# Projet Tableau de Bord Master 1 SID

## *L'évolution du e-commerce à l'ère du digital*

**Fournisseurs : Célia SARTORI, Samba Diallo WADE, Elhadji Fallou FALL, Serigne DIAW**

**Clients : Wahiba BASHOUN, Mokadem RIAD**

# Suivi des modifications

| VERSION | OBJET DE LA MODIFICATION | DATE |
|---|---|---|
| 1.00 | Initialisation du document | 16/01/18 |
| 1.01 | Problématique et Documents applicables et de référence | 26/01/18 |
| 1.02 | Initialisation Processus de Développement | 29/01/18 |
| 2.00 | Valorisation des données | 02/02/18 |
| 2.01 | Finalisation Valorisation des données | 05/03/18 |
| 2.02 | Finalisation du Processus de Développement – Gestion de configuration – Assurance et Contrôle qualité | 19/03/18 |
| 2.03 | Finalisation du rapport | 23/03/18 |
| 3.00 | Version finale | 28/03/18 |

# Sommaire

# Table des illustrations

# 1 Objet du document

L'objet de ce document est d'expliquer la démarche de développement que nous avons suivie et les résultats que nous avons produits dans le cadre de la mise en place du Projet Tableau de Bord au cours de l'année de Master 1 SID de notre cursus. Il est destiné au client (les enseignants) en réponse au Cahier des Charges qu'il a établi.

# 2 Problématique

Dans le cadre de ce projet Tableau de Bord, nous devons développer un système d'aide à la décision pour le client.

Pour ce projet, nous nous sommes intéressés au e-commerce.

Le e-commerce ou commerce électronique regroupe l'ensemble des transactions commerciales s'opérant à distance par le biais d'interfaces électroniques et digitales.

Le e-commerce englobe essentiellement les achats en ligne s'effectuant sur Internet à partir des différents types de terminaux (ordinateurs, tablettes, smartphones, consoles, TV connectées).

Ces dernières années, le commerce en ligne a pris de plus en plus de place dans l'économie et a radicalement bouleversé nos façons de consommer.

Nous avons ainsi décidé pour ce projet d'étudier l'évolution du e-commerce à travers des articles d'actualités.

Pour cela, nous avons effectué des analyses et produit une visualisation de données textuelles extraites à partir d'un site Web.

# 3 Documents applicables et de référence

## 3.1 Documents applicables

Les documents applicables définissent les documents qui seront créés pendant la vie du projet.

[CCO] Charte de codage

[CG] Charte graphique

[RR] Règles à suivre pour le rapport

[CR] Cahier de recette

## 3.2 Documents de référence

Les documents de référence sont les documents qui ne sont pas créés durant le projet, mais sur lesquels les documents applicables s'appuient.

[CDC] Cahier des charges
W. *Bahsoun, R. Mokadem*

[GL] Cours « Génie Logiciel »
*W. Bahsoun*

[GPRO1] Cours et TD « Gestion de Projet »
*W. Bahsoun*

[BDD2] Cours et TPS « SQL Server »
*R. Mokadem*

[DAWA] Cours « Data Warehouse »
*G. Hubert*

## 3.3 Terminologie

*Web Crawling :* Parcourir de manière programmée une collection de pages Web et extraire des données.

*Text mining* : La fouille de textes ou « l'extraction de connaissances » dans les textes est une spécialisation de la fouille de données et fait partie du domaine de l'intelligence artificielle.

# 4 Organisation de travail

## 4.1 Répartition des rôles

Nous avons réparti les fonctions assurées par les différents membres de l'équipe de cette façon-là :

| Fonctions assurées | Membres de l'équipe |
|---|---|
| Chef de projet | Serigne DIAW |
| Responsable de Gestion de Configuration | Samba Diallo WADE |
| Responsable Assurance et Contrôle Qualité | Célia SARTORI |
| Développeur | Elhadji Fallou FALL |

Au cours du projet, tous les membres de l'équipe ont participé aux différentes étapes, allant de la récupération des données aux analyses statistiques et à la rédaction du rapport.

Selon la dimension du projet, une même personne a pu prendre différentes responsabilités.

Pour organiser le travail nous avons utilisé l'outil "Trello" qui consiste en un tableau KANBAN interactif dans lequel nous avons identifié grâce à des couleurs les tâches déjà réalisés et celles en cours.

Nous avons cherché au maximum à répartir le travail pour que tous les membres du groupe travaille à la fois sur le côté Programmation et le côté BD. C'est le service GitHub qui nous a permis de mener à bien cet objectif.

Au début du projet, chaque membre du groupe a travaillé sur la partie du projet correspondant à ses préférences puis nous avons tourné aux différents postes : administration BD, Programmation WEB et statistique et génération des données. En début de projet nous avons fait beaucoup de réunions pour être sûr de partager la même vision du projet, de récupérer les données mais aussi de concevoir la BD.

Nous avons chacun consacré en moyenne 5h00 par semaine en plus des 4h00 de TP encadrées au projet.

## 4.2 Organisation et répartition des tâches

Afin d'organiser notre travail, nous avons créé un modèle SADT décrivant le déroulement de notre travail et un planning prévisionnel Gant afin de visualiser l'évolution des tâches. Nous avons également utilisé certaines méthodes de travail telles que SCRUM.

- **Modèle SADT**

Nous avons créé un modèle SADT afin de pouvoir visualiser les différentes phases nécessaires à la réalisation du projet.

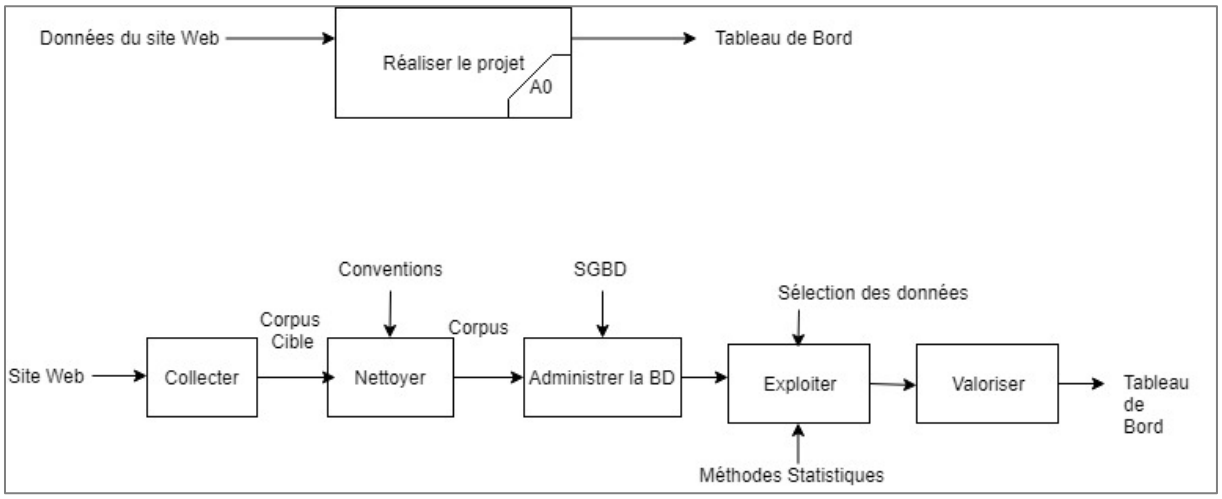


*Figure 1: Modèle SADT*

- **Planning prévisionnel**

Nous avons créé un diagramme de Gant afin de visualiser l'organisation des phases dans le temps:

