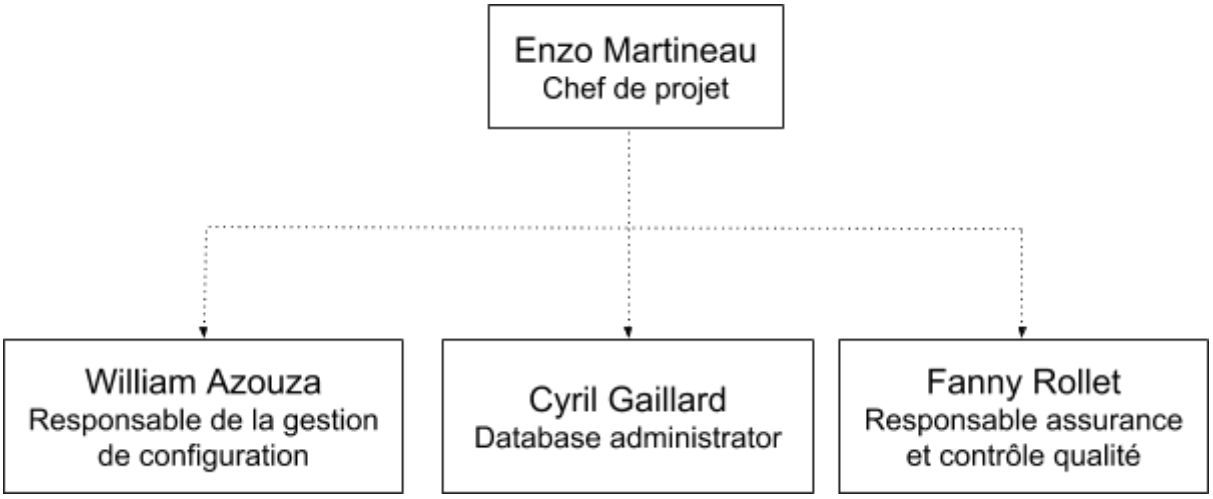
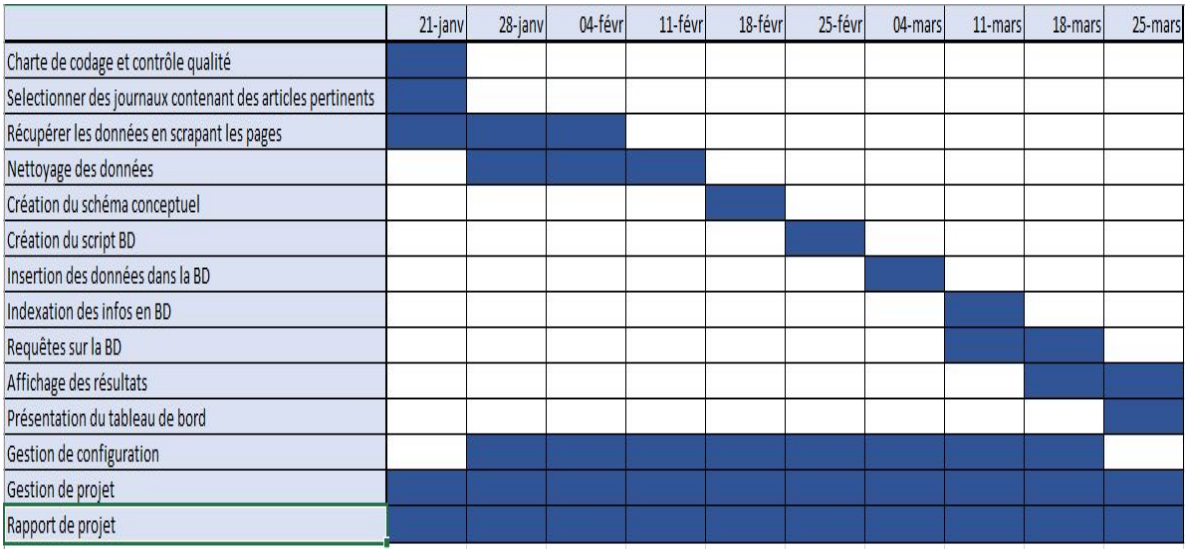


Figure 1 : Organigramme du groupe de projet

Concernant notre manière de travailler, chaque semaine nous faisions une réunion afin de voir les avancements du projet, puis nous listions les différentes tâches à effectuer au cours de la semaine suivante afin que chacun puisse avancer de son côté et que le travail ne soit pas fait en double.

B) Planification

Pour réaliser projet, nous avons eu a effectué différentes tâches. Celles-ci sont représentées dans le diagramme de Gantt ci dessous :



	21-janv	28-janv	04-févr	11-févr	18-févr	25-févr	04-mars	11-mars	18-mars	25-mars
Charte de codage et contrôle qualité	■									
Selectionner des journaux contenant des articles pertinents	■									
Récupérer les données en scrapant les pages	■	■	■							
Nettoyage des données		■	■	■						
Création du schéma conceptuel					■					
Création du script BD						■				
Insertion des données dans la BD							■			
Indexation des infos en BD								■		
Requêtes sur la BD								■	■	
Affichage des résultats									■	■
Présentation du tableau de bord										■
Gestion de configuration		■	■	■	■	■	■	■	■	
Gestion de projet	■	■	■	■	■	■	■	■	■	■
Rapport de projet	■	■	■	■	■	■	■	■	■	■

Figure 2 : Diagramme de Gantt

Le diagramme de Gantt nous permet de visualiser dans le temps les diverses tâches qui ont composé notre projet. Chaque tâche à été effectué par un ou plusieurs membres du groupe.

On retrouve cette répartition des tâches dans le tableau ci-dessous, qui représente le tableau de Gantt avec pour chaque tâche effectuée, le prénom du ou des membre(s) qui a/ont effectué cette tâche.

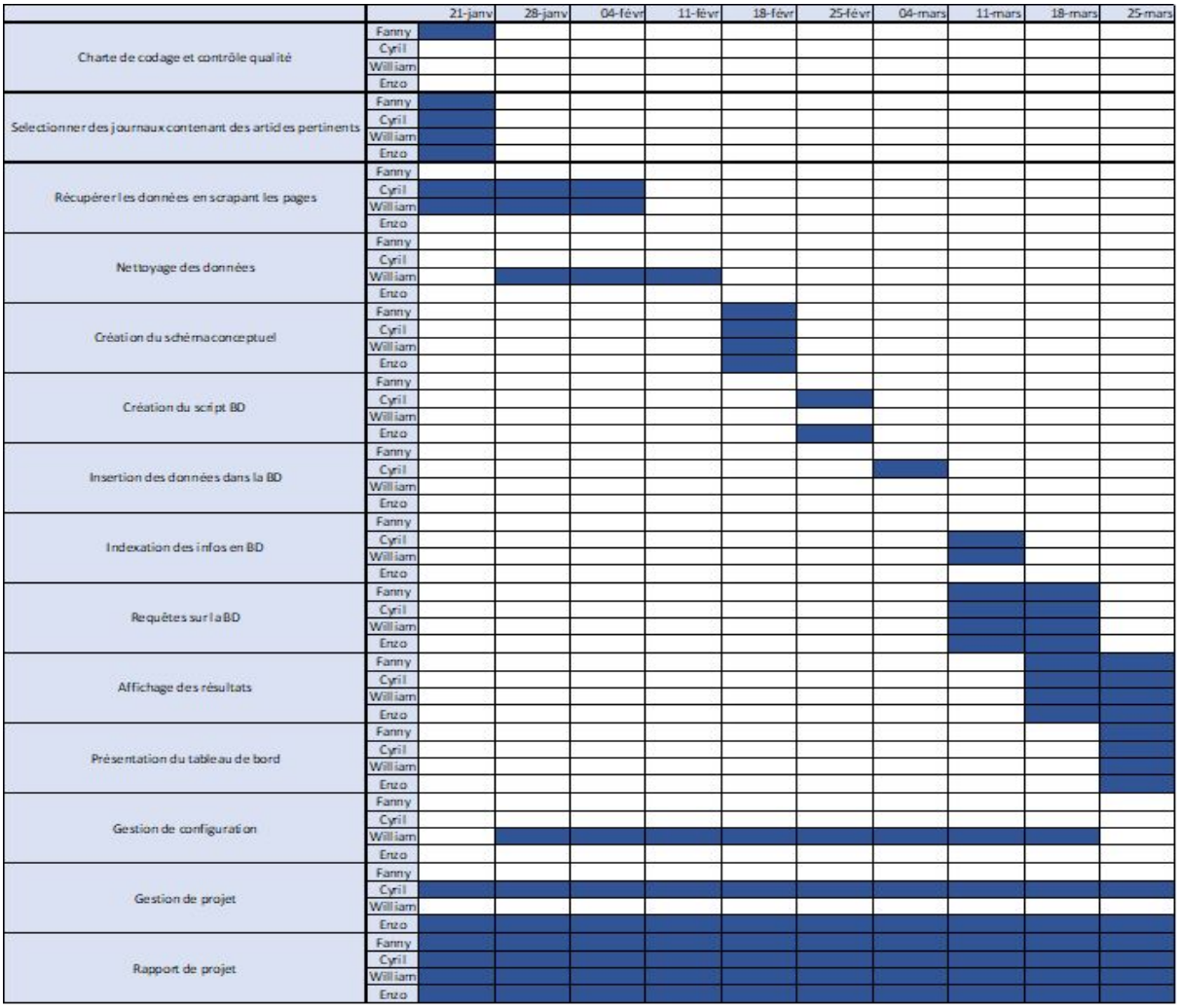


Figure 3 : Diagramme de Gantt détaillé

Suivi de ce diagramme de Gantt détaillé nous avons réalisé un graphe de Pertt afin d'estimer les retards éventuels que pourraient prendre certaines tâches et ainsi réagir de la meilleure manière pour y pallier

Ce graphe de Pert s'appuie sur un tableau de Pert que l'on retrouve ci-dessous.

Indexe activité	Activité	Date prévu	Date réel	Activités préalables
A	Sélectionner des journaux contenant des articles pertinents	21/01	21/01	
B	Récupérer les données en scrappant les pages	21/01	21/01	A
C	Charte de codage et contrôle qualité	04/02	04/02	
D	Nettoyage des données	11/02	11/02	C
E	Création du schéma conceptuel	18/02	18/02	D
F	Création du script BD	25/02	25/02	E
G	Insertion des données dans la BD	01/03	04/03	F
H	Indexation des infos en BD	11/03	04/03	G
I	Requêtes sur la BD	18/03	18/03	H
J	Affichage des résultats	25/03	25/03	I
K	Présentation du tableau de bord	25/03	25/03	J
L	Gestion de configuration	18/03	18/03	
M	Gestion de projet	25/03	25/03	
N	Rapport de projet	25/03	25/03	M,J

Figure 4 : Tableau de Pert

Le Graphe de Pert est ensuite déduit de ce tableau comme on peut le voir ci-dessous.

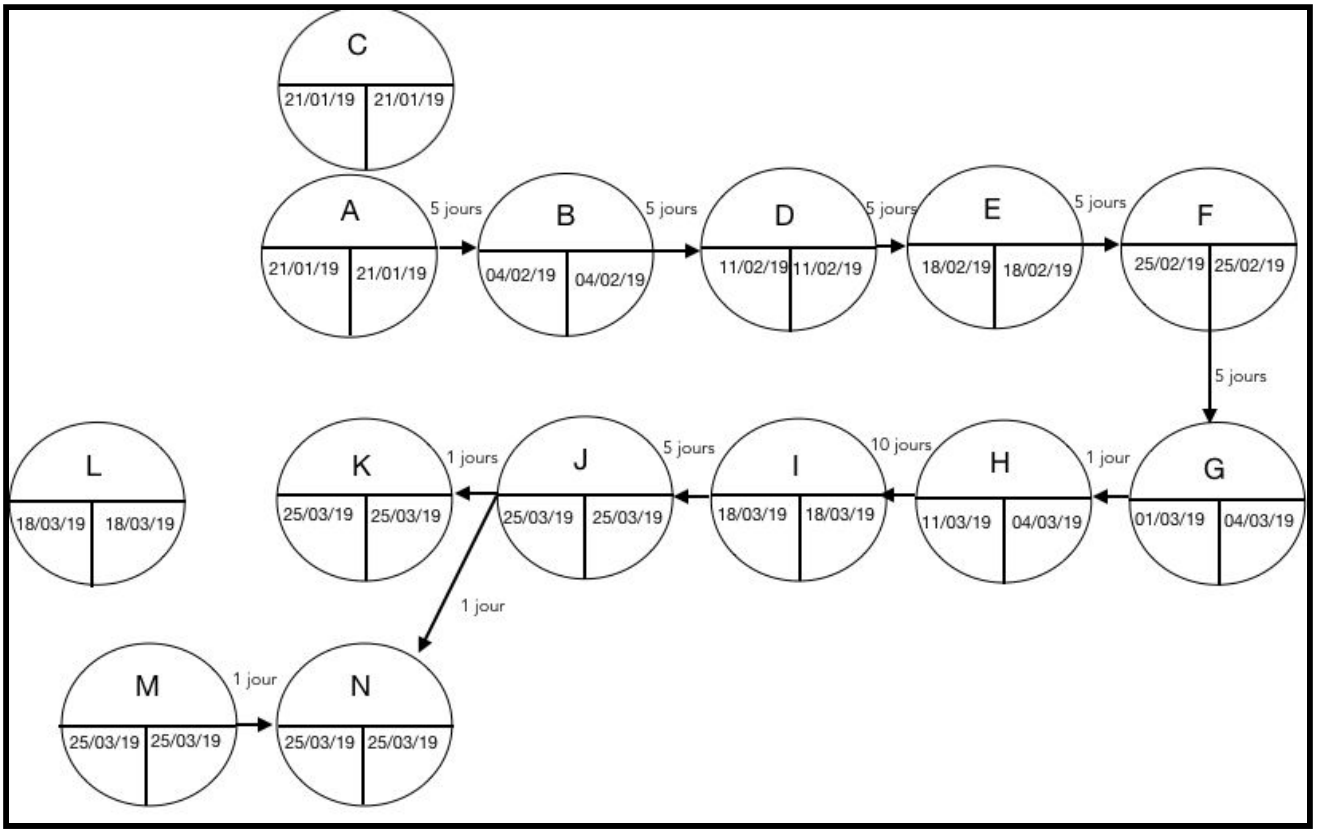


Figure 5 : Graphe de Pert

C) Ressources matérielles

Nous avons eu besoin d'un certains nombres de ressources matérielles afin de réaliser ce projet. Dans un premier temps chaque collaborateur disposait d'une machine sur laquelle il pouvait avoir l'entièreté des documents concernant le projet.

