Université Catholique de l'Afrique de l'Ouest
Faculté des Sciences de Gestion
INSTITUT SUPERIEUR DE GESTION SAINT MICHEL
Science – Foi – Action

Agrément : n° 05/AG/SAC/MESUCURRS/DES/DFS

Habilitation : N° RepSEN/Ensup-priv/HA/015-2017

SPECIALITE : INFORMATIQUE DE GESTION

MEMOIRE

Présenté par

Mme RABIYATOU DIOUF

Pour l'obtention du diplôme de

Master en INFORMATIQUE DE GESTION

SUJET

ETUDE ET CONCEPTION D'UN MOTEUR D'EXTRACTION AUTOMATIQUE DES ARTICLES ET AFFIRMATIONS JOURNALISTIQUES : LE CAS DE LA PRESSE EN LIGNE SENEGALAISE

Soutenu à UCAO/Saint Michel le X/X/2019 devant le jury composé de :

Président : Pr Cheikh Amadou Bamba GUEYE	Professeur Titulaire en Informatique	**UCAD**
Directeur de mémoire : Pr SAMBA NDIAYE	Professeur Titulaire en Informatique	**UCAD**
Co-encadreur : Dr Edouard Ngor Sarr	Assistant en Informatique	**UCAO**
Examinateur : Dr Reine Marie MARONE	Assistante en Informatique	**UCAO**

Année 2017-2018

DEDICACES

Je dédie ce mémoire

- *À ma mère fatou THIAM,*
- *À mon défunt père Mamadou DIOUF,*
- *À mon chéri Becaye Sonnar SENGHOR,*
- *Et à mes enfants Souleymane Sonnar SENGHOR et Siga Saly SENGHOR.*

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je commencerais par mon encadreur, le Dr Edouard Ngor SARR. Je tiens à vous exprimer mes plus vifs remerciements. Vous avez été pour moi un encadreur disponible malgré vos nombreuses charges. Vos compétences, votre rigueur et votre clairvoyance m'ont beaucoup appris et aidé à arriver au bout du tunnel.

J'exprime tous mes remerciements à l'ensemble des membres du jury.

Je remercie tous le personnel de l'Université Catholique de l'Afrique de l'Ouest (UCAO Saint Michel) de Dakar. Un grand merci à mon directeur des études M Rémy BASSE, à tous mes professeurs de l'UCAO. Ils m'ont fournies les outils nécessaires à la réussite de mes études universitaires. A mes amis, à la promotion Informatique de Gestion 2016.

Les mots les plus simples étant les plus forts, j'adresse toute mon affection à ma famille, Mamadou BAKHOUM, Adji, Khady, Coumba, Babacar, Ibrahima, à Malick Faye.

Je remercie aussi ma belle-famille Lamine, Souleymane, Awa, Marie, Yacine, Fatou particulièrement à ma belle-mère Ya Saly merci Yaya pour le soutien et les encouragements.

Et enfin, mon bien-aimé, mon mari, Becaye Sonnar SENGHOR, pour son amour, sa tendresse, sa présence et sa grande patience surtout, sans qui je ne serai peut-être pas arrivé au bout de cette aventure. Merci à toi chéri.

RESUME

Aujourd'hui le web est une mine d'or de données raison pour laquelle il est devenu le principal corpus des journalistes. Ces derniers, dans leur quête de la vérité, parcourent quotidiennement et le plus souvent de façon manuelle, des milliers de sources à la recherche d'informations pertinentes. Mais dans le contexte du web journalisme avec une quantité de données augmentant de façon exponentielle et en temps réel, cette fouille manuelle est devenu impossible et de plus en plus laborieuse. Une des techniques actuelles utilisées pour pallier ce problème est, l'extraction automatique d'informations sur les pages web, plus connue sous son nom anglais «web scraping ». L'objectif principal du *web scraping* est de faire ressortir d'une page web, des données ciblées et très structurées avec un effort humain réduit. Dans ce mémoire, nous présentons, un extracteur automatique d'articles de presse mis en application sur 30sites web sénégalais d'informations.

Mots-clés : Web Scraping, Journalisme, Fusion de données, JSON, CSV.

ABSTRACT

Today the web is a gold mine of data because of which it has become the main corpus of journalists. The latter, in their quest for the truth, travel daily and often manually, thousands of sources in search of relevant information. But in the context of web journalism with a growing amount of data exponentially and in real time, this manual search has become impossible and more and more laborious. One of the current techniques used to overcome this problem is the automatic extraction of information on web pages, better known by its English name "web scraping". The main goal of web scraping is to bring out from a web page, targeted and highly structured data with reduced human effort. In this job, we present, an automatic extractor of newspapers implemented on 30 Senegalese web sites of information.