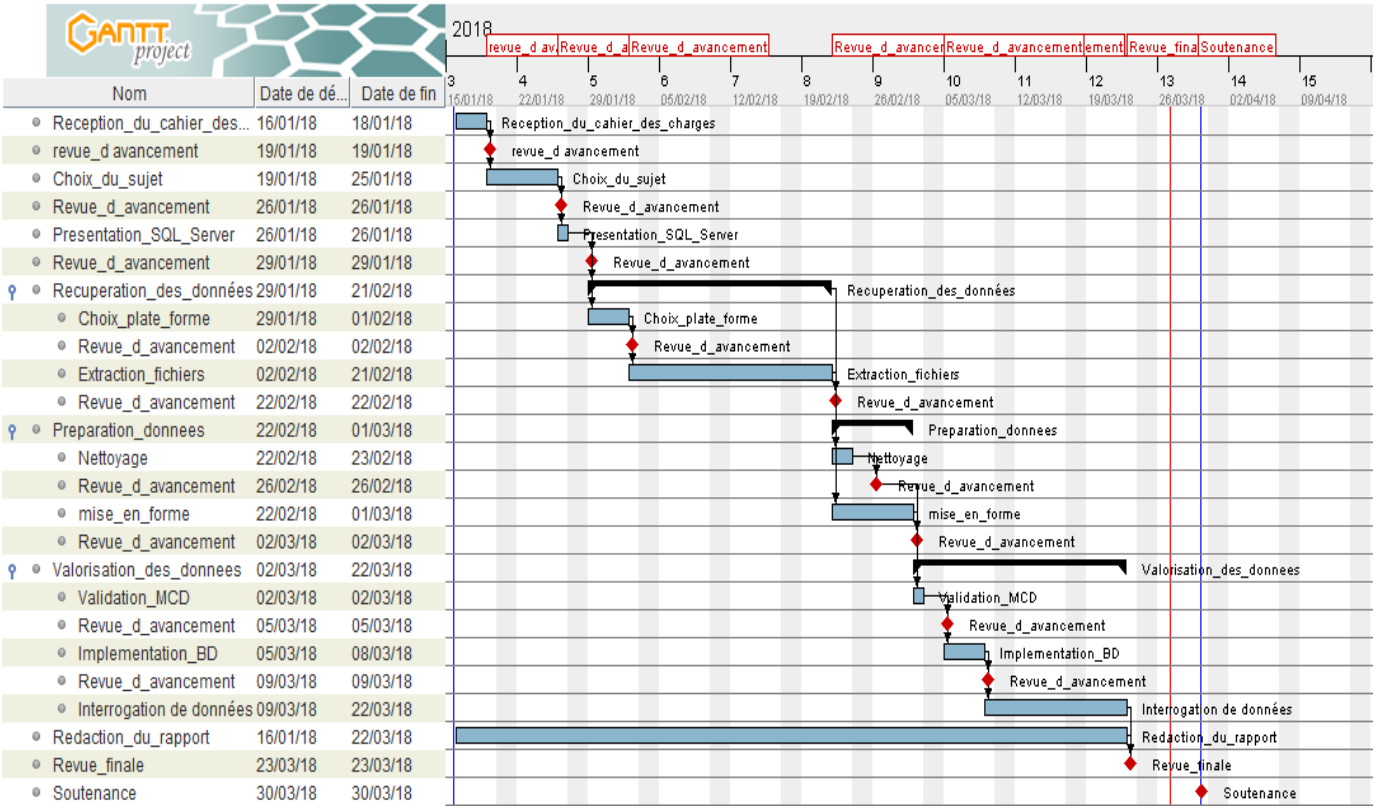


*Figure 2: Diagramme de Gant*

- **Méthode de travail**

Afin de suivre l'avancement de notre travail et de partager les documents de notre projet, nous avons utilisé des outils de gestion de projet en ligne : Github, Trello et Slack.

Pour le bon fonctionnement de notre projet nous avons choisi d'utiliser la méthode SCRUM. Ainsi une mêlée quotidienne était effectuée d'un ¼ d'heure pour voir l'état d'avancement de chaque tâche.

# 5 Processus de développement

## 5.1 Recherche d'informations et Collecte des données

Dans le cadre de ce projet, nous devons rechercher des informations sur le e-commerce et son évolution.

Afin de définir le périmètre du projet, nous avons choisis d'extraire les informations utiles à notre projet d'une unique source.

Nous avons ainsi décidé d'extraire les informations à partir du site du Journal Du Net (JDN) https://www.journaldunet.com/ .

Nous avons ensuite cherché la rubrique concernant le e-commerce accessible via le lien : https://www.journaldunet.com/ebusiness/commerce/list .

Nous avons définis notre équation de recherche de la forme suivante : journaldunet + ebusiness + commerce.

Cette rubrique relate des articles d'actualités concernant le e-commerce, depuis l'année 2006. Pour chaque année, les articles sont triés par mois.

**Web Crawling**

Nous avons utilisé Scrapy, qui est un Framework d'application open source permettant d'explorer des sites Web et d'extraire des données structurées qui peuvent être utilisées pour un large éventail d'applications utiles, telles que l'exploration de données, le traitement de l'information.

Ce qui nous a permis d'extraire nos articles concernant l'e-commerce à l'ère du digital. On a récupéré au total 2759 articles.

Pour chaque article on a récupéré les éléments nécessaires à notre analyse par la suite. C'est-à-dire par exemple : le titre, l'auteur, l'introduction, le contenu de l'article, ses tags et sa date.

Après avoir constitué notre corpus cible, nous avons procédé au nettoyage des textes.

## 5.2 Préparation des données

Avant d'insérer les données dans la base, il est nécessaire d'effectuer un traitement sur les données textuelles. En ce sens nous avons créé un algorithme qui effectue cette tâche.

Les principaux objectifs sont :

- Enlever les caractères spéciaux dans les contenus textuels
- Ne conserver que le texte
- Remplacer les articles n'ayant pas de contenu par « NAN »

Ce nettoyage a été effectué sur l'introduction des contenus textuels et sur les noms des éditeurs de chaque article avant d'être inséré dans la base donnée.

Cependant, pour obtenir les mots clés, un filtrage supplémentaire plus complexe a été effectué, intervenant à la suite de notre premier algorithme de filtrage. Les principaux objectifs de cet algorithme sont les suivants :

- Enlever les mots ayant une taille inférieure à 3 lettres
- Prédire les entités nommées (c'est-à-dire si le mot est un nom, un adjectif, nom commun etc…) pour chaque mot dans l'introduction d'un article
- Retenir les mots ayant les entités nommées suivantes (Nom commun, Nom ou Nom Propre)
- Enfin, retenir comme mot clé l'association des deux mots ayant le plus de fréquence dans l'introduction

Ainsi, cet algorithme nous permet après un filtrage très élaboré, d'avoir nos mots clés pour chaque article prêts à être insérés dans la base de données.

Puis nous avons structuré les informations afin de construire la base de données.

Le nettoyage de données a essentiellement été effectué avec le logiciel PYTHON grâce au package suivant : Natural Language Toolkit (NLTK) qui est une boîte-à-outil permettant la création de programmes pour l'analyse de texte.

## 5.3 Valorisation des données

Après avoir structuré les données, nous avons construit notre base de données. Pour cela, nous avons modélisé la base de données avec un MCD, que nous avons ensuite traduit en MLD.

Nous avons créé notre base de données afin de pouvoir stocker nos informations dans le but de faire des analyses statistique pertinentes. Ainsi nous avons créé trois tables (**Articles, Mots-clés et Auteurs**).

Il permet de modéliser la sémantique des informations, d'une façon compréhensible par l'utilisateur de la future base de données.

Nous avons utilisé le logiciel JMERISE pour effectuer notre Modèle conceptuel de données.