Concernant les ressources matérielles que nous avons utilisé, pour le développement nous avons utilisé le langage de programmation Python. Il nous a servi à développer les parties concernant le scrapping des données, le nettoyage des données ainsi que pour l'insertion de ces dernières dans une base données. Pour nous aider à réaliser ces tâches, nous nous sommes appuyés sur plusieurs librairies python. Puis, pour la partie base de données, nous avons utilisé le système de gestion de base de données SQL SERVER qui est développé par Microsoft. De plus, pour réaliser le MCD de notre base de données, nous avons employé le logiciel en ligne DB Diagram. Enfin, nous nous sommes appuyés sur Microsoft Excel de faire l'analyse de nos données à travers divers graphiques.

Concernant l'organisation du projet nous avons utilisé l'outil de gestion de projet en ligne Trello qui nous permet de lister les différentes étapes et tâches du projet indexées dans le temps et affectées à des ressources humaines. Cet outils nous a permis de pouvoir suivre facilement l'avancée du projet. Nous avons également utilisé l'outil Microsoft Excel afin de réaliser les différents diagrammes, tableaux et graphiques concernant la gestion de projet.

V. Démarche de développement

Nous avons choisi la méthode SADT comme processus de développement pour logiciel informatique. La méthode SADT est une méthode d'analyse par niveaux successifs d'approche descriptive d'un ensemble quel qu'il soit. On peut donc appliquer cette méthode à notre projet.

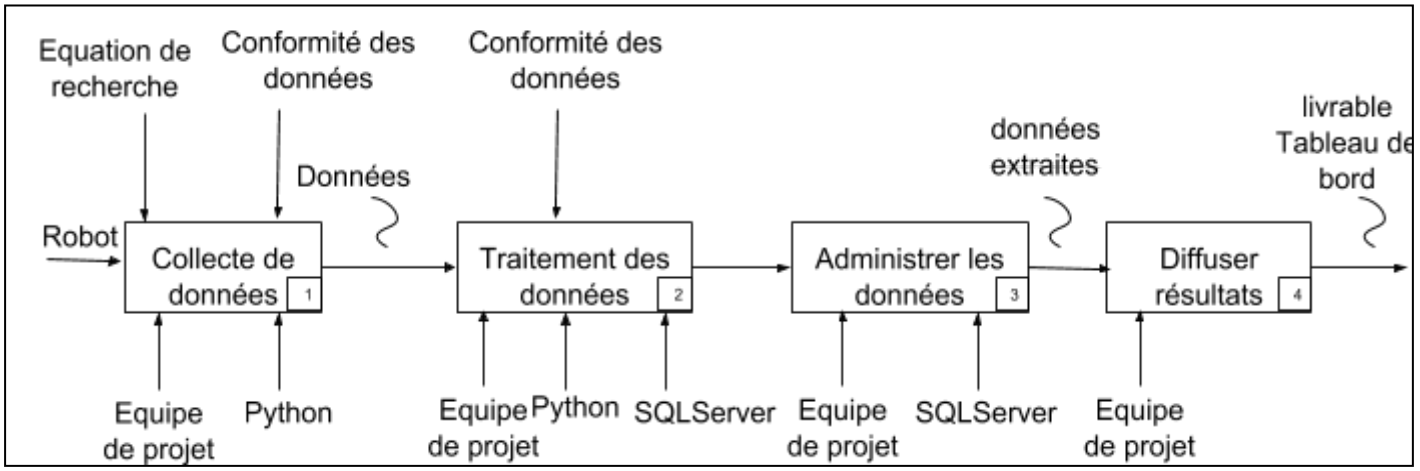


Figure 6 : A0 - Réalisation du projet

Le projet comprend 4 étapes :

- La collecte des données, qui consiste à trouver les données pertinentes pour le projet.
- Le traitement des données, qui est la phase de création de la base de données.
- L'administration des données, qui consiste à extraire les données propres de la bases de données afin de pouvoir les analyser par la suite
- La diffusion des résultats est la phase d'analyse des données qui permet la visualisation de ces données pour délivrer le tableau de bord.

Chacune de ces étapes est découpée en plusieurs tâches décrites dans les différents SADT présentés ci-dessous.

A) Recueil des données

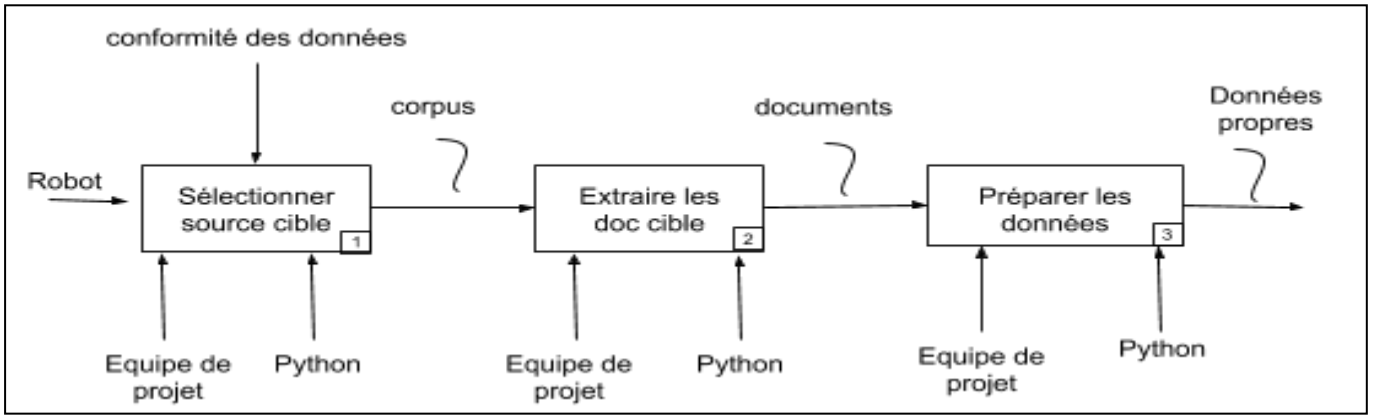


Figure 7 : A1 - collecter les données

Nous avons découpé l'étape de collecte des données en trois tâches consécutives :

- la sélection des sources cibles consistant à rechercher sur internet les sites qui peuvent faire l'objet d'analyse pour notre sujet.
- l'extraction des documents cibles choisis lors de la tâches précédente.
- la préparation de données propre afin de pouvoir les insérer dans la base de données par la suite.. (décrire)

1 - Sélection des sources cibles

La première étape du projet fut réfléchir à une idée de sujet sur lequel nous aurions assez de données afin d'en tirer des analyses. Nous avons donc décidé de consacrer les deux premières séances de travaux pratiques au choix du sujet. En effet le sujet est la partie la plus cruciale à choisir ; il déterminera toute la suite du projet. Nous avons pensé à plusieurs idées de sujet mais c'est le sujet concernant le 7ème continent qui nous intéressait le plus et sur lequel on pouvait récupérer le maximum d'informations.

Après avoir choisi le sujet nous avons dû lister l'ensemble des sources / liens internet qui nous semblaient pertinents et qui nous permettraient de construire notre analyse. Cette phase de recherche a été étalé sur une semaine; entre les deux premières séances de TP. L'ensemble de nos recherches se trouvent en annexe 1.

Nous avons retenu trois médias qui à eux trois nous permettraient de récupérer un peu plus de 1000 articles. Nous avons donc utilisé :

- http://www.journaldelenvironnement.net/
- https://www.lemonde.fr/
- http://www.septiemecontinent.com/

2 - Extraction des données

L'enjeu de cette partie était qu'à partir des sites que nous avions identifiés et trouvés pertinents, de récupérer les données des articles et leurs métadonnées associées.

Comme dit précédemment nous nous sommes principalement focalisés dans un premier temps sur les sites journalistiques : Le Monde, Le Journal de l'Environnement et Le 7e Continent. Ce sont tous les trois des sites de journaux qui nous semblaient simples à scrapper et regroupant assez d'informations pour tirer les analyses que nous voulions faire au départ.

i) Récupération des liens

Pour réaliser cette partie nous avons choisi de nous baser sur le code source de chaque page. Dans un premier temps et pour tous les sites de journaux cités précédemment, nous avons voulu récupérer les liens des journaux. C'est pourquoi pour chaque journaux, nous sommes allés sur leur page de recherche dédiée. Cela nous a permis de faire un premier tri sur les articles qui étaient susceptibles de nous intéresser.

Pour les sites Le Monde et 7e Continent nous avons placé en mots clés "plastique" et "océan". Pour le site journal de l'environnement il s'agissait du mot : "plastique" seulement. Ces mots clés nous ont servi de base pour sélectionner les articles intéressants pour notre sujet :

- Le Monde : environ 548 articles
- Le 7e continent : 153 articles
- Le Journal de l'environnement : 1100 articles

Ces articles n'étaient bien sûr pas tous localisés dans la même page, il a donc fallu trouver des méthodes de navigation de page en page, pour chaque site. La solution était similaire pour chaque site, elle consistait en un changement de numéro de page dans l'url de chaque page de recherche. Il suffisait alors de l'incrémenter jusqu'à atteindre la dernière page de recherche voulue sur chaque site.

ii) Recherche des métadonnées des articles pour chaque site

Une fois le problème de la navigation résolu, nous sommes passés à la seconde étape qui consistait en l'analyse des codes sources de chaque page, dans le but d'identifier les éléments qui nous intéressaient. Nous nous sommes mis d'accord pour récupérer pour chaque article, quelque soit le site, les informations suivantes :

- son titre
- son contenu
- son ou ses auteur(s)
- sa date de publication
- son lien

Ensuite, nous avons analysé le code HTML de chaque site. HTML est le langage interprété par les navigateurs web pour pouvoir afficher du contenu sur une page web, c'est un langage à balises. Nous avons donc dû décrire l'ensemble des balises pertinentes pour chaque site, afin de récupérer les informations importantes mentionnées plus tôt.

iii) Création des fichiers bruts

Nous avons choisi de regrouper l'ensemble des informations présentes dans chaque site (média numérique) dans un fichier distinct. Comme nous avions trois sources, nous aurons donc trois fichiers json. Nous avons utilisé la même structure pour chacun de nos trois fichiers json ; la voici :

```
{
"numero_article_1": {
        "authors":"auteur",
        "content":"contenu_article",
        "link":"lien_article",
        "newspaper":"nom_journal",
        "publication_date":"date_article",
        "title":"titre_article"
        },
"numero_article_2": {
        "authors":"auteur",
        "content":"contenu_article",
        "link":"lien_article",
        "newspaper":"nom_journal",
        "publication_date":"date_article",
        "title":"titre_article"
        },
}
```

Ceci étant fait nous avions pour chaque site, des fichiers brut de données qui étaient près à être nettoyés.

3 - Préparation des données

Cette phase consiste au nettoyage des données brutes que nous avions récupérées à partir des sites (etape précédente). En se basant sur la structure des fichiers json que nous avions bâti, le but était maintenant de modifier leur format et de nettoyer les données pour faciliter leurs insertions en base de données.

i) Nettoyage des contenus, titres et descriptions de chaque article

Les clés de nos fichiers json contenant du texte brut, devaient être nettoyées en premier. Ce nettoyage consistait à enlever l'ensemble des mots vides présents dans le corps des articles, leur titre, et leur description. Nous avons donc dû construire une liste de mots vides en français, et nous l'avons combiné à d'autres listes (prépositions, conjonctions, déterminants et pronoms).

A partir de ces listes, nous avons pu obtenir un premier nettoyage sur les différents éléments composants nos articles. Le but étant de sélectionner les mots ne figurant pas dans ces listes pour permettre de faire une dernière sélection d'articles.

ii) Construire un vocabulaire par article

En résultat de la première étape nous possédions un vocabulaire par article, ce qui nous a permis de faire des traitements d'enrichissement sur les mots contenus dans ce vocabulaire. En effet, à l'aide des librairies nltk et numpy de python, nous avons pu enrichir les mots avec des différents éléments : son tf (term frequency) et son pos tagging. Ces deux éléments nous permettent :

- de connaître le nombre de fois que le mot apparaît dans le contenu, le titre ou la description de son article
- de savoir à quelle type de mots le mot en question appartient (exemple : un nom commun , un nom propre, un adverbe, etc.)

