

Projet Tableau de Bord Master 1 SID

*L’évolution du e-commerce à l’ère du digital*

Suivi des modifications

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Objet de la modification** | **Nom** | **Date** |
| 1.00 | Initialisation du document | Célia SARTORI, Samba Diallo WADE, Elhadji Fallou FALL, Serigne DIAW | 08/03/18 |
|  |  |  |  |
|  |  |  |  |

Sommaire

[1 Objet du document 4](#_Toc509959019)

[2 Problématique 4](#_Toc509959020)

[3 Documents applicables et de référence 5](#_Toc509959021)

[3.1 Documents applicables 5](#_Toc509959022)

[3.2 Documents de référence 5](#_Toc509959023)

[3.3 Terminologie 6](#_Toc509959024)

[4 Organisation de travail 7](#_Toc509959025)

[4.1 Répartition des rôles 7](#_Toc509959026)

[4.2 Organisation et répartition des tâches 7](#_Toc509959027)

[5 Processus de développement 9](#_Toc509959028)

[5.1 Recherche d’informations et Collecte des données 9](#_Toc509959029)

[5.2 Préparation des données 9](#_Toc509959030)

[5.3 Valorisation des données 10](#_Toc509959031)

[5.4 Visualisation des données 13](#_Toc509959032)

[5.4.1 Analyse des mots les plus fréquemment utilisés dans les articles 13](#_Toc509959033)

[5.4.2 Analyse des sentiments 13](#_Toc509959034)

[5.4.3 Analyse des enseignes proposant des services de e-commerce 16](#_Toc509959035)

[5.4.4 Analyse des pays les plus fréquemment cités dans les articles 18](#_Toc509959036)

[5.4.5 Evolution du nombre d’articles publiés par année 19](#_Toc509959037)

[6 Gestion de configuration 20](#_Toc509959038)

[7 Assurance et Contrôle qualité 21](#_Toc509959039)

[8 Bilan du projet 23](#_Toc509959040)

Table des illustrations

[Figure 1: Modèle SADT 7](#_Toc509959005)

[Figure 2: Diagramme de Gant 8](#_Toc509959006)

[Figure 3: Modèle conceptuel de données 11](#_Toc509959007)

[Figure 4: Modèle logique de données 12](#_Toc509959008)

[Figure 5: Nuage de mots représentant les mots les plus fréquemment utilisés dans les articles 13](#_Toc509959009)

[Figure 6: Répartition des articles selon leur polarité 14](#_Toc509959010)

[Figure 7: Evolution de la polarité des articles 14](#_Toc509959011)

[Figure 8: Nuage de mots représentant les mots les plus fréquemment utilisés dans les articles orientés négativement sur le e-commerce 15](#_Toc509959012)

[Figure 9: Nuage de mots représentant les noms d’enseignes proposant du e-commerce les plus fréquemment citées dans les articles 16](#_Toc509959013)

[Figure 10: Diagramme circulaire montrant la répartition des sites de e-commerce cités dans les articles selon leur secteur d'activité 17](#_Toc509959014)

[Figure 11: Nuage de mots représentant les pays les plus cités dans les articles 18](#_Toc509959015)

[Figure 12: Histogramme montrant l'évolution du nombre d'articles publiés par année 19](#_Toc509959016)

[Figure 13: Cycle de vie des états d'un fichier 20](#_Toc509959017)

[Figure 14: Historique des modifications sur GitKraken 20](#_Toc509959018)

# 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.

# Problématique

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

Nous avons ainsi choisi 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.

# Documents applicables et de référence

## 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

## 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*

## Terminologie

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

# Organisation de travail

## Répartition des rôles

Nous avons réparti les fonction 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.