**Modèle conceptuel de données**

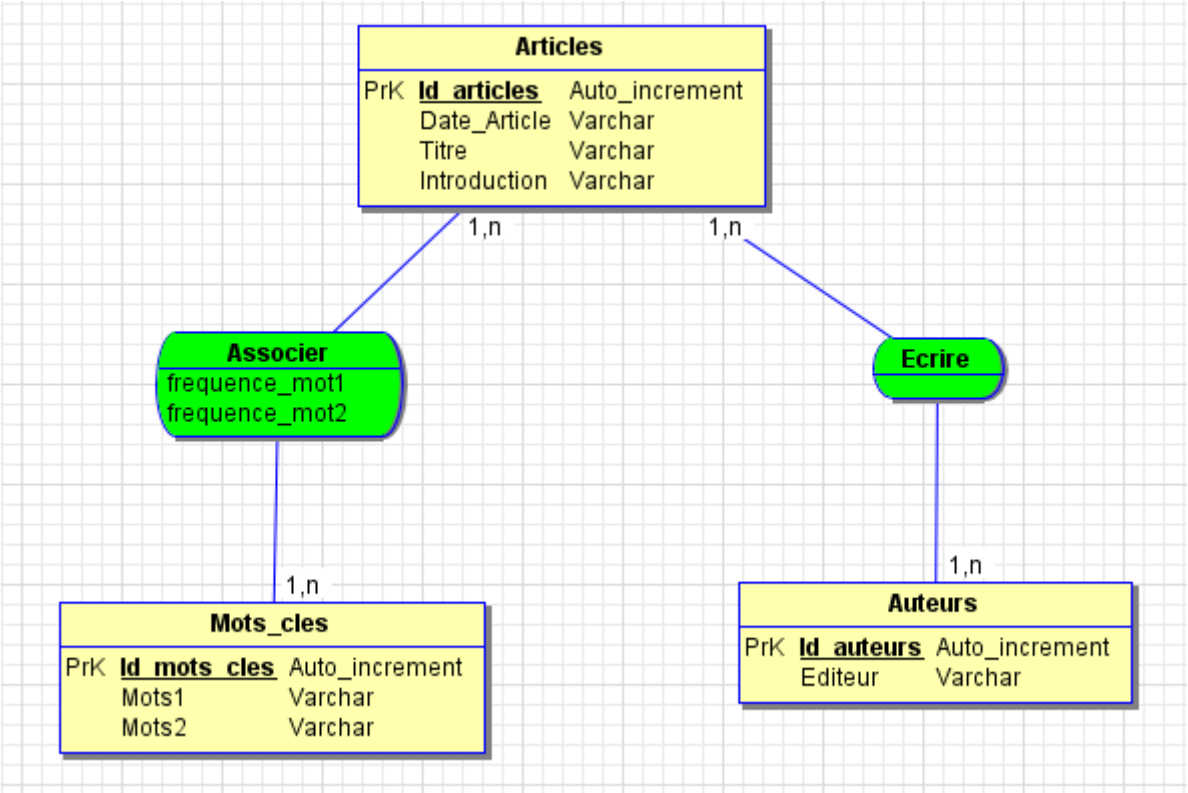


*Figure 3: Modèle conceptuel de données*

Sur ce schéma on peut voir que la table articles est composée des attributs suivants :

**Id_articles :** Identifiant de la table Articles

**Date_article :** la date de publication de l'Articles

**Titre :** Titre de l'Article

**Introduction** : Court résumé du contenu de l'Article contenant quelques mots clés

La table Auteurs a les attributs suivants : Id auteurs, Editeur.

**Id auteurs :** Identifiant de la table Articles

**Editeur :** les auteurs ou la maison d'édition de l'article

La table Mots_cles est composée des attributs suivants :

**Id_mots_cles :** Identifiant de la table Articles

**Mots1 :** Mot le plus fréquent de l'Article

**Mots2 :** Second mot le plus fréquent de l'Article

Nous avons deux relations (**Ecrire, Associer**) :

- Les articles sont écrits par un ou plusieurs auteurs et les auteurs écrivent un ou plusieurs articles
- Les articles sont associés à un ou plusieurs mots_cles et les mots_cles sont associés à un ou plusieurs articles. La relation Associer présente des attributs (frequence_Mot1, frequence_Mot2) car le mot clé associé à un article a une fréquence d'apparition dans l'article.

**Modèle logique de données**

Il permet de modéliser la structure selon laquelle les données seront stockées dans la base de données sous SQL SERVER.

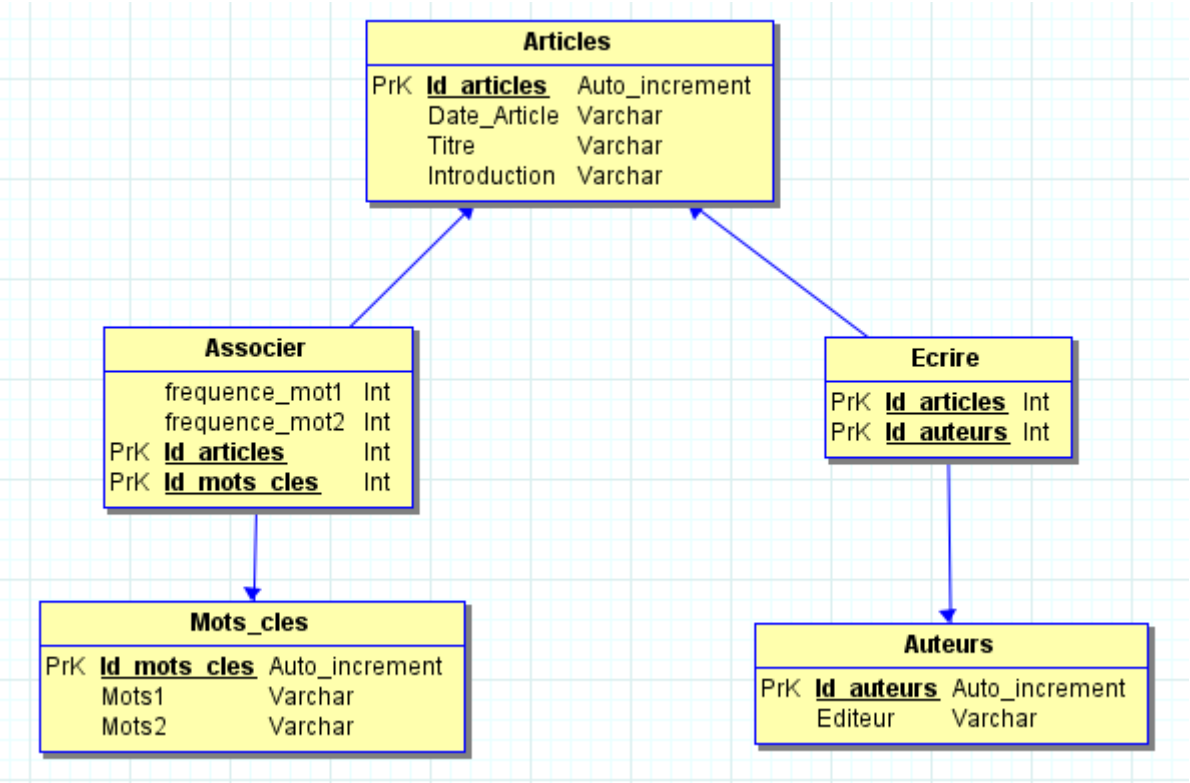


*Figure 4: Modèle logique de données*

La base de données a été construite comme convenu sous SQL Server. On a utilisé l'application SQL Server Management studio. Pour alimenter notre base de données on a dû générer les scripts d'insertion à l'aide de python.

Text mining : un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes.