De plus, nous ne voulions pas perdre le fait de savoir qu'un mot faisait parti soit de la description, soit du titre. En effet cette information est précieuse pour les traitements que nous voulions réalisé par la suite. C'est pourquoi nous avons ajouté deux autres éléments pour chaque mots du vocabulaire :

- is_title : savoir si un mot fait parti du titre de l'article ou non
- is_description : savoir si un mot fait parti de la description ou non

iii) Créer les fichiers json finaux

Basé sur les quelques ajouts fait précédemment, nous pouvions désormais bâtir nos fichiers finaux. Avec la mise en place d'un vocabulaire par article et les différents éléments associés aux mots, les clés "content", "description", "title", nous étaient inutiles désormais. Ainsi, nous les avons supprimé pour alléger le poids des fichiers finaux. En effet pour des raisons d'optimisation nous devions au maximum diminuer la taille de ce fichier json, afin de diminuer drastiquement le temps d'exécution des procédures d'insertion en base de données. La forme générale des fichiers json finaux est décrite ci-dessous :

```
"numero_article":  {
        "authors":"7emeContinent",
        "link":"lien_article",
        "newspaper":"nom_journal",
        "publication_date":"date_article",
        "words": {
                "mot_1": {
                        "tf":1,
                        "is_title":0,
                        "is_description":1,
                        "pos_tag":"pos_tagging"
                                },
                "mot_2": {
                        "tf":2,
                        "is_title":1,
                        "is_description":0,
                        "is_country":0,
                        "pos_tag":"pos_tagging"
                        }
        }
}
```

B) Traitement des données

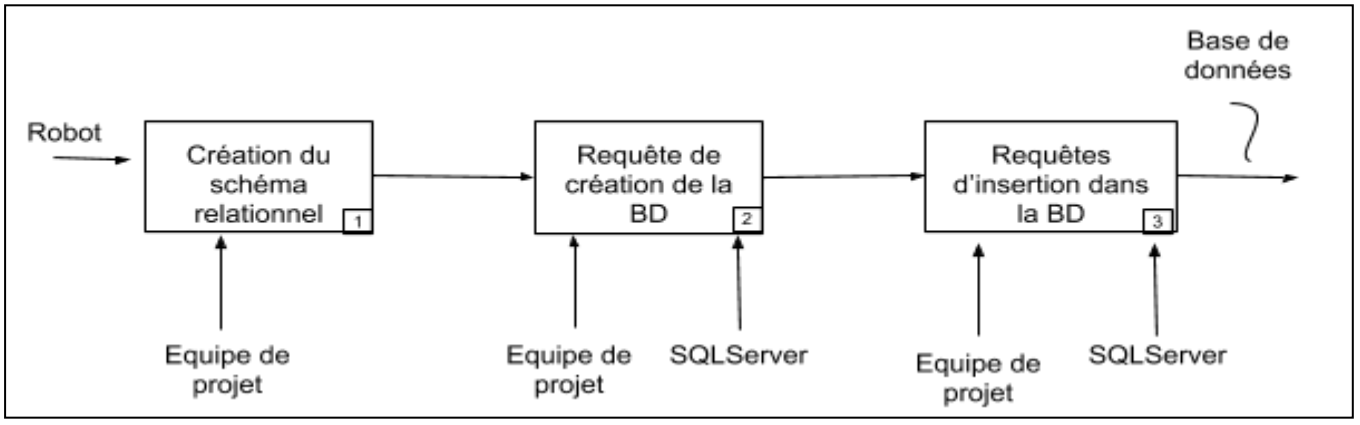


Figure 8 : A2 - Traiter les données

Le traitement des données est composé de 3 étapes :

- La création du schéma relationnel afin de pouvoir par la suite insérer correctement toute les données propres et préparées dans la phase de recueil des données.
- La création de la bases de données dans SQL Server, grâce à des requêtes SQL de création de tables.
- La création des requêtes d'insertion dans la base de données pour insérer les données propres préparé dans la phase de recueil des données.

Suite à la phase de recueil, nous avions nos fichiers permettant d'optimiser l'insertion en base données, nous devions construire le schéma conceptuel de cette base de données. C'est pourquoi nous avons fait un schéma entité association.

1 - Création du schéma relationnel

En se basant sur les éléments contenus dans les fichiers et les différents traitements que nous voulions réalisés, nous avons produit une première version de notre MCD que voici ci-dessous :

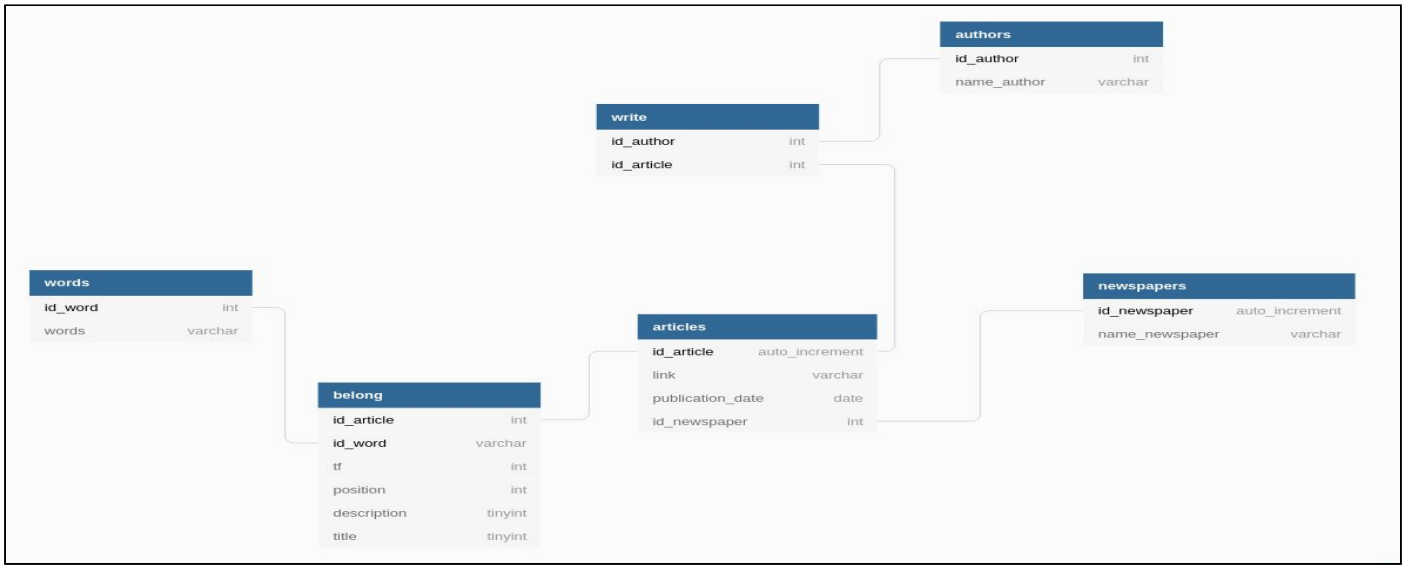


Figure 9 : MCD V.0

Ensuite, après réflexion avec nos enseignants encadrants nous l'avons fait évoluer afin qu'il s'adapte aux mieux aux besoins du clients, ce qui nous donne le MCD ci-dessous :

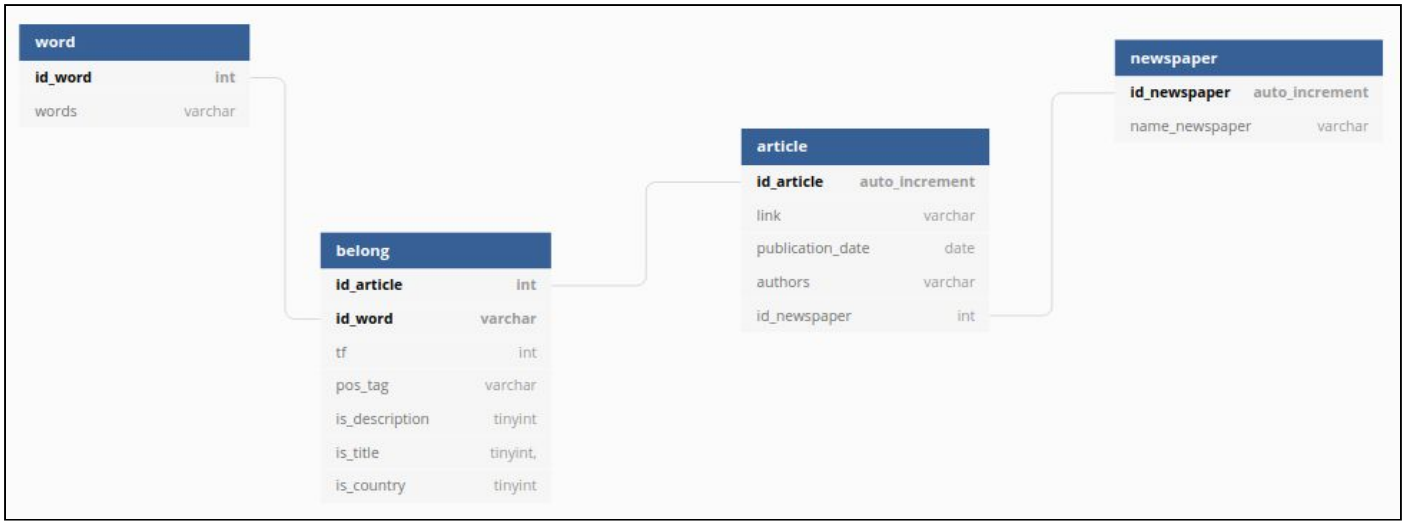


Figure 10 : MCD V.1

2 - Création de la base de données

La table centrale de notre MCD est la table Article ; celle-ci regroupe l'ensemble des articles que l'on a récupéré grâce à l'étape de *Collecte des données*. Dans cette table sont stockés le lien unique de l'article (link), la date de publication (publication_date), l'auteur(s) (author) ainsi que l'id du journal auquel appartient cet article.

La table Article est donc relié à la table Newspaper qui comprend la liste des journaux que l'on a choisi lors de la phase de *Recherche de données*. On y retrouve donc l'id du journal (id_newspaper) ainsi que son nom (name_newspaper). Les tables Article et Newspaper sont liées par la présence d'une clé étrangère dans Article (id_newspaper). En effet, un article appartient à un et un seul journal.

La table suivante est la table Word, elle regroupe tous les mots non vides présents dans l'ensemble de nos articles. On y retrouve donc l'intitulé du mot (label_word) ainsi que son pos_tagging (pos_tag) c'est à dire la catégorie du mot.

Enfin nous retrouvons la table Belong, c'est celle-ci qui nous permettra de faire toutes nos analyses. On y retrouve l'id de l'article (id_article), l'id du mot (id_word), le tf (tf) c'est à dire la fréquence d'apparition du mot dans l'article concerné, le fait de savoir si le mot est dans la description (is_description) de l'article, savori si le mot est un pays ou non (is_country), et enfin déterminer si le mot est présent dans le titre de l'article (is_title). Ces deux derniers champs sont des booléens, ils peuvent être défini à True tous les deux. Cette table fait le lien entre la table Words et la table Articles. on y retrouve donc deux clés étrangères id_article et id_word. C'est donc dans la table Belong qui nous pourrons déterminer quels mots sont les plus importants pour quels articles grâce notamment à leur fréquence d'apparition mais aussi dû au fait qu'il soit présent dans la description de ce dernier ou bien dans son titre.

Une fois le MCD validé, nous avons créé le script de création de notre base de données. Il suffisait alors de suivre le MCD et de créer les tables dans l'ordre :

- Newspaper
- Article
- Word
- Belong

Il fallait également pensait à mettre les contraintes de clés primaires et étrangère sur les tables. Nous avons également fait le choix de créer un index sur la table Belong sur les champs id_article et id_word afin d'accélérer le temps de calcul lors de nos requêtes. (*cf. script de création de la base de données en annexe)*

3 - Peuplement de la base de données

Après avoir créé notre base de données sur SQL Server en local, l'étape suivante fut de peupler notre base de données. Pour se faire, nous avons fait le choix de créer un script python qui insérera les données. Afin de faciliter les choses dans ce script python, nous avons commencé par créer quatre procédures (une pour insérer des données dans chaque tables) ; en plus d'insérer dans la table concernée, ces procédures vérifiées que l'on ajoutait pas les données en double dans les tables afin d'éviter d'avoir des redondances et des incohérences.

Une fois la connexion établie entre le script python et la base de données en local, nous parcourions les fichiers json obtenu à l'étape précédente en insérant où il faut les données au fur et à mesure. Le temps total pour insérer l'ensemble de nos données (environ 1000 articles) est d'environ 30 minutes.