Keywords: Web Scraping, Journalism, Data Fusion, JSON, CSV.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

- SGBD : Système de Gestion de Base de données
- RI : Recherche d'information
- TANL : Traitement automatique de la langue naturelle.
- JSON : JavaScript Object Notation
- CSS : Cascading Style Sheets (Feuilles de Style en Cascade)
- NLTK : Natural Language Toolkit (boîte à outils en langage naturel)
- CSV : Comma-separated values (Valeurs Séparées par des Virgules)
- SQL : Structured Query Language (Langage de Requête Structurée)
- HTML : HyperText Markup Language
- XLM: Extensible Markup Language (Langage de Balisage Extensible)
- URL:Uniform Resource Locator (Localisateur Uniforme de Ressource)
- HTPP : HyperText Transfer Protocol (Protocole de Transfert Hypertexte)
- XSS : Cross-Site Scripting
- DOM :Document Object Model
- XHTML : Extensible HyperText Markup Language
- API :Application Programming Interface (interface de programmation applicative)
- REST : Representational State Transfer
- BSD : Berkeley Software Distribution
- SAAS : Software As A Service (Logiciel en tant que service)
- WDI : Interface de composant de WinDev
- QT : Bibliothèque Logicielle Orientée Objet.
- OFX : Open Financial Exchange,
- QIF : Quicken Interchange Format (format d'échange de données financière)
- CMD : Interpréteur de Commandes (Interface en Ligne de Commande)
- DOS : Disk Operating System (Système d'Exploitation **PC-DOS**)
- JDK : Java Development Kit

INTRODUCTION GENERALE

1. Contexte de l'étude

Aujourd'hui, les données les plus pertinentes pour les journalistes se trouvent dans les sites web, les encyclopédies en ligne ou les réseaux sociaux. Or, la plupart de ces données sont non-structurées ou semi-structurées au meilleur des cas. Cette situation rend très complexe la tâche d'extraction via des méthodes classiques telles que l'extraction manuelle ou l'utilisation des requetés SQL des bases de données relationnelles.

Pour contourner ce problème, une des techniques utilisée aujourd'hui est l'extraction directe de données sur les pages web plus connue sous son nom anglais web scraping [1].Un scrapeur est un logiciel qui simule la navigation humaine sur le web pour collecter les données spécifiques provenant de différents sites. C'est un sous domaine de l'Extraction de l'information (EI).

Il permet d'extraire une connaissance à partir d'un texte par la transformation d'un format non structuré à un format structuré [1]. Le principe de base du Scraping est simple : transformer des informations non structurées présentes dans des pages web en données structurées facilement exploitables [2]. Il permet donc d'extraire des données structurées à partir de données non structurées [3]. Le Web Scraping est la technique d'automatisation de ce processus.

Au lieu de copier et coller manuellement les données des sites web, le logiciel web Scraping effectuera la même tâche dans une fraction de temps sur plusieurs milliers de pages sans effort. Le logiciel de Scraping va automatiquement charger et extraire les données de plusieurs pages de sites web en fonction de vos besoins. Ils sont en général conçus sur mesure pour un site web spécifique afin de récupérer un certain type d'information.

Une fois le processus terminé, il sera possible de récupérer la totalité des données extraites sous plusieurs formats, .json ou .csv... . Le Scraping peut-être très utile pour faire du retargeting, trouver les profils de candidat, faire une étude de marché ou faire une étude des prix sur les sites e-commerces.

Ainsi il sera plus aisé d'exploiter les données pour une analyse ou autre [4]. Une telle réalisation permet aux données web d'être accessible avec un effort humain réduit. Le web Scraping est évidement une tâche très complexe et très couteuse en temps et en ressources surtout lorsqu'elle est traitée manuellement.

2. Problématique

L'extraction de données à partir du web est un problème important qui a été étudié au moyen de différents outils scientifiques et dans une large gamme d'applications [5]. Les sites web actuels sont construits sur des infrastructures JavaScript qui facilitent l'utilisation de l'interface utilisateur, mais sont moins accessibles aux scrapeurs.

Le web Scraping, l'exploration web et toute autre forme d'extraction de données web peuvent être compliqués. Entre l'obtention de la page source correcte, l'analyse correcte de la source, le rendu de JavaScript et l'obtention des données sous une forme utilisable, il y a beaucoup de travail à faire [4]. Il existe plusieurs sites et programmes permettant de faire du web Scraping.

Ils se différencient par leurs utilisations, la façon d'extraire les informations sur le web. L'utilisation d'API étant probablement la meilleure façon d'extraire des données d'un site web. Presque tous les grands sites comme Twitter, Facebook, Google... fournissent les API pour accéder à leurs données de manière plus structurée [4]. De nombreuses approches d'extraction de données à partir du web ont été conçues pour résoudre des problèmes spécifiques [5, 6, 7].