La fonction sent2clean nous a permis à premier temps de supprimer les caractères spéciaux dans titre et introduction. Puis pour créer la table des mots clés on a supprimé tous les mots de tailles inférieures à 3. En suite les deux mots clés ont été définis par ordre de fréquence décroissante dans les articles.

Pour réaliser les requêtes d'interrogation des données on a utilisés la librairie python *pypyodbc* pour sélectionner et exporter les données sous format JSON afin de réaliser des graphiques sous python et d'autres sous EXCEL nécessitant au préalable une modification de la structure des données.

Voici les requêtes nécessaires pour réaliser les analyses statistiques ci-dessous.

```
26 #%%
27 # les nuages de mots (titre intro et date)
28 #Requête avec plusieurs résultats
29 SQLQ = "SELECT titre,introduction,Date_Article
30         FROM [dbo].[Articles] order by Date_Article "
31 cursor.execute(SQLQ)
32
33 # Method 1, simple reading using cursor
34 titre_intro_date_articles=[]
35 while True:
36     row = cursor.fetchone()
37     if not row:
38         break
39     else:
40         #print(row)
41         titre_articles.append(row)
```

*Figure 5: Requêtes SQL (1)*

Les données retournées par cette requête ont été utilisées pour réaliser les graphiques ci-dessous :

- nuages de mots des enseignes utilisant le plus le E-commerce
- nuages de mots des pays les plus cités dans les articles
- diagrammes circulaires de la répartition des enseignes dans les différents secteurs d'activité.
- diagramme en barre de l'évolution de la polarité des introductions
- diagramme en barre de l'évolution de la polarité des introductions

```
46 #%%
47 # les nuages de mots (titre intro et date)
48 #Requête avec plusieurs résultats
49 SQLQ1 = "
50         SELECT Mots1,Mots2,frequence_Mot1,frequence_Mot2,Date_Article
51         FROM [dbo].[Mots_cles] M, [dbo].[Articles] A,[dbo].[Associer] Ass
52              WHERE M.Id_mots_cles=Ass.Id_mots_cles
53              AND Ass.Id_articles=A.Id_articles "
54 cursor.execute(SQLQ1)
55
56 # Method 1, simple reading using cursor
57 mots_cles=[]
58 while True:
59     row = cursor.fetchone()
60     if not row:
61         break
62     else:
63         #print(row)
64         mots_cles.append(row)
```

*Figure 6: Requêtes SQL (2)*

Les données retournées par cette requête ont été utilisées pour réaliser les graphiques ci-dessous :

- Nuages de mots les plus fréquent le plus le E-commerce
- Diagramme en barre de l'évolution du nombre d'articles publiés par année
- Diagramme en barre de l'évolution de la polarité des mots clés
- Nuage de mots sur les mots clés
- Nuage des mots les plus cités sur les articles classés négatifs

**Interaction entre les différentes technologies utilisées :**

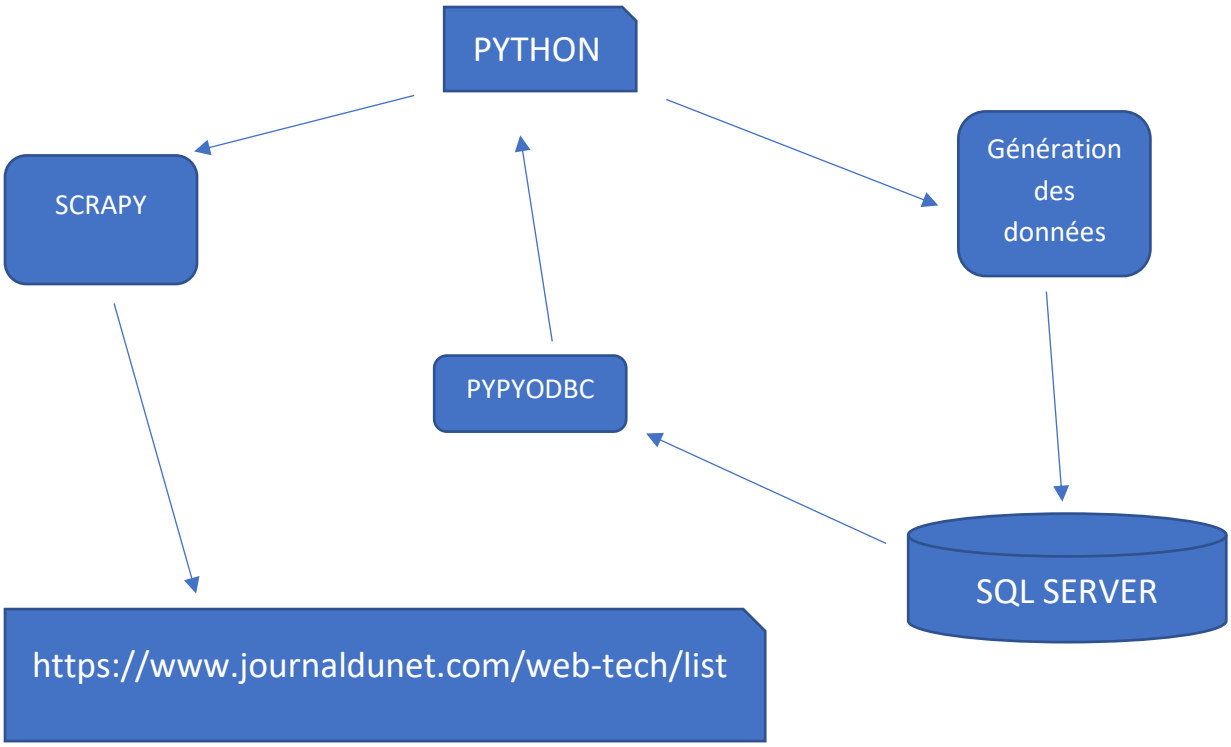


*Figure 7: Schéma montrant l'interaction entre les différentes technologies utilisées*

## 5.4 Visualisation des données

Dans cette partie, nous développons l'ensemble des analyses que nous avons effectuées.

### 5.4.1 Analyse des mots les plus fréquemment utilisés dans les articles

Nous avons analysé les mots les plus fréquemment utilisés dans nos articles afin de les mettre en valeur avec un nuage de mot.