C) Administration des données

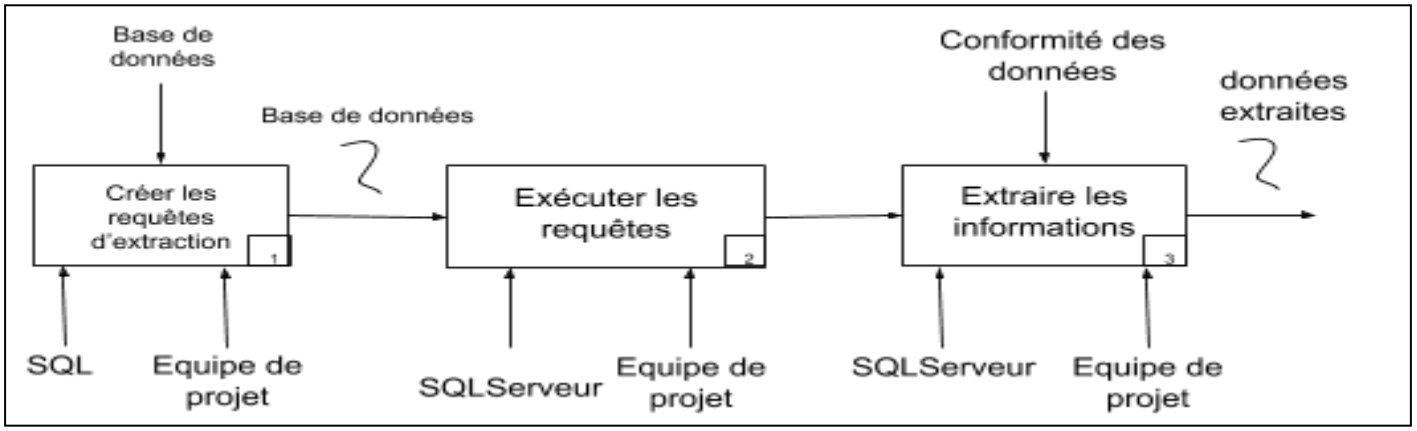


Figure 11 : A3 - Administrer les données

Cette étapes est elle aussi composée de 3 tâches :

- La création des requêtes d'extraction des données pertinentes.
- L'exécution des requêtes.
- L'extraction des informations pertinentes à visualiser pour les tableaux de bords

1 - Création des requêtes

Dans un premier temps, nous voulions trouver des requêtes qui correspondaient à l'analyse que nous voulions faire de notre projet. Elles devaient nous permettre de répondre à notre problématique de départ.

Dans le tableau ci-dessous on retrouve l'ensemble de nos requêtes. En vert les requêtes que l'on conserve, et en rouge celles que l'on n'a pas utilisé, en effet nous avons essayé de les utiliser mais elles ne nous apportaient pas d'informations pertinentes.
Les requêtes elles mêmes se trouvent en annexe (**cf. page ??**)

Rôle de la requête	Remarques sur la requêtes
- Q01 : Récupération du tf cumulé de chaque mot (les 30 premiers)	Tri nécessaire sur certains mots, car jugés non pertinents par nos soins.
- Fréquence chaque mot dans chaque article	Le vocabulaires des articles diversifié, paramétrisation des requêtes trop importantes pour obtenir des résultats

	pertinents
- Nombre d'articles par auteur	Beaucoup d'auteurs similaires pour certains journaux, ou alors aucun auteurs n'avaient pu être récupéré lors du scrapping des sites
- Q02 : Le nombre de fois que chaque pays apparaît par article (30 plus grands tf)	
- Q03 : Les 30 plus grands tf group by mot/journal	
- Les mots qui apparaissent le plus par date	Ne permet pas de faire une analyse fine, trop de paramétrisation au niveau de la requête.
- Q04 : Le nombre de pays par date (année)	
- Q05 : Le nombre de mots par année	
- Q06 : Le nombre de fois qu'un mot pertinent apparaît dans un article	
- Q07 : pays par mois avec label associé à chaque tranche de date	
- Q08 : pays par année avec label associé à chaque tranche de date	
- Q09 : nombre de pays par mois avec label associé à chaque tranche de date	
- Q10 : nombre de pays par année avec label associé à chaque tranche de date	
- Q11 : mots qui apparaissent le plus dans les descriptions	Permet de déterminer l'importance de certains mots
- Q12 : mots qui apparaissent le plus dans les titres	Permet de vérifier que nos articles sont cohérents dans l'ensemble
- Q13 : mots qui apparaissent le plus dans les titres et description d'articles	On est sûr que ces mots sont très pertinents

Pour la partie diffusion des données, la majorité de nos requêtes furent ordonnancés pour permettre une lecture plus rapide des données.

2 - Exécution des requêtes

Ces requêtes furent ensuite exécuté sur la base données en local (contenant au préalable l'ensemble des données scrappées). Cela nous a permis de récupérer les résultats de ces dernières.

3 - Extraction des informations

Pour pouvoir analyser les données et en tirer l'information pertinente nous devions récupérer l'ensemble des résultats de requête dans un fichier excel. C'est pourquoi, nous avons créé un script qui insère automatiquement les résultats de requêtes dans un fichier excel. Chaque résultat de requête était ainsi stocké sur sa propre page excel, et était prêt à être analysé.

D) Diffusion des données

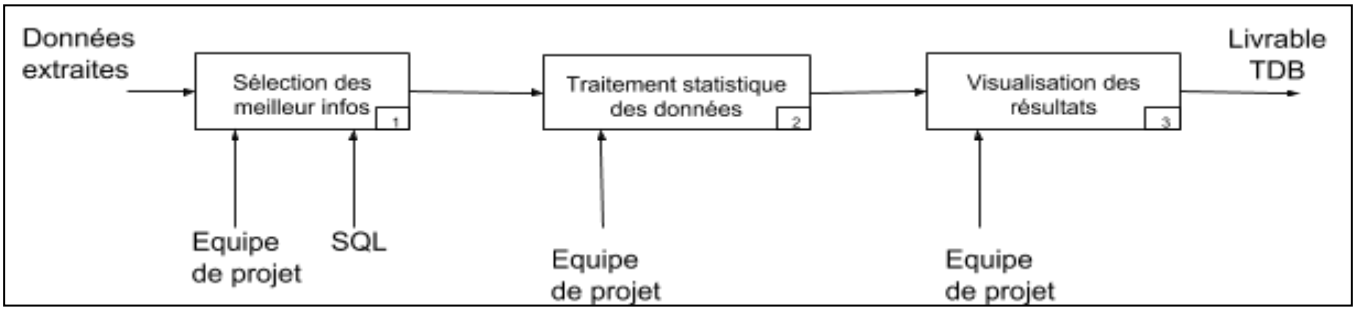


Figure 11 : A4 - Diffuser les résultats

La dernière étape du projet est la validation et la diffusion des résultats. Pour cela nous avons là aussi séparé cette étape en 3 tâches différentes :

- La sélection des informations les plus importantes
- Le traitement statistique des données afin d'en sortir les informations les plus pertinente pour répondre à notre problématique.
- La visualisation des résultats qui consiste à mettre en forme les résultat. La création du tableau de bord.

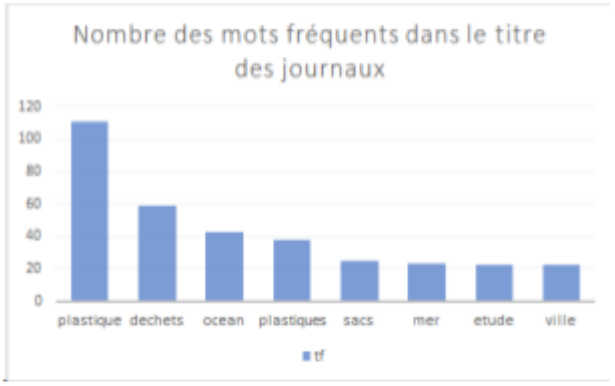
1 - Sélection des informations et traitement statistique

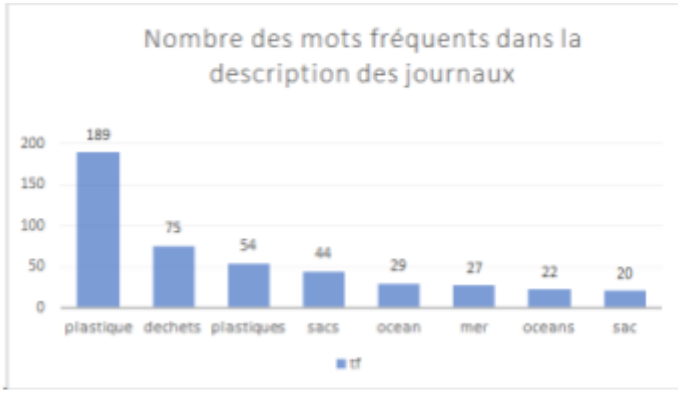
Durant cette étape nous avons essayé de tirer des résultats de requête, de l'information pouvant être analyser. Nous nous sommes donc basé sur des des tableaux excel contenant l'information, et nous avons ensuite créé de nouveaux tableaux regroupant le plus souvent des données croisées. Ces tableaux croisés nous ont permis de bâtir des graphiques et illustrations plus simplement en regroupant l'information pertinente au bon endroit.

De plus, les données nous étant livré de manière triée, nous n'avions plus beaucoup de manipulations à faire concernant l'affichage sous forme de graphique de ces dernières.

2 - Visualisation des résultats

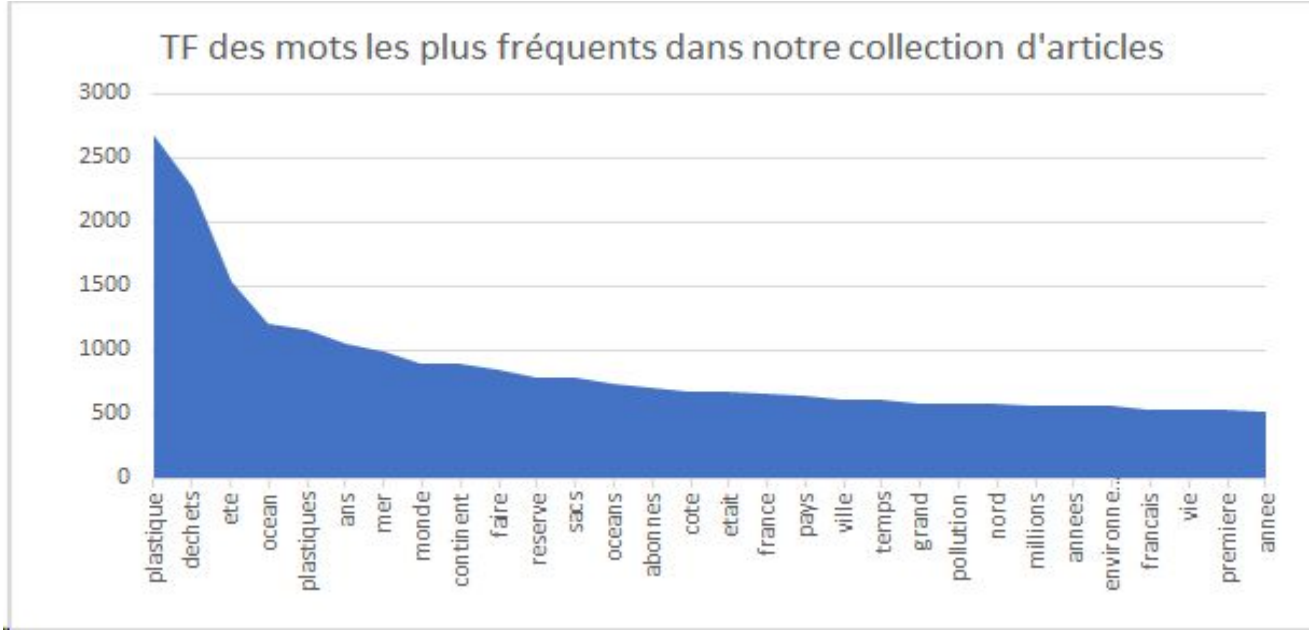
Cette ultime étape regroupe l'ensemble des résultats du projet Tableau de Bord. C'est une étape majeure car elle résume l'ensemble des travaux qui ont été fait dans ce projet. A l'aide du fichier Excel nous avons produit différentes visualisations que nous allons analyser afin de répondre notre problématique, les voici ci-dessous :

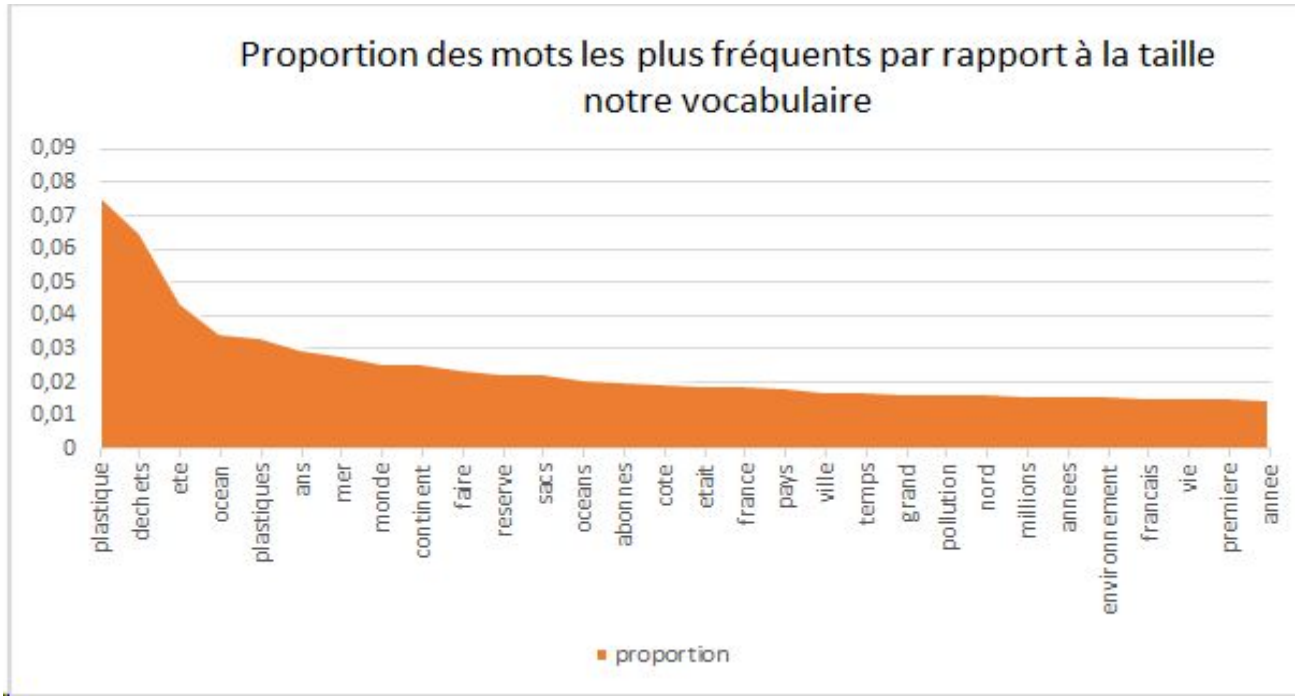




Ces 3 graphiques sont issus des requêtes Q11, Q12 et Q13. Ils nous permettent de nous rendre compte de la pertinence de nos articles tout au long de notre analyse. En s'appuyant simplement sur la description et le titre de chaque article, on identifie un vocabulaire commun lorsque l'on regarde les mots les plus employés. Toujours en haut de la liste le "plastique" avec les "déchets", ainsi que les "sacs" et "océan" qui sont tous des mots en lien direct avec notre sujet. A noter la présence du mot "étude" pour les mots les plus fréquents dans les titres qui relèvent l'aspect scientifique de certains articles.