## 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**

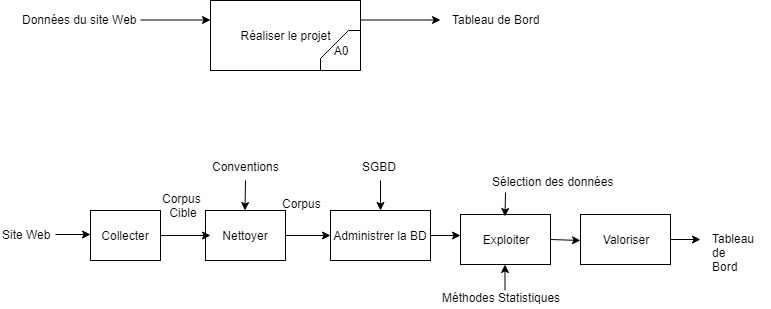
****

Figure : Modèle SADT

* **Planning prévisionnel**

Nous avons créé un diagramme de Gant :

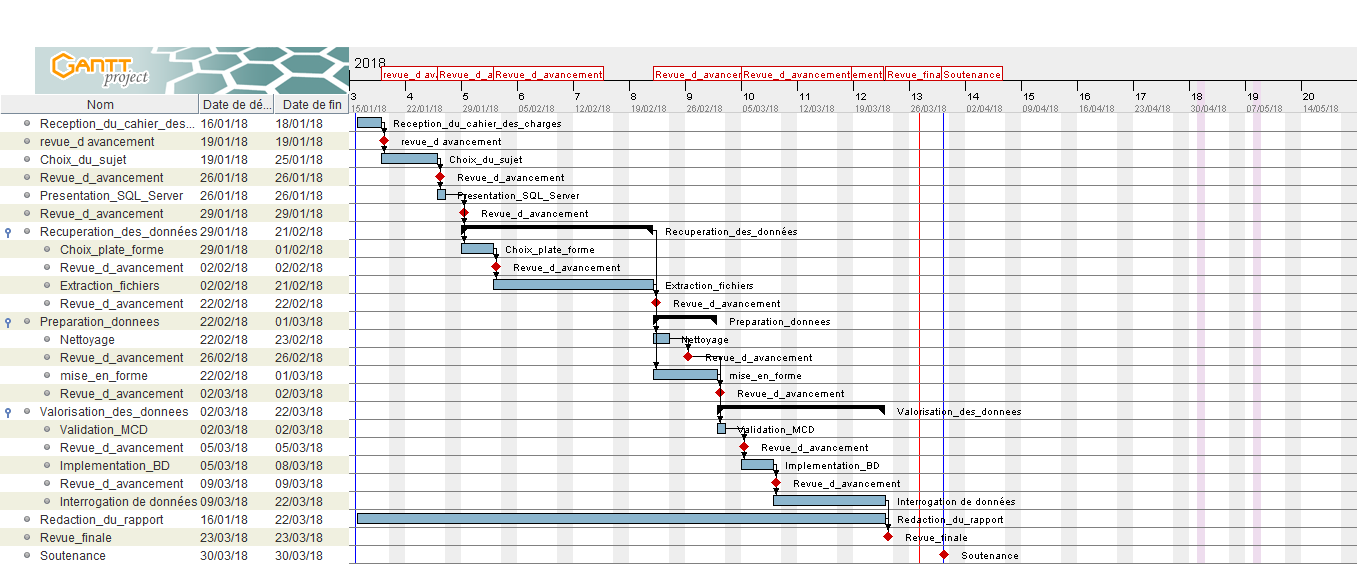


Figure : Diagramme de Gant

* **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.

SCRUM

# Processus de développement

## 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.

## 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.

## 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**

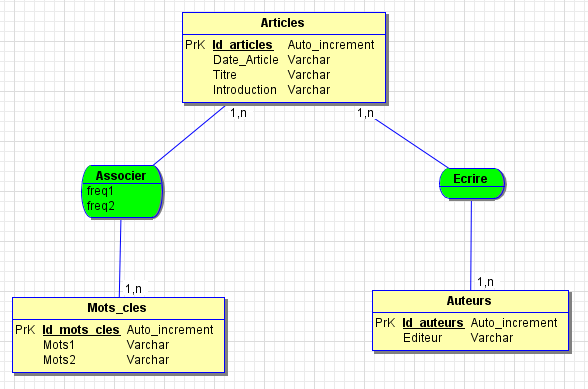


Figure : Modèle conceptuel de données

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

Date\_article, Titre, Introduction.