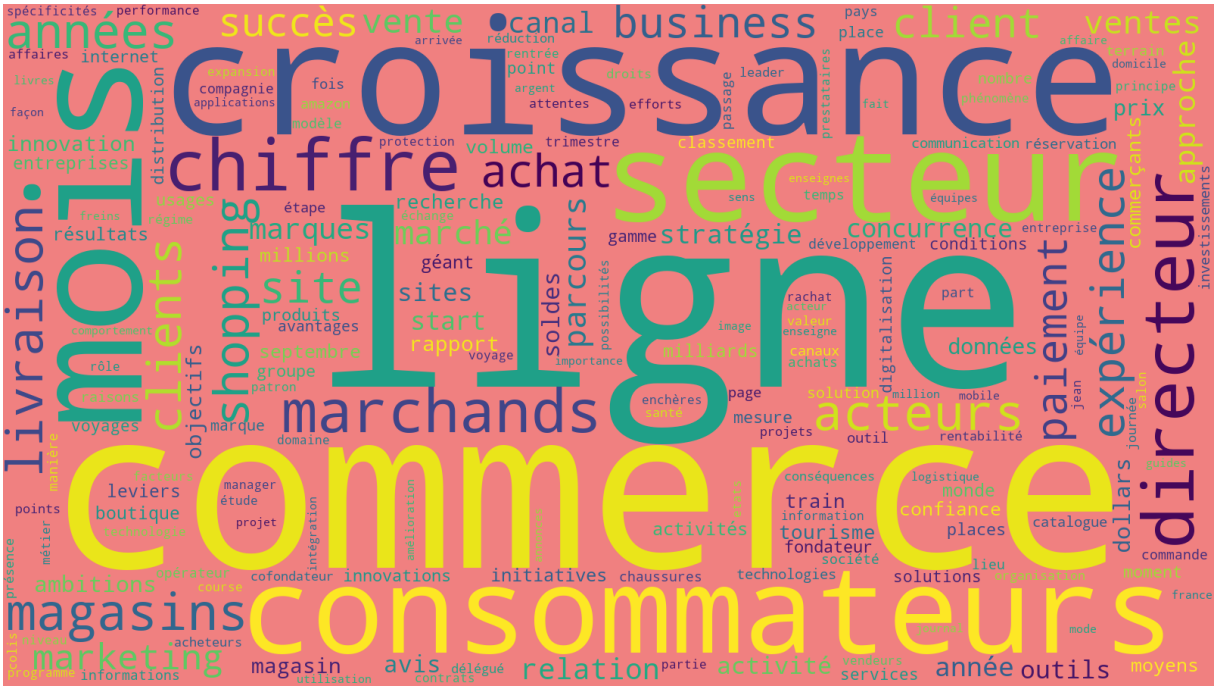


*Figure 8: Nuage de mots représentant les mots les plus fréquemment utilisés dans les articles*

**Interprétation :**

Sur ce graphique on a les mots les plus utilisés dans nos articles de 2006 à 2018, on constante que parmi les mots les plus fréquents on a « Commerce, ligne, Croissance » ce qui nous permet de déduire que le E-commerce est en plein expansion.

### 5.4.2 Analyse des sentiments

Pour prédire la polarité d'un article nous avons utilisé le logiciel PYTHON grâce au package textblob fr.

Dans cette partie nous avons essayé de comprendre le ressentis des personnes vis à vis de l'e-commerce à travers nos articles depuis 2006 à 2018.

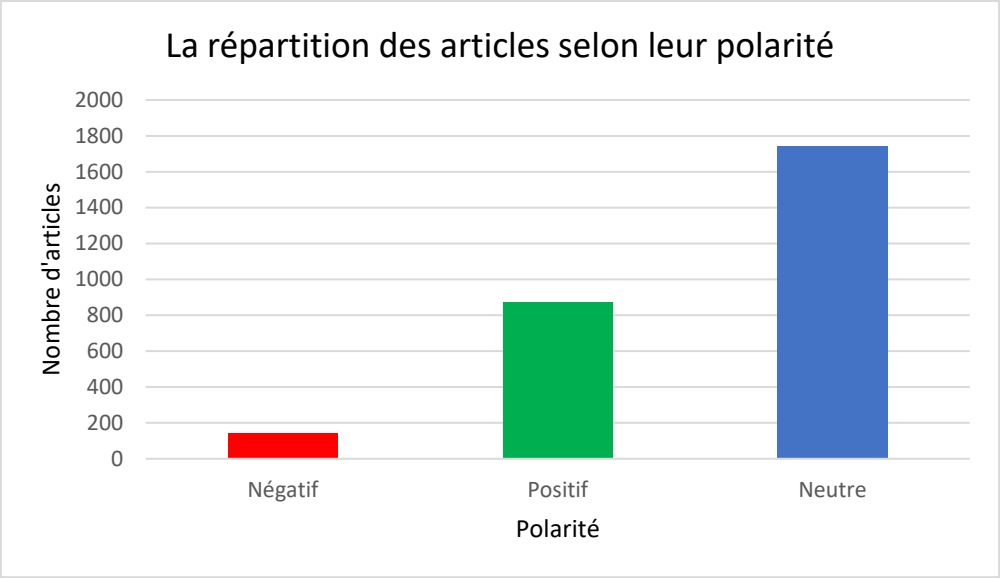


*Figure 9: Répartition des articles selon leur polarité*

**Interprétation :**

Ainsi nous avons constaté que dans nos articles la plupart ne porte pas de jugement sur ce nouveau phénomène, ce sont des articles que l'on considère comme neutres. On a 1744 articles considérés comme étant neutres. Par contre on a 1015 articles qui portent un jugement sur cette nouvelle tendance. Parmi les 1015 il y a 872 publications qui ont une opinion positive et 143 articles qui ont une opinion négative.

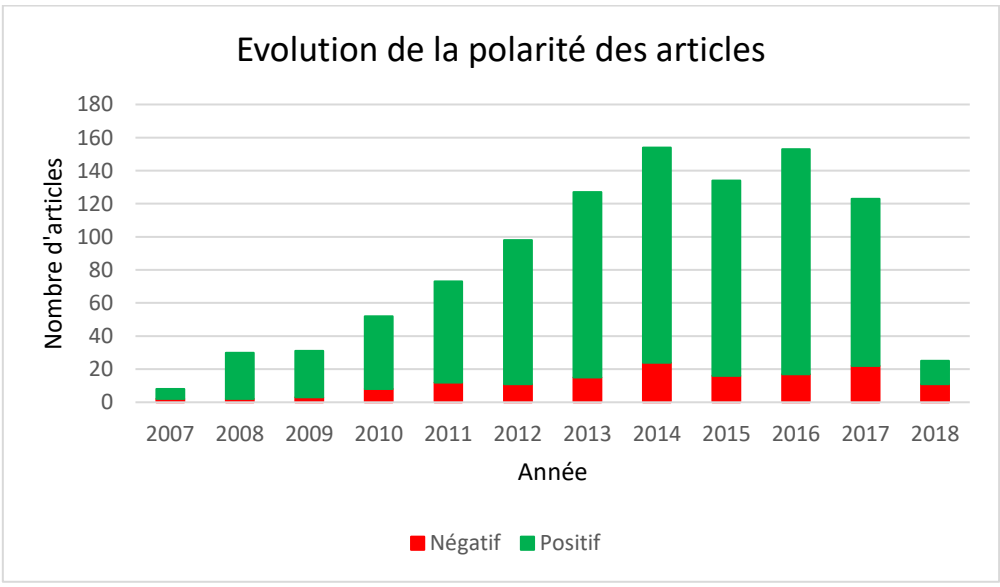


*Figure 10: Evolution de la polarité des articles*

**Interprétation :**

Ce graphique permet de voir le pourcentage d'évolution de la polarité des articles chaque année depuis 2006. On constate les années où il y a le plus d'articles portant un jugement sur le e-commerce sont de 2013 à 2017. En particulier l'année 2014 est l'année où il y a plus d'articles subjectifs. Globalement les articles positifs ont une large dominance ce qui nous permet de déduire que l'E-commerce est en plein expansion. Pour autant, les articles négatifs ne sont pas négligés car on peut voir que ces dernières années leur proportion augmente légèrement tandis que celui des articles positifs diminue légèrement.

**Analyse des mots les plus fréquemment utilisés dans les articles ayant une opinion négative sur le e-commerce**


*Figure 11: Nuage de mots représentant les mots les plus fréquemment utilisés dans les articles orientés négativement sur le e-commerce*

**Interprétation :**

Sur ce nuage de mots on a les mots les plus utilisés dans les articles ayant une opinion négative sur le e-commerce de 2006 à 2018. On constate que les mots les plus cités dans ces articles sont «Commerce, Ligne, France, Euros, Internet etc..».

### 5.4.3 Analyse des enseignes proposant des services de e-commerce

**Analyse des enseignes de e-commerce les plus citées dans nos articles**

Nous avons analysé les enseignes proposant du e-commerce et nous avons ainsi relevé les plus citées dans les articles.