Dans la requête Q1, nous avons produit deux graphiques, ils sont étroitement liés car ils permettent de mettre en évidence les mots qui sont les plus employés au sein de notre collection d'articles.

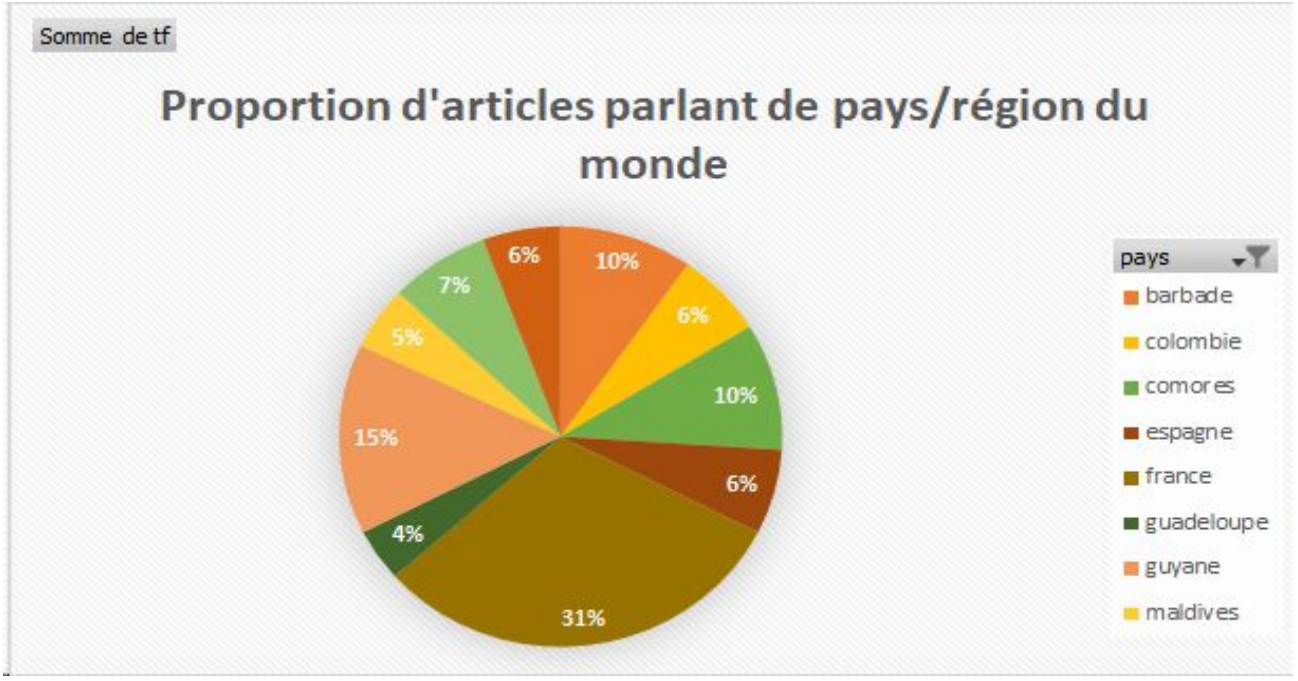




Ainsi, on retrouve les mots tels que : plastique, déchets, océan, mer et continent. Ce n'est pas étonnant car ce sont de mots que l'on qualifie de "mots clés" pour notre sujet. Il est donc normal que ce soient les mots qui apparaissent le plus dans notre corpus d'articles. De plus, on remarque aussi que la proportion de ces mots sur le vocabulaire total de notre collection, s'évalue à environ 20%. C'est à dire qu'en moyenne 20% de nos articles sont composés de ces mots, ou en d'autres termes, 20% du vocabulaire de chaque article comprend en moyenne ces mots. Ce qui témoigne de la pertinence de nos articles ; en majorité, ils traitent bien du sujet choisi et pour lequel nous avons scrappé l'ensemble de ces articles.

A partir de ces deux graphiques, on peut également constater que les mots : monde, pays, france, ville, français et nord, ont un poids important dans nos articles, ce sont donc des mots très significatifs. Ils nous permettent ainsi de comprendre que les questions et débats liés à la pollution plastique concernent un grand nombre de pays et de zones géographiques. Comme nous avons sélectionné des articles en français, il n'est pas anormal de voir des mots tels que france et français ressortir souvent.

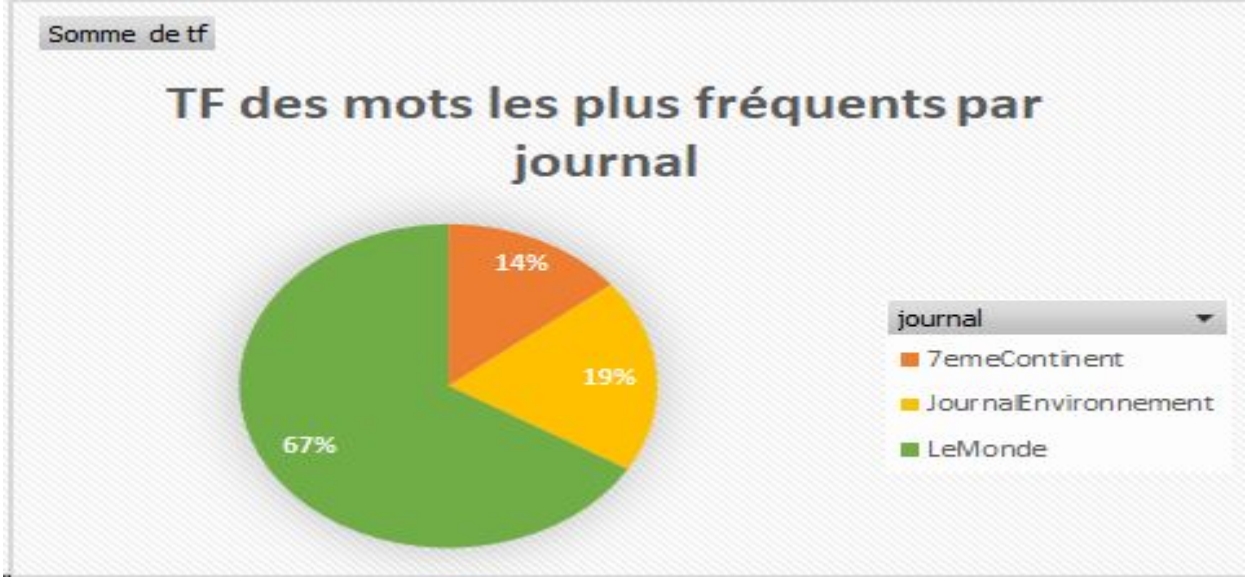
Puis en se basant sur les résultats de la requête Q2, on a pu dégager la proportion de chaque pays ou région du monde en fonction des articles qui contiennent des pays ou des régions du monde dans leur vocabulaire.



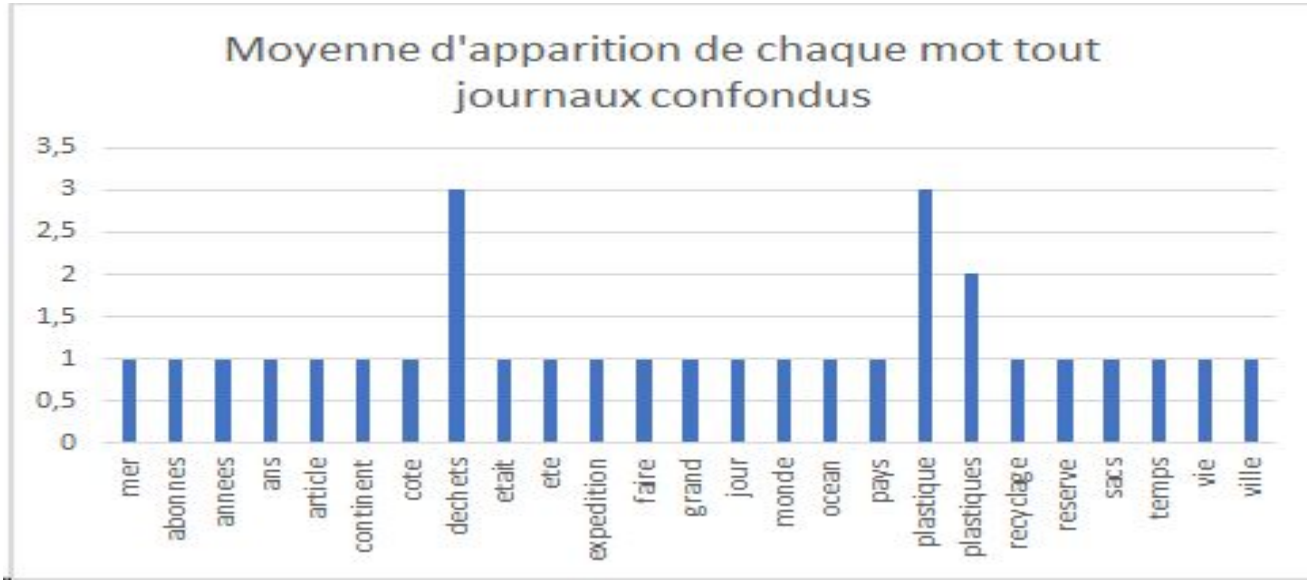
L'objectif étant ici de mettre en évidence le fait que la pollution maritime est un sujet mondial. On peut également déduire de ce graphique quelles sont les principales victimes de ce phénomène. On peut constater grâce à ce graphique que de nombreuses îles sont présentes dans nos articles, ce qui n'est pas étonnant. En effet, la pollution

maritime aurait pour première victime la faune et la flore, mais il ne faut pas oublier les victimes secondaires de ce phénomène. Ces victimes secondaires sont les habitants sur des îles et c'est ce qu'on peut clairement identifier ici.

Ensuite la requête Q3, nous a permis de rassembler par journaux le nombre de mots fréquents afin de comparer les proportions.

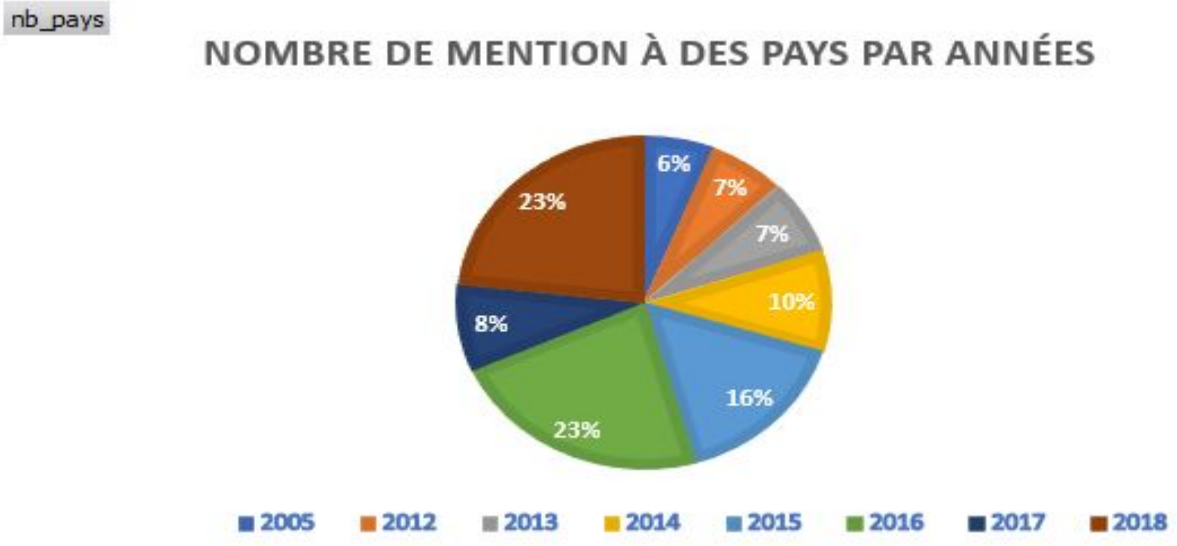


On constate donc qu'en moyenne les articles du journal Le Monde sont bien plus pertinent que les articles des deux autres sources journalistiques que sont : le 7ème continent et le Journal de l'Environnement. Cela s'explique en majeure partie car les articles du site Le Monde sont bien plus long que ceux des deux autres sites. De plus, c'est du site lemonde.fr que la majorité de nos articles proviennent.