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

La table Mots\_cles est composée des attributs suivants : Id\_mots\_cles, Mots1, Mots2.

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 (freq1, freq2) car le mot clé associé à un article a une fréquence d’apparition dans l’introduction de l’article

**Modèle logique de données**

Il permet de modéliser la structure selon laquelle les données seront stockées dans la future base de données.

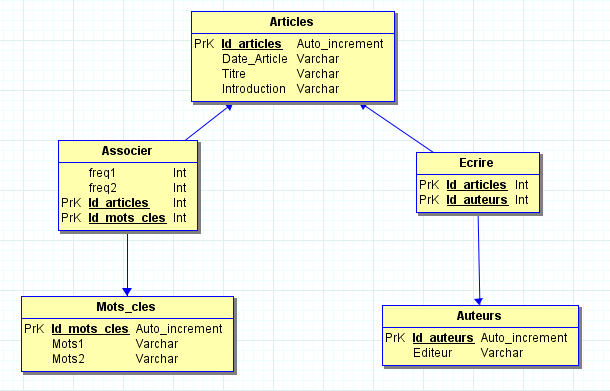


Figure : Modèle logique de données

Abdoul

* Construction + Administration BD → SGBD
  + MCD

## Visualisation des données

### 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.

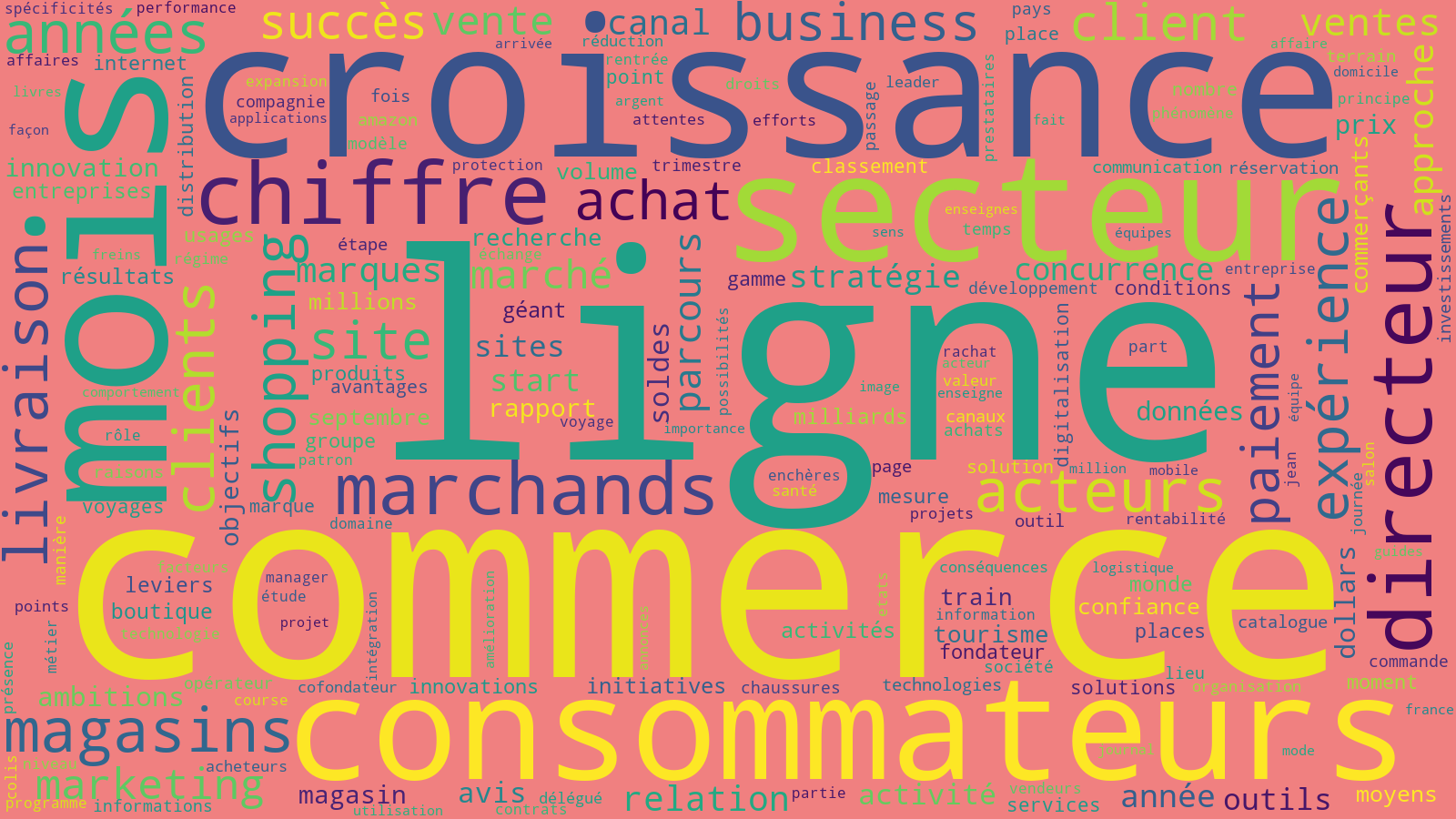