*Figure 12: Nuage de mots représentant les noms d'enseignes proposant du e-commerce les plus fréquemment citées dans les articles*

**Interprétation :**

Sur ce nuage de mots, on peut voir que les enseignes les plus mises en valeur donc les plus citées sont des enseignes telles que : « Amazon, Ebay, Priceminister, Cdiscount, Carrefour, Groupon, Pixmania, Fnac, Airbnb, … ».

On remarque que « Amazon » se démarque largement par rapport aux autres enseignes. C'est en effet le leader du commerce en ligne.

On remarque aussi que les enseignes les plus mises en valeur sont notamment des enseignes dites de « pure players » c'est-à-dire des enseignes qui commercialisent exclusivement leurs produits en ligne, par exemple : « Amazon, Ebay, Priceminister, Cdiscount, Groupon, … » contrairement aux enseignes qui disposent également de boutiques physiques telles que : « Carrefour, Fnac, Auchan, Monoprix, Darty, Sephora, … » qui apparaissent moins souvent dans nos articles.

Cela confirme la place actuelle des sites exclusifs de e-commerce comme leaders du marché.

**Répartition des sites de e-commerce cités dans nos articles dans les principaux secteurs d'activité**

Afin de visualiser la répartition dans les principaux secteurs d'activité des sites de e-commerce cités dans nos articles, nous avons produit un diagramme circulaire. Il présente les proportions d' enseignes de e-commerce citées dans nos articles dans différents secteurs d'activité.

Nous avons choisis sept secteurs d'activité:

- Général : Les enseignes proposant tout type de produits. Ex : Carrefour, Amazon, Auchan, Leclerc, …
- Mode/Beauté : Les enseignes proposant des vêtements, produits de beauté, cosmétiques, bijoux, … telles que : Sephora, BrandAlley, Showroomprivé, Zalando,...
- High-tech : Les enseignes vendant des produits de nouvelles technologies telles que : Fnac, Darty, Boulanger, Conforama,…
- Voyage : Les enseignes proposant des produits en rapport avec le voyage telles que : Airbnb, Booking, Blablacar, Accorhotels, …
- Alimentaire : Les enseignes proposant des produits alimentaires (notamment pour des livraisons de repas) tels que : Ubereats, Foodora, Deliveroo, Chronoresto, …
- Automobile : Les enseignes proposant des produits en rapport avec l'automobile. Ex : AramisAuto, Oscaro, Allopneus,…
- Autre : les autres secteurs d'activité tels que le jardinage (OoGarden, Interflora,…), la billeterie (Ticketnet,…), la décoration, le bricolage par exemple.

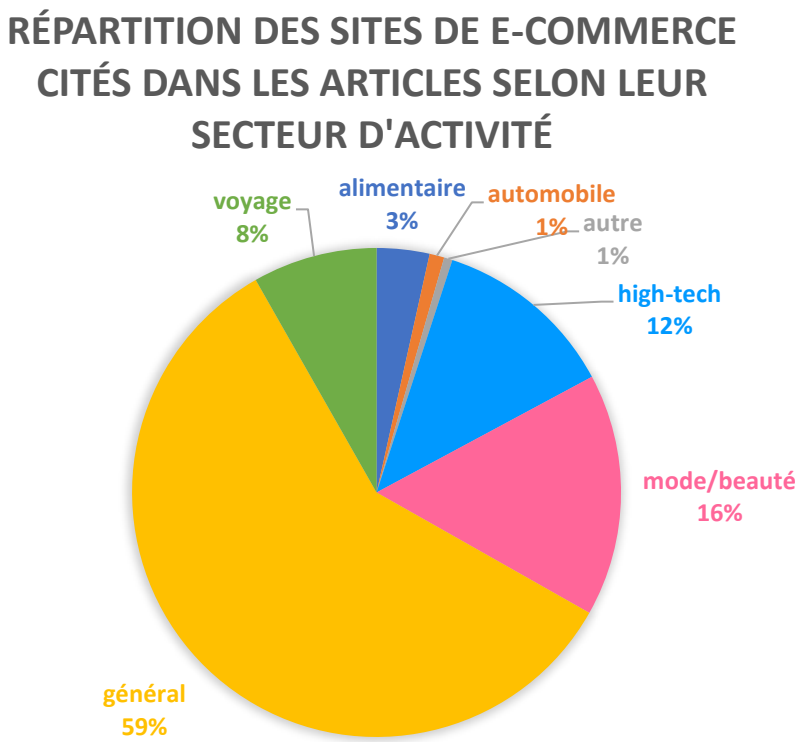


*Figure 13: Diagramme circulaire montrant la répartition des sites de e-commerce cités dans les articles selon leur secteur d'activité*

**Interprétation :**

On voit que les enseignes de e-commerce les plus citées sont des enseignes généralistes (à 59%). En effet, ce sont des enseignes qui vendent tout type de produits et qui sont très connues et répandues.

Les secteurs de la Mode/beauté du High-Tech et du Voyage se détachent également représentant respectivement 16%, 12% et 8% de la proportion d'enseignes citées dans les articles. En effet, les sites de e-commerce de ces secteurs-là se développent de plus en plus et attirent de plus en plus de clients.

Enfin, les enseignes des secteurs de l'automobile et d'autres secteurs comme le jardinage ou la billetterie par exemple sont peu représentés dans nos articles.

Cette répartition est cohérente et semble bien refléter les types de sites de e-commerce les plus utilisés par les clients.

### 5.4.4 Analyse des pays les plus fréquemment cités dans les articles

Nous avons recherché les noms de pays les plus cités dans nos articles. Nous avons ainsi pu produire le nuage de mots suivant :

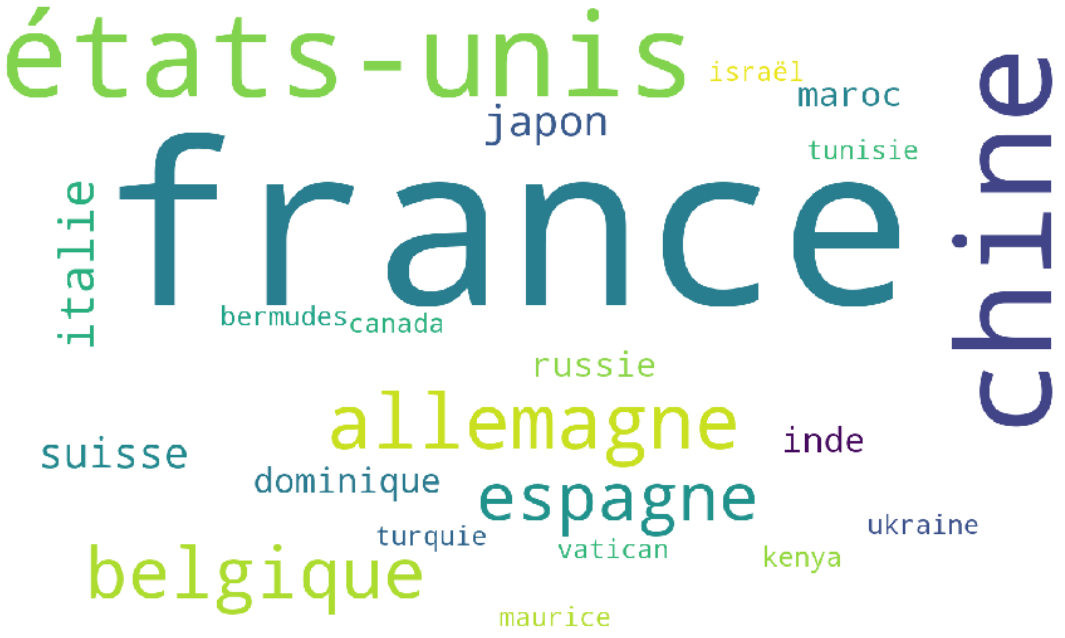


*Figure 14: Nuage de mots représentant les pays les plus cités dans les articles*

**Interprétation :**

Ce nuage de mot met particulièrement en avant les pays suivants : France, Etats-Unis, Chine, Allemagne, Espagne, Belgique,…

On peut en déduire que la plupart des évolutions, tendances, nouveautés dans le e-commerce ont lieu dans ces pays-là, et en particulier en France et aux Etats-Unis.