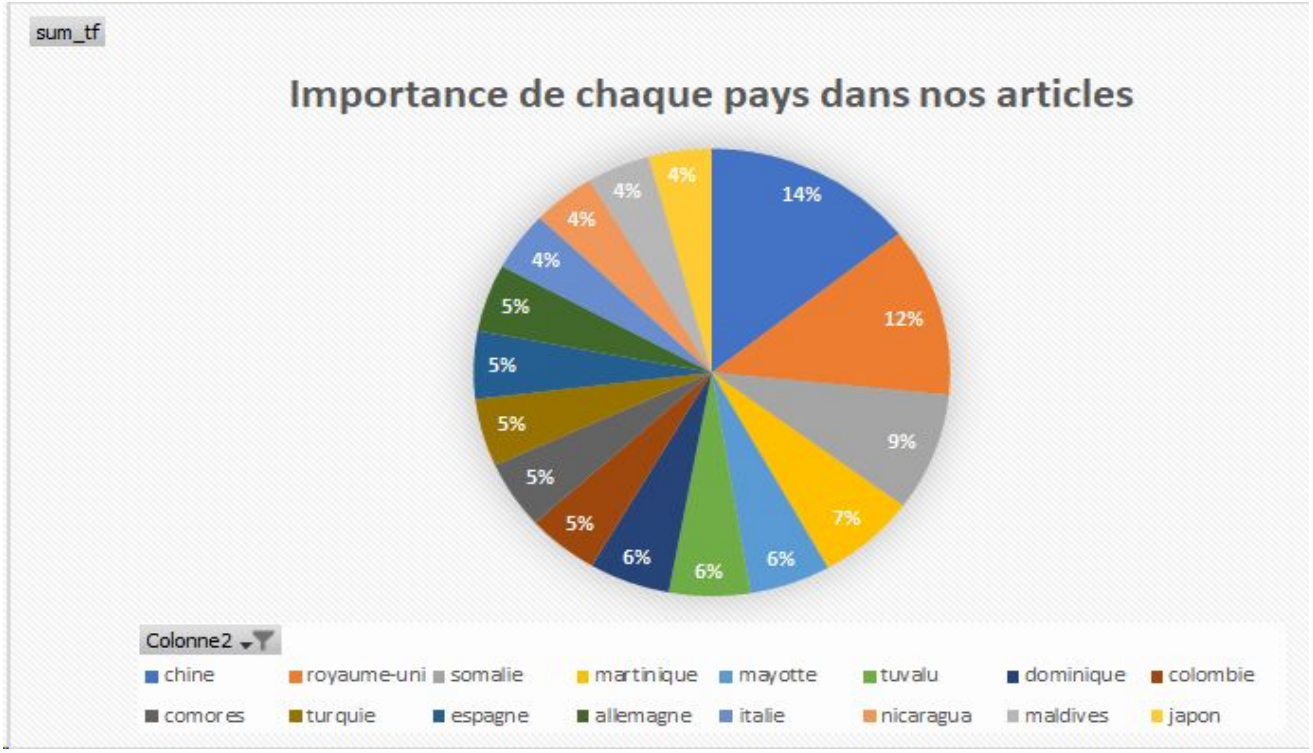
Ici on peut voir que les mots jugés très pertinents par la requête Q1 ressortent du lot. Avec des moyennes d'apparition d'environ deux fois à 3 fois par articles quelque soit la source, les mots "plastique" et "déchets" se font remarquer. On peut aussi remarquer que des mots comme "expéditions" et"recyclage" qui font parti du vecteur concernant les solutions du problème de la pollution en milieu maritime apparaissent en moyenne 1 fois par articles, toutes sources confondues.

Nous avons récupéré ensuite les résultats de la requête Q4 pour obtenir ceci :



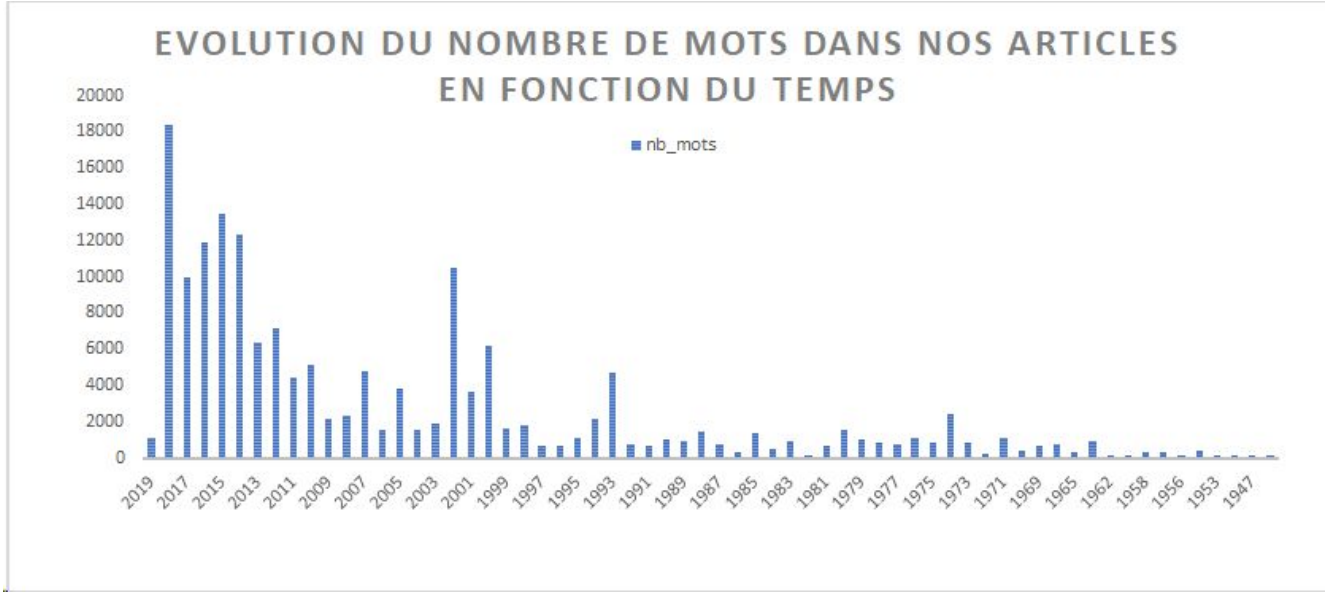
Grâce à ce graphique on peut voir une évolution des questions tournant autour de la pollution maritime au fil du temps. En effet, on peut remarquer que la pollution maritime au fil du temps ne cesse de faire intervenir des acteurs internationaux que sont

les pays. Ces augmentations se corrèlent parfaitement avec des évènements mondiaux sur le climat tels que la COP 24 en 2018 ou encore les accords de Paris (COP21).



Sur ce graphique nous avons fait le choix d'enlever la france car elle avait un poids beaucoup trop important (dû au fait que nos articles soit en français) nuisant à la bonne lecture du graphique. Nous pouvons ainsi voir que la Chine, le Royaume Uni et de nombreuses autres îles ont une certaine importances dans nos articles. Il n'est pas étonnant de retrouver la Chine car elle serait la principale source de production de plastiques aux mondes. De plus, nous retrouvons également les îles car elles doivent au quotidien faire face aux problèmes liés à la pollution plastique maritime.

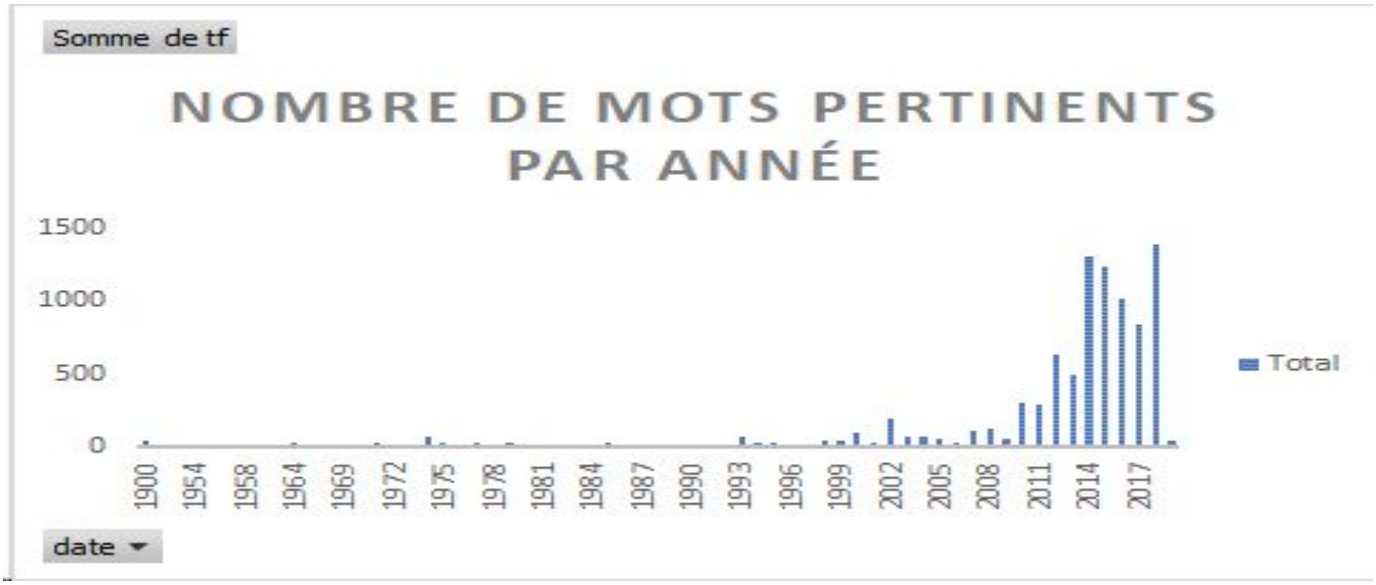
La requête Q5 quant à elle, nous permet de nous rendre compte, qu'à certains moments ou encore pendant certaines périodes, l'actualité traitaient du sujet de la pollution plastique en milieu maritime.



On observe ainsi des pics d'audience de ce sujet en 1992 (risques évoqués par la CEI), en 2000, en 2002 (pétrolier Le Prestige), en 2016 (accord de Paris) ou encore dernièrement en 2018 (COP 24). Et si nous avions tracé une courbe de tendance, cette dernière aurait été à la hausse. Donc on peut voir qu'au fil du temps, la pollution plastique des fonds marins prend une place de plus en plus importante dans la société.

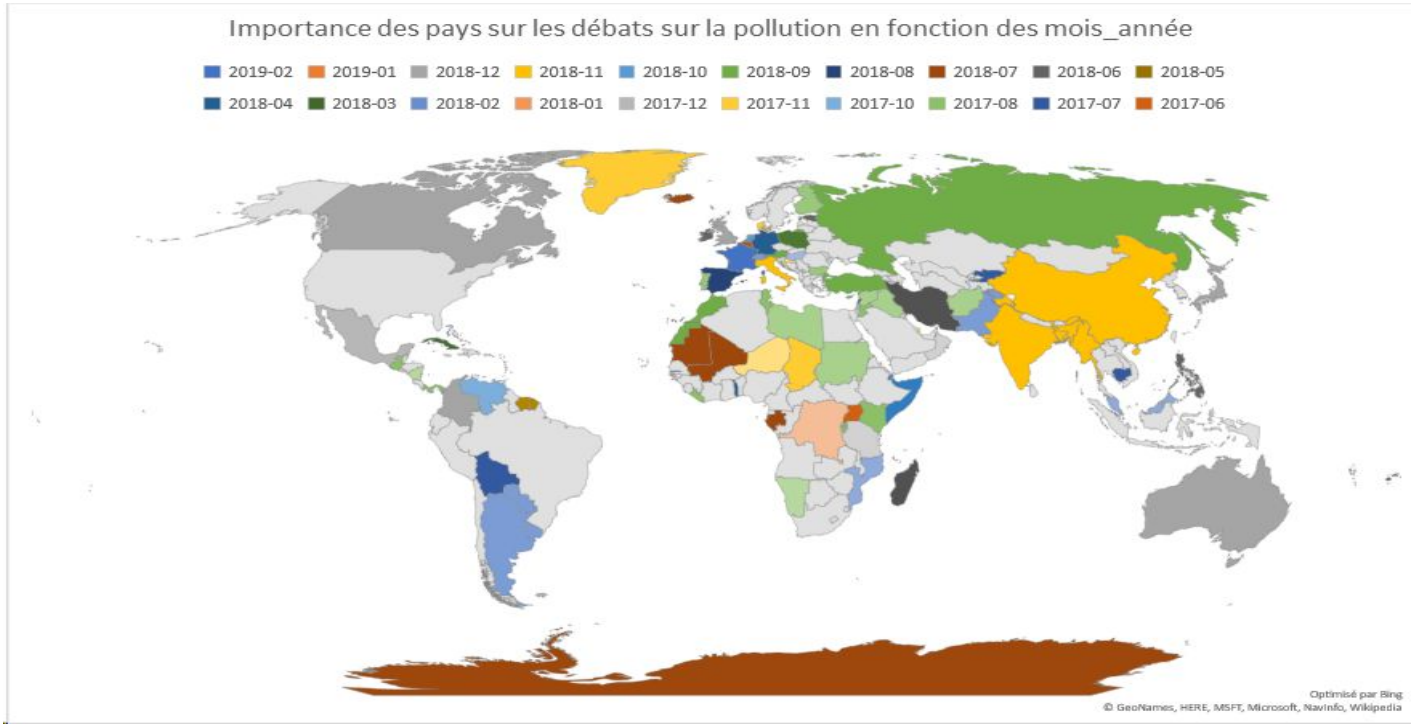
La requête Q6 consistait à rédiger une liste mots pertinents et nous les avons ensuite compté par années pour obtenir une évolution au fil du temps.