D'autres approches réutilisent les techniques et les algorithmes développés dans le domaine d'extraction d'information comme les flux RSS mais ils sont limités dans leur utilisation. Les données affichées par la plupart des sites web ne peuvent être visualisées qu'avec un navigateur web. Ceux-ci n'offrent pas la fonctionnalité pour enregistrer une copie de ces données à des fins autres que visuelles. La seule option est de copier et coller manuellement les données.

Un travail très fastidieux qui peut prendre plusieurs jours voire semaines pour terminer suivant la taille des sites. Aujourd'hui, peu de travaux sont orientés dans le Scraping des données journalistiques web. La seule solution existante est préconfigurée pour la langue anglaise et il nécessite une implémentation pour l'adapter au français [8].

En plus, celle-ci se limite juste à la phase de collecte de l'article dans sa généralité laissant à la charge du journaliste, la tâche de prétraitement, de segmentation et de structuration des données. Il serait important de disposer d'un outil capable d'extraire les articles à partir des sources journalistiques et de segmenter chaque article en plusieurs affirmations.

3. Proposition

Nous proposons dans ce document, un moteur d'extraction automatique des faits journalistiques dans la presse sénégalaise en ligne. Notre moteur est composé de trois modules:

- ***Un module d'extraction d'articles:*** C'est un collecteur d'articles à partir des Url des sites journalistiques en ligne mis en entrée. Ce module utilise la librairie *Newspaper* [1] de Python pour réaliser l'extraction de tous les articles des sites web sélectionnés. Pour chaque article, le titre, les auteurs, la date de publication, les mots-clés, le texte et le résumé sont alors extraits et enregistrés dans un ensemble de dossiers en local ;
- ***Un module de fusion d'articles et d'affirmation :*** Ce module attaque, nettoie, segmente en affirmations, fusionne et stocke tous les articles et affirmations issus du module d'extraction ;
- ***Module de stockage :*** La partie de stockage est composée d'un ensemble de fichiers JSON et CSV constitués d'articles et d'affirmations.

4. Intérêt

La mise en place de tel extracteur sera un atout considérable pour les journalismes et l'ensemble des données récoltées pourront être réutilisés à des fins d'analyse et/ou de factchecking automatisé (vérification des faits). En effet, dans la pratique, le factchecking consiste à interroger plusieurs sources de données, analyser les résultats pour déterminer la véracité ou non d'un fait [9].

5. Annonce du plan

Ce document est organisé en deux grands chapitres hormis l'introduction générale et la conclusion. Dans l'introduction générale nous présentons le contexte de notre étude, la problématique, les objectifs, et l'annonce du plan du document.

Dans le premier chapitre nous présentons concepts clés du sujet et l'état de l'art du web Scraping. Dans le deuxième chapitre, nous présentons notre solution, son architecture et son mode de fonctionnement.

CHAPITRE 1 : CONTEXTE DE L'ETUDE ET ETAT DE L'ART

1.1. Clarification des concepts clés du sujet

1.1.1. Le web journalisme

Le ***journalisme en ligne*** est une forme de journalisme utilisant Internet comme principal support, par le biais notamment de versions électroniques de médias traditionnels, ou bien de journaux en ligne. Le web journalisme utilise des milliers de sources qui donnent des opinions simultanément sur un même sujet.

Il met à jour en temps réel des articles et des brèves. Il ne s'oppose pas au journalisme traditionnel, mais propose une autre manière de faire qui utilise des données numériques comme corpus. En effet, l'usage du « data » est un atout éditorial puissant pour le travail du journaliste [10].

Le journaliste Web travaille pour des médias numériques. Il alimente les sites Internet en contenus éditoriaux : textes, photos, vidéos, sons… Il peut aussi participer aux choix de la ligne éditoriale et à la rédaction de newsletters envoyées aux abonnés. De ce fait, il doit être polyvalent et extrêmement réactif. Car, sur un site contrairement à un journal traditionnel, il peut réactualiser un article après sa publication et choisir ses sujets en fonction des réactions des internautes.