### 5.4.5 Evolution du nombre d'articles publiés par année

Nous avons produit un graphique montrant l'évolution du nombre d'articles publiés au fil des années :

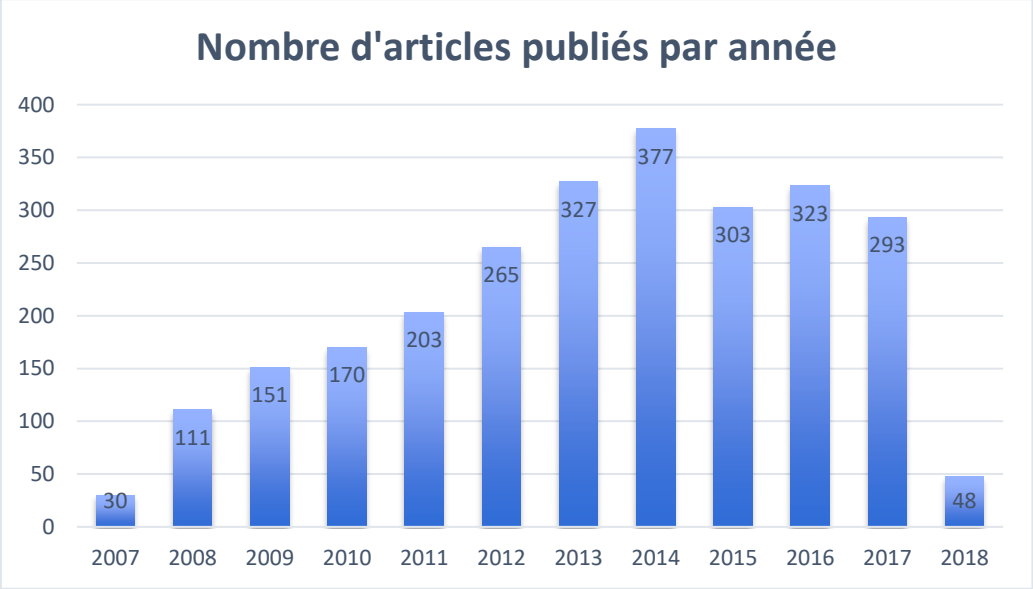


*Figure 15: Histogramme montrant l'évolution du nombre d'articles publiés par année*

**Interprétation :**

Cet histogramme montre une augmentation du nombre d'articles de 2007 à 2014 puis une légère baisse de 2015 à 2017. Cette analyse reflète une expansion progressive du e-commerce jusqu'à 2014 puis une stagnation de la croissance du e-commerce ces dernières années.

# 6 Gestion de configuration

La gestion de configuration peut être utilisée à plusieurs fins. Nous l'utilisons afin de stocker et tracer les différentes versions ou révisions de toute information destinée à être utilisée par notre système (matériel, logiciel, document, donnée unitaire, etc.). En d'autre terme, la gestion de la configuration gère les évolutions du produit pendant tout son cycle de vie, en termes d'adéquation entre ce qui est spécifié et ce qui est réalisé. Ainsi nous avons utilisé le logiciel Git HUB par le biais de Git Kraken pour pouvoir synchroniser notre travail et assurer la traçabilité de chaque élément produit.

Le principe de la gestion de configuration de Git hub pour un fichier est illustré par le schéma suivant :

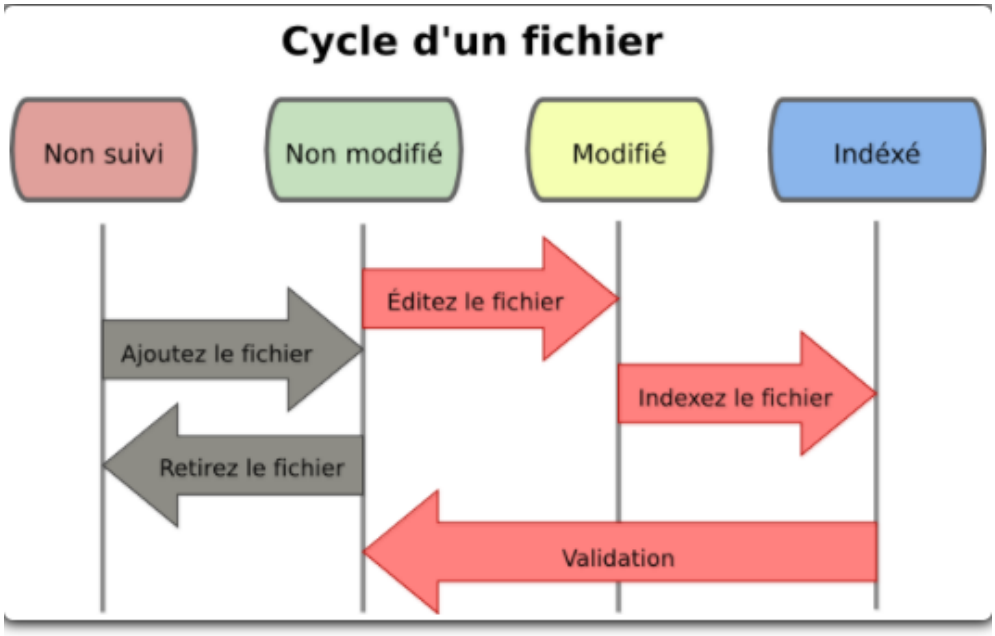


*Figure 16: Cycle de vie des états d'un fichier*

Le schéma ci-dessous permet de voir la structure de la gestionnaire de configuration des fichiers par GitKraken :

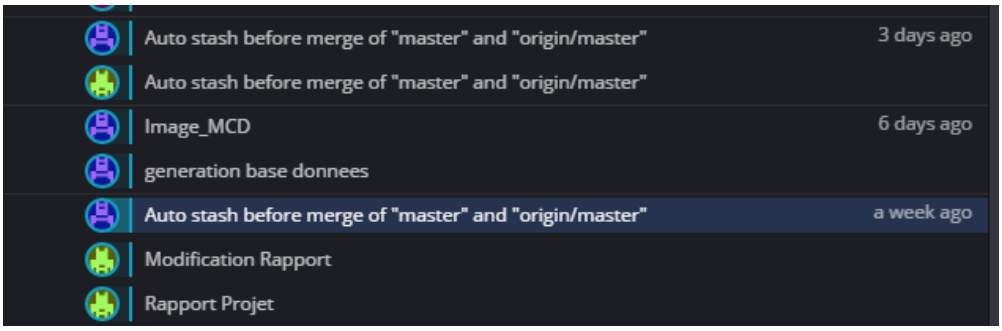


*Figure 17: Historique des modifications sur GitKraken*

Pour chaque fichier quand on le sélectionne on a la description de la modification, la version, la date et l'auteur qui apparait.

# 7 Assurance et Contrôle qualité

**Démarche Qualité**

La démarche qualité consiste à concevoir et développer un service de qualité correspondant aux attentes des utilisateurs.

Plus précisément, le processus de qualité consiste à identifier les exigences de qualité et les normes à respecter pour le projet et ses livrables.

Pour cela, nous avons produit de la documentation sur la façon dont le projet démontre sa conformité aux exigences et aux normes de qualité appropriées. C'est-à-dire les documents suivants : la charte de codage, la charte graphique, les règles à suivre pour le rapport ainsi que le cahier de recette.