:



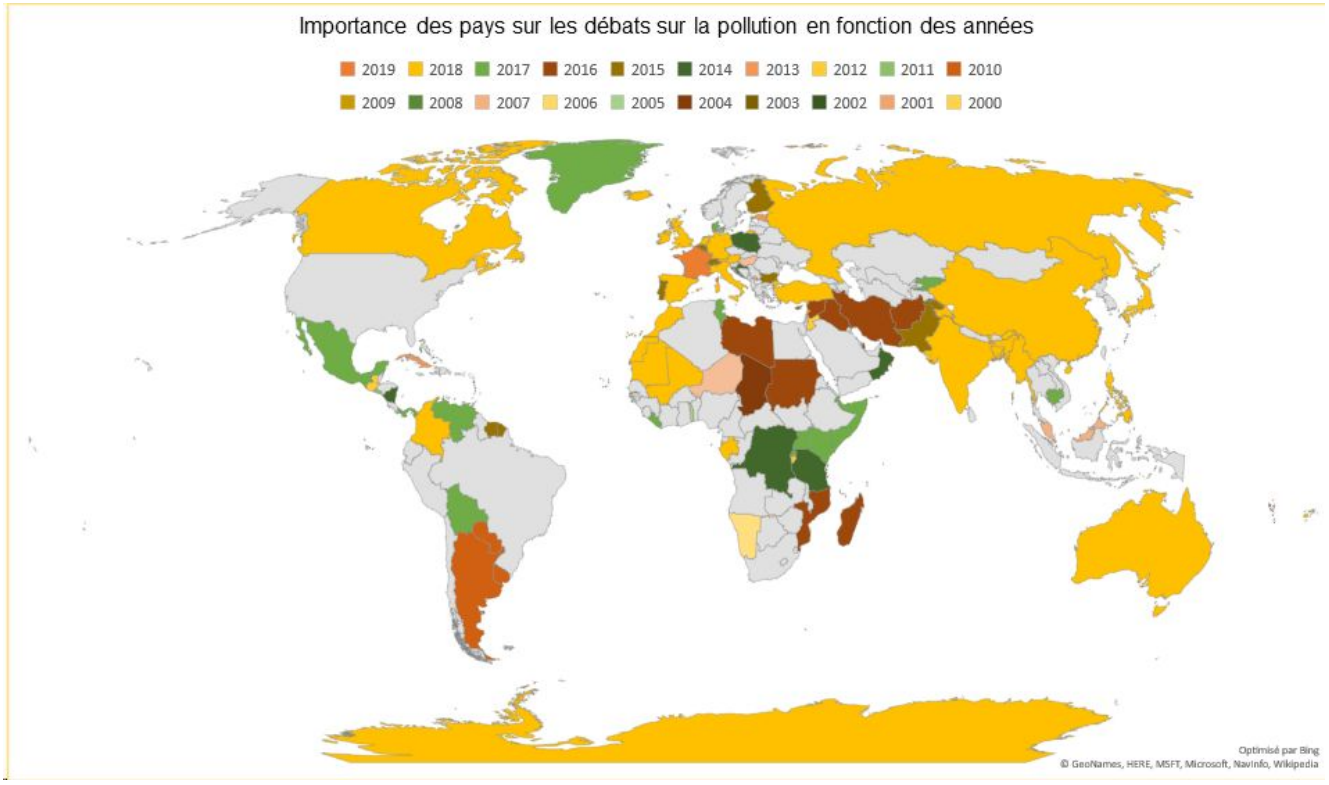
Ce graphique rejoint un peu l'analyse faite lors de la requête Q5, soit qu'au fil de temps on parle de plus en plus du sujet de la pollution maritime. Mais elle se base elle sur un nombre de mots présélectionnés.

A partir de la requête Q7, nous voulions analyser sur une période (2017 - 2019) les pays qui étaient directement ou indirectement liés à la pollution maritime.



Comme on peut le voir, en analysant la presse uniquement française, et sur une période restreinte, la majorité des continents sont touchés, révélant ainsi que la pollution plastique de nos océans et mers, est bel et bien un problème mondial et non uniquement localisé, il faut donc voir plus loin que le 7ème continent.

Sur le même principe que la requête Q7, la requête Q8 se focalise seulement sur les années et sur une période plus vaste que la requête Q7 (de 1900 à nos jours)



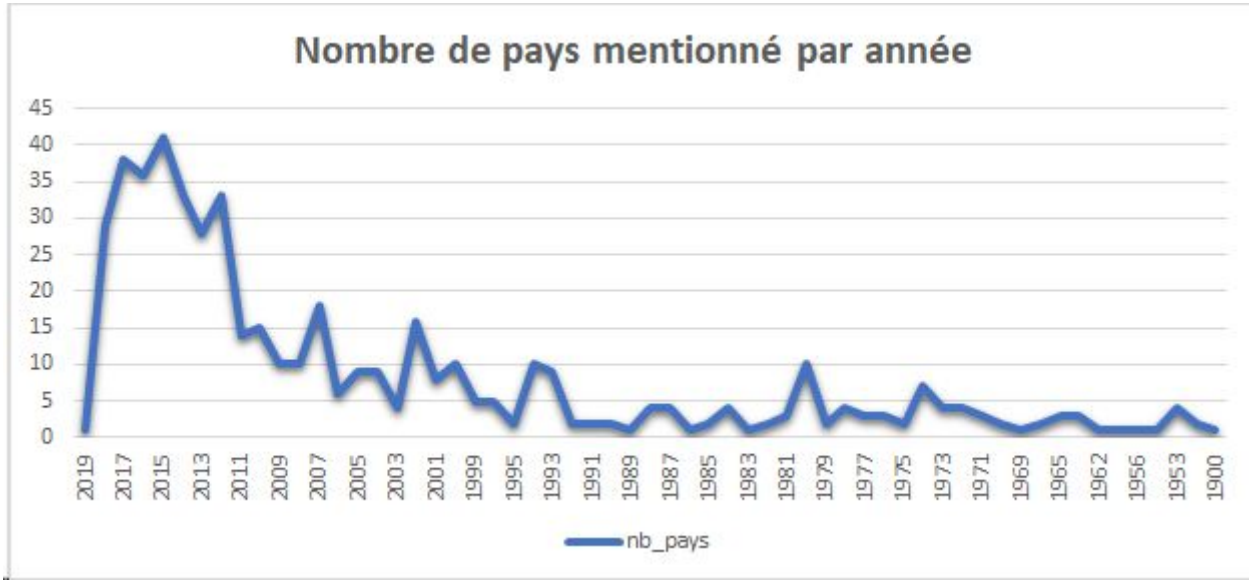
On peut voir que très peu de pays sont insensibles au problème mondial qu'est la pollution maritime. On peut ainsi observer en analysant la presse française, que la France, ainsi que d'autres pays européens (le Royaume Uni, l'Espagne, l'Italie et l'Allemagne), mais également des pays d'autres continents (le Canada, la Russie, la Chine, l'Inde et l'Australie pour les plus importants), sont au coeur du débat depuis 2018 et veulent proposer des solutions afin résoudre ce casse-tête à grande échelle.

Cependant il reste à noter qu'avec cette simple analyse certaines puissances restent insensibles aux problèmes environnementaux (les Etats Unies ou le Brésil). Deux explications sont possibles à ce manque d'information :

- nos sources journalistes ne parlent jamais ou trop peu de ses pays en lien avec la pollution plastique en milieu maritime
- ces pays ne s'intéressent pas à la pollution plastique

En se renseignant un peu, on a remarqué que les pays grisé ne proposaient pas de réelles solutions pour palier aux problèmes de la pollution plastique, ou alors qu'ils n'étaient pas directement impliqués par les conséquences.

Enfin, il reste à noter que des zones de faible peuplement sont souvent mentionnée dans nos articles car ce sont des zones qui en majorité sont sensible aux changements climatiques. (Groenland, Antarctique et archipels)



Avec ce graphique on peut voir que de plus en plus de pays s'intéressent au sujet de la pollution plastique en milieu maritime. Environ 40 pays en parlent de manière quotidienne de nos jours, et seulement en analysant la presse française. Donc il y a beaucoup d'acteurs concernés aujourd'hui.

Pour conclure cette phase de diffusion des données au travers de graphiques, on peut dire la pollution plastique en milieu maritime est devenu un problème mondial majeur qui intéresse de plus en plus de personnes au fil du temps afin de proposer des solutions.

VI) Gestion de configuration

Ce projet nécessitait un logiciel de versionning assez complet pour éviter de perdre du code, pouvoir proposer une architecture commune à tous les collaborateurs, et assurer une sécurité ainsi qu'une pérennité du code.

C'est pourquoi nous avons utilisé un logiciel GIT (avec github). Nous l'avions déjà manipulé durant les divers projets universitaires auxquels nous avons pu participer. C'est un logiciel simple et très complet qui répond parfaitement à nos besoins. Les seuls prérequis étaient que tous les collaborateurs possèdent un compte github et GIT sur leur machine.

On a mis en place sur le github une architecture commune à tous les membres du groupe projet, comme on peut la retrouver ci-dessous :

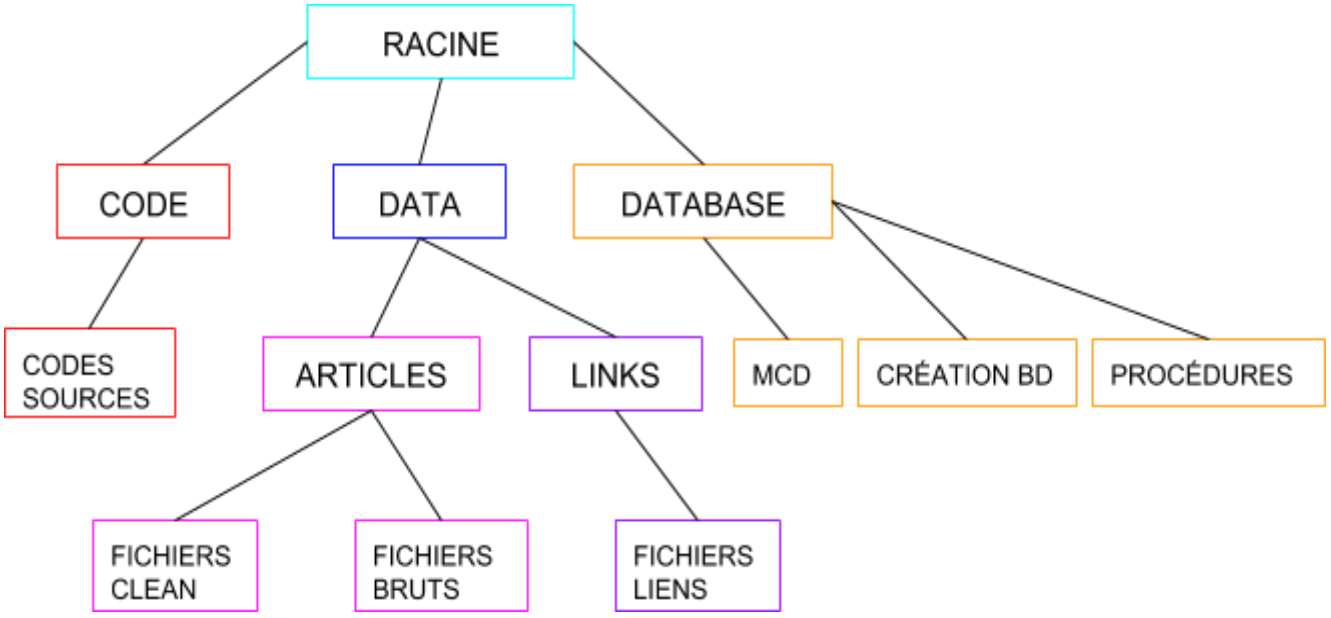


Figure 12 : Architecture du projet

Cela nous a autorisé à envoyer du code (cf. push), à le récupérer (cf. pull) sans conflits (cf. conflict) apparents. En effet les membres de l'équipe projet ne faisaient pas les modifications sur les mêmes fichiers, cela éliminait donc les risque de conflit. Chacun jouant un rôle différent dans ce projet, il était indispensable de se mettre d'accord sur toutes modifications avant de push des fichiers sur le git afin de garantir une version stable sur ce dernier.

VII) Assurance et contrôle qualité

Nous avons réalisé plusieurs réunions au cours du projet pour évaluer l'avancement et valider les différentes étapes de ce dernier. Certaines de ces réunions ont eu lieu avec les clients.

21 Janvier 2019

Ordre du jour

Définir les bases de nos recherches dans le but de déterminer le sujet de notre projet.

Commencer à faire quelques recherches sur les différents sujets auxquels nous pensions, pour voir si les sujets disposent de ressources nécessaires

Réalisations

Pour ce projet, nous avons cherché des sujets d'actualité qui concernent le plus grand nombre. Nous avions plusieurs idées en tête telles que : La coupe du monde de football 2018, le Burn Out, l'écologie, l'astronomie, la conquête de mars ou encore la pollution.

Parmis ces idées, seules deux ont été retenues : La conquête de Mars et la pollution maritime

Nous avons finalement décidé de sélectionner le sujet concernant la pollution maritime, ce sujet correspondait à tous les membres du groupe ainsi qu'aux enseignants.

Nous avons trouvé quelques sites de journaux susceptibles de nous intéresser.

A faire

Il nous restait à chercher des sources intéressantes concernant notre sujet

28 Janvier 2019

Ordre du jour

Vérifier que les articles sélectionnés sont pertinents et suffisants pour réaliser une analyse complètes.