Figure : Nuage de mots représentant les mots les plus fréquemment utilisés dans les articles

**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.

### 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 : 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 : 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 : 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..».

### 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.

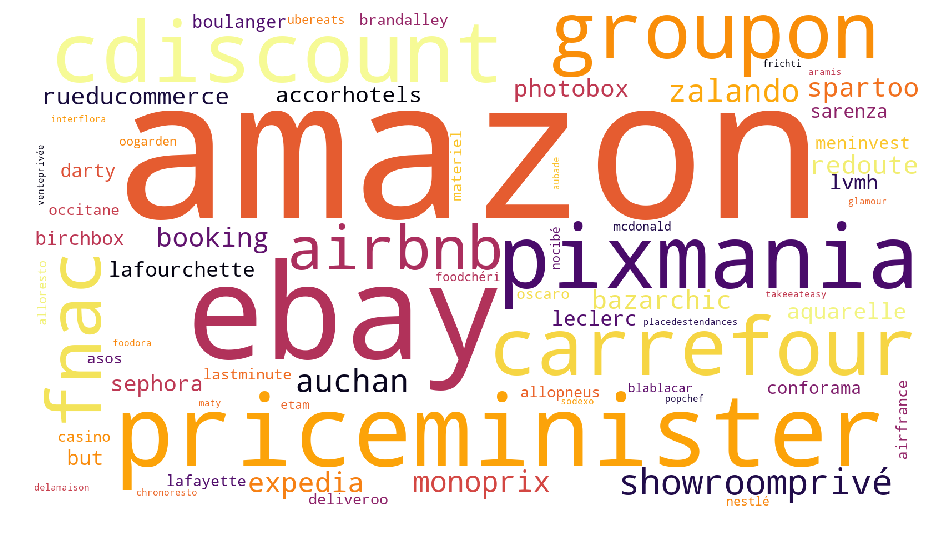


Figure : Nuage de mots représentant les noms d’enseignes proposant du e-commerce les plus fréquemment citées dans les articles