L'intérêt principal de ce processus est qu'il fournit les directives et les orientations de management et de validation de la qualité tout au long du projet.

**Assurance et Contrôle Qualité**

L'assurance qualité est, selon la définition ISO 9000, la partie du management de la qualité visant à donner confiance en ce que les exigences pour la qualité soient satisfaites.

Il s'agit donc de piloter le processus de développement en mettant en place un mécanisme de prévention des défauts qui consiste à définir, au début du projet, les activités de vérification et de validation du cycle de développement.

Ce mécanisme d'assurance qualité s'appuie sur l'engagement conjoint du client et du fournisseur.

Les revues incluses dans la démarche d'assurance qualité, portant à la fois sur le processus et sur le produit, permettent de s'assurer de la conformité du produit.

Elles permettent au client de suivre, d'apprécier et d'anticiper l'avancement du projet, et à l'équipe de projet d'organiser son travail.

**Revues Client-Fournisseur**

- Revue d'avancement du 19/01/2018 : Choix du sujet et de la problématique

Après avoir reçu le cahier des charges du client, nous avons discuté sur l'organisation de notre projet.

Nous avons notamment réfléchis à plusieurs sujets d'études potentiellement intéressants.

Nous nous sommes intéressés au commerce en ligne, qui évolue particulièrement ces dernières années.

Nous avons finalement choisis d'étudier « L'évolution du e-commerce à l'ère du digital ».

- Revue d'avancement du 26/01/2018 : Validation du sujet et Choix du site Web

Nous avons soumis notre idée de sujet au client qui l'a ensuite validé.

Nous avons ainsi réfléchi à notre équation de recherche afin de trouver des informations pertinentes.

Nous avons fait de nombreuses recherches afin de choisir le site web duquel nous allions extraire les données. L'objectif étant de trouver un site Web contenant un grand nombre d'articles, homogènes, et les plus pertinents possibles par rapport à notre sujet.

Nous avons finalement choisis la rubrique concernant le e-commerce du site du Journal Du Net, présentant au fil des années des articles d'actualités du secteur du e-commerce.

Le client a ensuite validé le site Web choisis.

- Revue d'avancement du 29/01/2018 : Choix de la plate-forme

Nous avons ensuite dû choisir une plate-forme nous permettant d'extraire les données du site Web préalablement choisis.

Nous avons fait des recherches sur les différents outils de Web Crawling existants tels que HTTrack et Scrapy et nous avons finalement décidé de choisir Scrapy. Nous l'avons choisis car c'est un outil open source performant, ergonomique et cohérent avec le langage de programmation « Python » que nous utilisons puisqu'il suffit d'installer le package afin de l'utiliser avec Python.

- Revue d'avancement du 02/02/2018 : Extraction des données

Après avoir défini le site Web nous permettant d'extraire les données et de produire notre corpus, ainsi que l'outil de Web Crawling nous permettant de récupérer les données du site, nous avons réfléchi aux données qui seraient pertinentes à extraire.

Nous avons ainsi choisis de récupérer certaines données des articles en particulier.

Nous avons également discuté avec les enseignants de la structure du rapport.

- Revue d'avancement du 22/02/2018 : Nettoyage des données

Après avoir extrait les données du site Web via Scrapy, nous avons réfléchi à des règles de nettoyage des données. Nous avons ainsi fixé des conventions de filtrage à ce moment-là.

- Revue d'avancement du 26/02/2018 : Mise en forme des données

Nous avons discuté de la mise en forme des données que nous avons extraites puis nous les avons structurées de façon à produire ensuite le MCD. Nous avons en amont discuté des analyses statistiques à produire à partir des données par la suite.

- Revue d'avancement du 02/03/2018 : Validation du Modèle Conceptuel de Données (MCD)

Nous avons réfléchi à une modélisation de notre base de données telle que nous puissions faire les analyses statistiques voulues par la suite. Nous avons ainsi produit et fais validé le MCD par le client par la même occasion. Notre MCD comporte trois tables et deux associations.

- Revue d'avancement du 05/03/2018 : Implémentation de la Base de Données

Après la phase de préparation des données, comportant le nettoyage ainsi que la mise en forme des données, nous avons créé et implémenté la base de données. Nous avons utilisé SQL Server pour cela.

- Revue d'avancement du 09/03/2018 : Interrogation de la Base de Données

Nous avons effectué des requêtes interrogeant la base de données nécessaires pour effectuer nos analyses statistiques. Puis nous avons établi une liste d'analyses statistiques à produire. Nous avons réfléchi aux graphiques que nous pouvions faire.

Nous avons produit des nuages de mots notamment car ils permettent de mettre en valeur les mots ou concepts ressortant le plus des articles.

Nous avons ainsi généré des graphiques variés.

- Revue d'avancement du 23/03/2018 : Finalisation du projet

Lors de cette dernière phase, nous avons finalisé les dernières analyses statistiques ainsi que le rapport et vérifié tous nos contenus et documents.

**Tests unitaires**

Voici ci ci-dessous en capture d'écran les tests unitaires réalisés sur nos fonctions :

```
#test unitaire
def test_sentense2cleanTokens():
    text = 'le voiture %% est en panne/00'
    return sentense2cleanTokens(text) == 'le voiture est en panne'

def test_rp_vide():
    data={'0':[''],'1':['le voiturer est en panne']}
    return rp_vide(data) == ['Nan', 'le voiture est en panne']

def test_stop_word():
    list_phrase=['Nan', 'le voiturer est en panne']
    return stop_word(list_phrase) == {0: ['Nan'], 1: ['voiture', 'panne']}

def test_key_word():
    data={0: ['Nan'], 1: ['voiture', 'panne']}
    return key_word(data) == {0: [('Nan', 1)], 1: [('panne', 1), ('voiture', 1)]}
```

*Figure 18: Capture d'écran des test unitaires effectués (1)*

```
if test_sentense2cleanTokens()== True:
    print('test unitaire sentense2clean Valide')
if test_rp_vide()== True:
    print('test unitaire test_rp_vide Valide')
if test_stop_word()== True:
    print('test unitaire test_stop_word Valide')
if test_key_word()== True:
    print('test unitaire test_key_word Valide')
```

*Figure 19: : Capture d'écran des test unitaires effectués (2)*

Résultats obtenus : tous les tests sur les fonctions sont valides pour être utilisés dans la suite du projet.

```
....
test unitaire sentense2clean Valide
test unitaire test_rp_vide Valide
test unitaire test_stop_word Valide
test unitaire test_key_word Valide
```

*Figure 20: Capture d'écran des résultats des test unitaires effectués (3)*

# 8 Bilan du projet

En conclusion on peut dire que le projet nous a beaucoup apporté, tant en terme de compétences acquises qu'en terme de savoir-faire et de savoir-être. Vous trouverez ci-après la liste de ces points principaux.

Les principales compétences que nous avons acquises sont les suivantes : prise en main des technologies GitHub, Trello, Slack, des librairies python (scrapy , pypyodbc, ..). Concernant le savoir-faire nous avons pris conscience de l'importance de l'entraide au sein du groupe notamment entre les membres du groupe ayant des profils orientés programmation et les autres orientés base de données. De plus, nous avons su bien communiquer et cela a été indispensable au bon avancement du projet, nous avons donc pu constater que la communication est essentielle. Enfin, pour ce qui est du savoir-être, nous avons développé notre sens de l'écoute.

Une chose particulièrement importante est le fait que nous avons pris différents postes (administration BD, programmation, génération des données). Bien que cela nous ait ralenti, cela a permis à tous les membres du groupes de gagner en polyvalence ce qui est un plus non négligeable.