Réalisations

Nous avons parcouru différentes sources internet (toutes les sources trouvées se trouvent en annexe) afin de trouver le maximum d'articles traitant de notre

problématique. Nous avons sélectionné les médias numériques suivant : Le monde, le journal de l'environnement, septiemecontinent.

A faire

Une fois les articles trouvés et validés nous devons créer un code pour récupérer le contenu de ces articles dans un fichier.

08 Février 2019

Ordre du jour

La récupération des articles est finie, nous devons maintenant vérifier que le code respecte la charte de codage (cf annexe 2).

Réalisations

Nous avons créé des fonctions avec le langage de programmation python propre à chaque média numérique afin de récupérer automatiquement les articles qui traitent notre problématique.

A faire

Nous devons maintenant créer une base de données et peupler la base avec les articles.

18 Février 2019

Ordre du jour

Valider le MCD de la base de données par les enseignants.

Réalisations

Nous avons fait un MCD qui nous permet de donner une représentation conceptuelle de toute la base de données relationnelle.

A faire

Nous devons maintenant réaliser le script de création de la base de données (langage SQL).

25 Février 2019

Ordre du jour

Valider la conformité du script de la création de la base de données avec la charte qualité (cf annexe 2).

Réalisations

Nous avons réalisé le code pour créer la base de données

A faire

Nous devons maintenant créer le script qui permettra d'alimenter la base données

04 Mars 2019

Ordre du jour

Création du script de la base de données
Mise en place de celle-ci en local

Réalisations

Nous avons réalisé le code pour alimenter la base de données.

A faire

Nous devons maintenant réaliser des requêtes et réaliser des analyses

15 Mars 2019

Ordre du jour

Faire un point sur l'avancement du rapport.
Valider le plan et vérifier la cohérence des parties.

Réalisations

Nous avons réalisé le plan du rapport.

A faire

Apporter des modifications au rapport et créer le diaporama pour la présentation oral de notre projet.

18 Mars 2019

Ordre du jour

Valider les requêtes
Déterminer des analyses statistiques pertinentes.

Réalisations

Sélectionner les analyses pertinentes.

A faire

Mettre en forme les résultats dans le rapport de projet.

22 Mars 2019

Ordre du jour

Valider le diaporama
Se préparer à la présentation oral.

Réalisations

Nous avons apporté des modifications au diaporama et nous nous sommes attribués des parties pour l'oral

A faire

Réaliser la soutenance.

VIII) Bilans de projet

A) Bilans personnels

1 - William

Lors de ce projet j'ai pu appliquer l'ensemble des connaissances vues lors de cette année universitaire et lors des précédentes. De plus, j'ai participé au projet de sa conception à sa réalisation. Ainsi, j'ai pu avoir une vision globale de ce projet.

Durant le projet, nous avions tous un rôle défini. En étant gestionnaire de configuration, j'avais des responsabilités concernant la sauvegarde du code source, la conservation des versions stables, et pouvoir revenir à ces dernières sans trop de difficultés. Le logiciel GIT Kraken m'a bien aidé pour réaliser cette tâche.

Ensuite, il y avait un réel challenge au niveau des deadlines que nous nous étions fixées, mais nous avons réussi à les respecter sans trop de problèmes. J'ai consacré beaucoup d'heures de travail pour que l'on puisse les respecter (hors créneaux prévus à cet effet).

Cependant il reste quelques points à améliorer. En effet, la mise en place de la base de données avec le SGBD SQL Server (licence Microsoft) fût une source de retard dans notre projet car nous ne disposions pas tous des possibilités pour pouvoir l'exploiter pleinement (ex : SQL Server n'est pas disponible sous OS Linux, accès à SQL Server seulement à l'Université ou en local sur une machine sous OS Windows).

Pour finir, les aspects organisationnels de ce projet (ex : la mise en place d'un planning prévisionnel) ont été des éléments clés à la réussite de ce projet.

2 - Cyril

Ce projet a été d'un point de vue personnel très enrichissant, car il m'a permis d'apprendre à gérer un projet de sa conception à sa phase de restitution. Nous avons dû travailler en équipe pour faire face à des contraintes au niveau des délais ou encore au niveau de l'utilisation de certains logiciels ce qui est une application très concrète du monde du travail. De plus, le fait que chaque membre du groupe "joue" un rôle défini nous plonge au coeur d'un réel projet, tel qu'on aura l'occasion d'en réaliser dans nos futurs stages.

Je pense cependant que certains points de ce projet sont à améliorer pour les années suivantes. Tout d'abord, le nombre d'heures dédiées à ce projet me semble insuffisant, j'ai pour ma part dû beaucoup travailler en dehors de ces heures afin de terminer ce projet. D'autre part, je trouve que le fait d'imposer le SGBD SQL Server est une erreur, en effet beaucoup d'étudiants ne possèdent pas Microsoft sur leurs machines ce qui est donc pénalisant pour l'ensemble des groupes.

3 - Enzo

D'un point de vue personnel le projet m'a permis de me mettre en situation dans un réel projet au sein d'une équipe de travail. J'ai pu voir tous les aspects de gestion de projet, ainsi que les aspects technique. Le fait de travailler en équipe nous oblige à nous organiser afin de nous répartir les tâches et de respecter les délais.

Concernant les points d'améliorations, l'utilisation du SGBD Sql Server peut défavoriser certains groupes puisqu'il faut des ordinateurs avec Microsoft ou alors le faire à l'université, ce qui n'est pas toujours possible puisque les salles ne sont pas forcément disponibles en dehors des créneaux réservés.

4 - Fanny

D'un point de vue personnel, ce projet a été très enrichissant. Le fait de réaliser un projet dans son intégralité nous permet de voir tous les aspects de la conception du projet jusqu'à sa finalisation. Nous avons travaillé en équipe où chacun avait un rôle prédéfinie, ce qui nous a permis de s'organiser et de se répartir le travail en fonction de notre spécialité. Le projet étant réalisé dans un concept client/fournisseur, c'est un bon exemple de projet que l'on traitera dans la vie professionnel.

J'ai été confronté à des difficultés pour la base de données, nous devions réaliser celle-ci sur le SGBD SQL Server, présent uniquement sur Microsoft, ne possédant pas ce système d'exploitation, je n'ai pas pu travailler sur celle-ci depuis mon ordinateur personnel.

B) Bilan global

Au cours de ce projet, nous avons été confrontés à un certain nombre de difficultés.
La première d'entre elles concernait le scrapping des données, en effet nous avons choisi trois sites de média numérique différents, et nous avons donc dû analyser les sites un par un afin d'adapter nos méthodes et fonctions de scrapping à chacun des sites.
Nous avons également eu des difficultés pour la base de données, en effet, le SGBD à utiliser était imposé, il s'agissait de SQL Server qui est un logiciel développé par Microsoft. Cependant, deux membres du groupes ne possédaient pas WIndows sur leur machine ; ils ne pouvaient donc pas posséder la base de données sur leur ordinateur. Nous avons pallié à ce problème en créant la base de données sur la machine de Cyril qui pouvait exécuter un script python pour générer le résultat des requêtes et que les autres membres du groupe puissent faire des analyses.
La partie nettoyage des données a également été difficile, comme nous avons choisi de récupérer des données venant de média français, nous avons dû utiliser des librairies python très peu complètes, en effet , les librairies sont beaucoup plus adapté pour traiter les textes en anglais. Notre nettoyage de données n'est pas le plus optimal, mais reste suffisant pour faire nos analyses.

Insérer en base de données nous a pris beaucoup de temps alors que nous avions relativement peu de données. La création de procédures et l'insertion des données a impliqué un traitement assez lourd alors que nos fichiers json étaient bien construit et auraient donc pu être interrogé directement via python sans passer par le SGBD SQL Server.

Finalement, ce projet nous aura permis de voir l'importance de travailler en équipe pour mener à bien un projet. En effet, nous avons dû nous diviser le travail, nous écouter les uns les autres afin de sélectionner les meilleurs idées de chacun et ainsi faire les choix qui nous semblaient les plus judicieux. De plus, nous avons pu appréhender le fait d'avoir des deadlines et d'autres contraintes, ce qui nous a obligé à surmonter un certain nombre de difficultés, mais également à nous adapter aux exigences et changements. Nous ressortons donc plus fort et plus apte pour affronter les futurs challenges de nos vies professionnelles.

Annexe

Annexe 1 : Travail préparatoire

Recherche d'article pertinents sur la pollution maritime

https://www.ifop.com/publication/les-francais-et-le-septieme-continent/
https://www.planetoscope.com/eau-oceans/1889-pib---production-economique-marine.html
https://information.tv5monde.com/info/le-7eme-continent-un-monstre-de-plastique-1863
https://www.lsa-conso.fr/
https://www.europe1.fr/sciences
http://maplanete.blogs.sudouest.fr/archive/
https://www.pleinevie.fr/vie-quotidienne/environnement
https://www.gurumed.org/
https://www.notre-planete.info/actualites/
https://www.encyclopedie-environnement.org/eau/
https://www.notre-planete.info/actualites/
https://www.wwf.fr/champs-daction/ocean
https://www.lemonde.fr/pollution/
https://www.greenpeace.fr/protection-des-oceans/
https://www.encyclopedie-environnement.org/rubrique/eau/
https://www.planetoscope.com/environnement/eau-oceans
https://www.futura-sciences.com/planete/
http://www.septiemecontinent.com/
https://www.pleinevie.fr/vie-quotidienne/environnement
https://www.liberation.fr/planete/
http://www.regardsurlemonde.fr/blog/tag/plastique
https://www.nationalgeographic.fr/environnement/
https://www.lesechos.fr/05/06/2018/lesechos.fr/
https://www.data.gouv.fr/en/datasets/tonnages-dechets-menagers-et-assimiles-valorises1/

Annexe 2 : Charte de codage

Ce document décrit les règles à suivre pour coder le plus lisiblement possible. Pour que n'importe quel utilisateur puisse comprendre, modifier, poursuivre la construction du code facilement.

A) Convention logiciel

1 - Langage Python

Tous les acteurs du projet utilise la même version de python pour éviter les problèmes de compatibilité. Nous utilisons l'application Jupyter NoteBook pour réaliser le code.

2 - Langage SQL

Nous manipulerons la base de données avec SQL Server.

B) Convention de nommage des variables, fonctions

Tous les noms de variables et noms de fonctions doivent être

- En une seule langue : Anglais
- En minuscule, si les noms contiennent plusieurs mots, utiliser les underscores « _ ». Exemple : my_var

Ce qu'il faut éviter :

- Les variables trop longues (3, 4 parties maximum). Exemple: long_variable_like_this_is_wrong.
- Les variables qui sont peu compréhensibles
- Les variables quasi-semblables. Exemple: my_var et my_var2.

C) Commentaires

Les commentaires permettent d'expliquer au lecteur de votre code le fonctionnement de votre programme. Ils sont d'une aide précieuse pour relire le code après quelque temps ou si si quelqu'un essaye de comprendre le code.

Veillez à éviter les commentaires inutiles.

Toutes vos fonctions doivent être commentées, de façon à indiquer ce que prend la fonction en entrée et ce qu'elle retourne, suivie d'un petit résumé de ce qu'elle effectue.

Exemple : def function f(x):
"""
input : explain what the parameters are. ex.: x: json data.
ouput : explain what the function returns. This is what the function is doing.
"""

- Les commentaires ne doivent pas dépasser 30% de la longueur du code.
- Indiquez dans les commentaires ce que fait le code, et non comment il le fait (le code
  doit être suffisamment clair pour que le « comment » se lis tout seul).

D) Lisibilité du code

- Les variables doivent être séparées de chaque opérateurs par un espace
  Exemple : var = var1 + var2
- Le code doit être encodé en utf-8.
- Le code doit être indenté pour avoir une meilleur lisibilité
- Sur une ligne, 80 caractères semblent être la limite actuelle pour ne pas utiliser la barre de défilement afin de voir la fin du code. Dans le cas d'une ligne qui dépasse 80 caractères, découpez-la après les opérateurs.
- Le code doit être le plus simple possible pour être facilement compréhensible par tous.
- Préférez les méthodes simples et les fonctions qui prennent un petit nombre de paramètres. Évitez les méthodes et les fonctions qui sont longues et complexes.
- Évitez de mettre beaucoup d'idées sur une seule ligne de code.
- Toutes les variables définies ainsi que les packages importés doivent être utilisés.

1 - Langage Python

- Sautez une ligne après chaque méthode.
- Une fonction doit toujours retourner quelque chose.

- Un seul « return », il ne doit pas se situer pas dans un « if » dont la condition n'est pas toujours vérifiée.
- Les imports de plusieurs modules doivent être écrit sur plusieurs lignes.
- Pour importer vos modules, Préférez des chemins relatifs. Exemple : « import sys » et non « from sys import * ».
- Ne pas mettre d'espace avant les deux points et les virgules, mais après.
- Ne pas mettre d'espace à l'intérieur des parenthèses, crochets ou accolades.

2 - Langage SQL

- Le nom des tables doivent être :

  - Toujours en minuscule
  - Toujours au singulier
  - Sans accents
  - Sans abréviations
  - Le nom des tables ne doit pas être un mot réservé de SQL.

- Le nom des colonnes doivent être:

  - Une clef primaire : id_ + nom de la table
  - Un libellé : lib_ + nom de la table.

- Les contraintes doivent respecter les règles de nommage:

  - Les contraintes des clés primaires doivent commencer par «pk_» suivi du nom de la table.
  - Les contraintes des clés étrangères doivent commencer par «fk_» suivi du nom de la table fille puis de la table mère.
  - Pour les autres contraintes, commencez par « ck_ ».