**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 : 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.

### 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 :

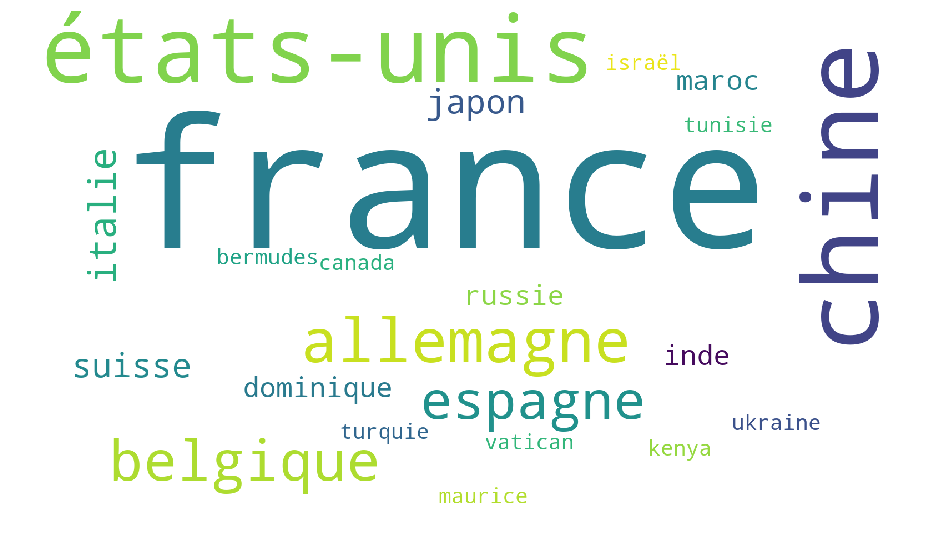


Figure : Nuage de mots représentant les pays les plus cités dans les articles

**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.

### 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 : 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.

# 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 :

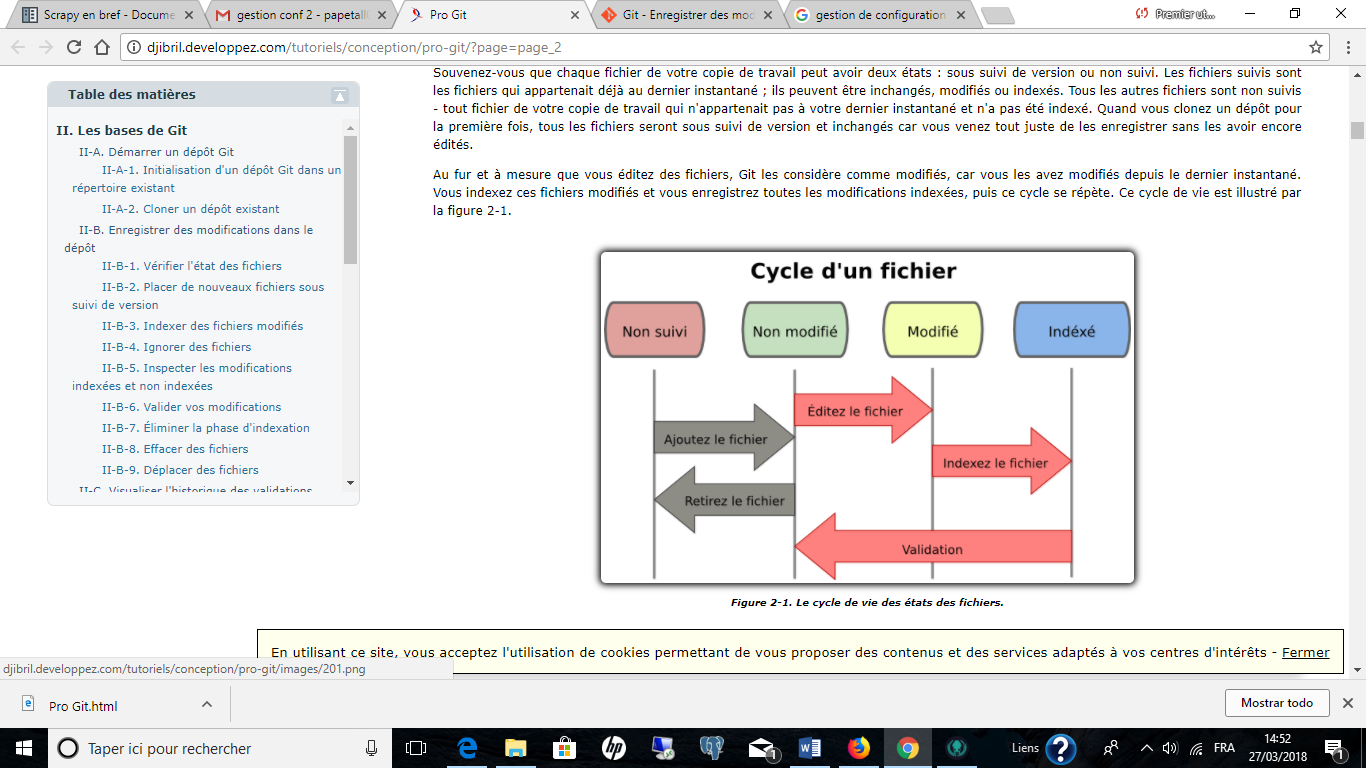


Figure : Cycle de vie des états d'un fichier

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

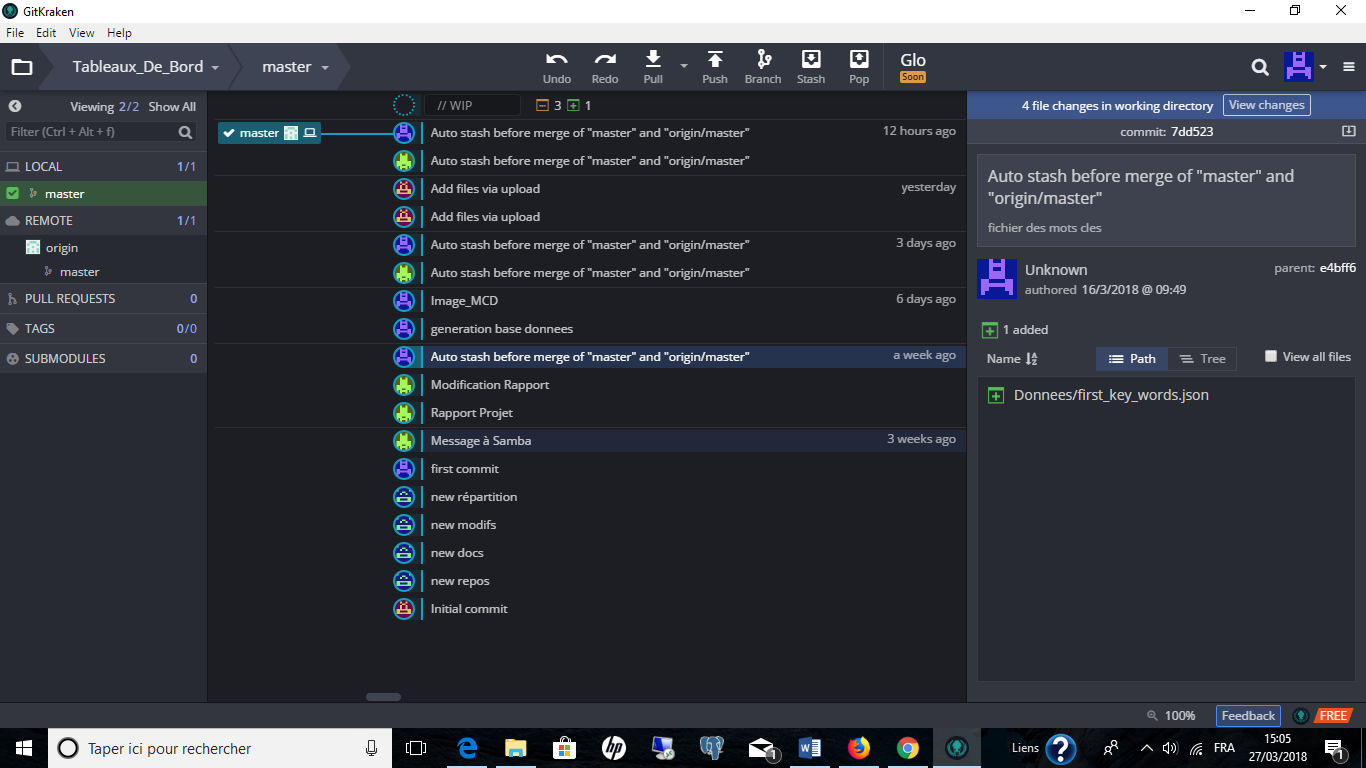


Figure : Historique des modifications sur GitKraken

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

# 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.

* 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.

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

* 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 05/03/2018 : 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 05/03/2018 : 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.

# Bilan du projet