1.1.2. Article de presse

En journalisme, un article est un texte qui relate un événement, présente des faits ou expose un point de vue. Il s'appuie pour cela sur différentes sources d'informations orales ou écrites (Selon Wikipédia). Objectif : Ecrire un article de presse nécessite de respecter certaines règles de présentation et d'écriture, ceci dans le but d'atteindre trois objectifs : délivrer une information claire et précise, éveiller la curiosité du lecteur et rechercher sa complicité. Vu sous l'angle informatique, un article est un fichier HTML. L'article de presse normal comporte :

- Un surtitre : Placé au-dessus de l'article, il s'agit d'une phrase qui permet de situer le cadre général de l'article ;
- Un titre : Il doit viser à l'efficacité et à la brièveté, c'est pourquoi on privilégie la nominalisation. Le titre peut être informatif (il ne cherche qu'à renseigner le lecteur) ou incitatif (il fait réagir le lecteur par un effet de surprise, le sourire ou l'intrigue).
- Un chapeau : Placé sous le titre, il résume l'essentiel de l'information présentée.

- Le corps de l'article : C'est-à-dire l'article en lui-même ; celui-ci suit un plan précis.

L'article de presse en ligne contient en plus de ces informations des images et/ou des vidéos.

Figure 1. Exemple : Exemple d'article en ligne [42]

1.1.3. Le web Scraping

Aujourd'hui, les données les plus pertinentes pour les journalistes se trouvent dans les sites web, les encyclopédies en ligne ou les réseaux sociaux. Le World Wide Web contient toutes sortes d'informations d'origines différentes; certains sont sociaux, financiers, sécuritaires et universitaires.

La plupart des gens accèdent à l'information par internet à des fins éducatives. Les informations sur le web sont disponibles dans différents formats et à travers différentes interfaces d'accès. Or, la plupart de ces données sont non-structurées ou semi-structurées au meilleur des cas. Cette situation rend très complexe la tâche d'extraction via des méthodes classiques telles que l'extraction manuelle ou l'utilisation des requetés SQL des bases de données relationnelles.

Par conséquent, l'indexation ou le traitement sémantique des données sur des sites web pourrait être fastidieux. Pour contourner ce problème, une des techniques utilisée aujourd'hui

est l'extraction directe de données sur les pages web plus connue sous son nom anglais web scraping. Le web Scraping est une technique permettant l'extraction des données d'un ou plusieurs sites web via un programme, un logiciel automatique ou un autre site [1].

L'objectif du web Scraping est donc d'extraire le contenu d'une page d'un site de façon structurée. Le scraping permet ainsi de pouvoir réutiliser ces données. Le principe de base du web scraping est simple : transformer des informations non structurées présentes dans des pages web en données structurées facilement exploitables.

En effet, un scrapeur est un logiciel qui simule la navigation humaine sur le web pour collecter les données spécifiées provenant de différents sites. Il s'agit donc d'attaquer une page web, d'extraire son contenu et de présenter le résultat suivant la structuration et le format souhaités. Le Scraping peut être effectuée à l'aide d'un logiciel, d'un outil ou d'une application. Mais elle peut aussi être réalisée par des développeurs.

Il existe diverses techniques de récupération web, notamment le copier-coller traditionnel, la capture de texte et la correspondance d'expression régulière, la programmation HTTP, l'analyse HTML, l'analyse DOM, le logiciel de conversion web, les plateformes d'agrégation verticale, la reconnaissance d'annotation sémantique et les analyseurs de pages web de vision informatique [17].

Le copier-coller traditionnel est la technique de base et fastidieuse de scraping de la toile où les utilisateurs doivent supprimer de nombreux jeux de données. Le logiciel de Scraping web est la technique de Scraping la plus simple, car toutes les autres techniques, à l'exception du copier-coller traditionnel, nécessitent une certaine expertise technique. Il existe aujourd'hui des centaines de logiciels de Scraping web, la plupart conçus en utilisant Java, Python et Ruby.

Il existe également des logiciels de scraping web open source et des logiciels commerciaux. Les outils de web Scraping permettent d'automatiser la collecte de données sur Internet. Pour la plupart, et bien que leur portée soit limitée, ils sont accessibles en libre-service, et peuvent être tout simplement utilisés depuis le navigateur de votre ordinateur (Chrome ou Firefox).

Si la taille des données dont vous avez besoin est limitée, ou que les sites web sont particulièrement faciles à atteindre (site statique…), alors ces addons présentent un rapport qualité-prix très intéressant, et peuvent constituer un choix judicieux [17].

Toutefois, si vous avez beaucoup de données à extraire, que vous souhaitez les extraire de façon périodique, ou que les sites web présentent des systèmes de défense robustes, alors ces outils gratuits s'adapteront mal à vos besoins.

Le coût, certes dérisoire, de ces outils sera toutefois vite contrebalancé par le temps et les efforts qui vous seront nécessaires afin d'implémenter, d'exécuter et de maintenir ces outils de web Scraping dans le temps. Dans de telles situations, la meilleure solution technique et économique reste de faire appel à un fournisseur de solution d'extraction clés en main [17].

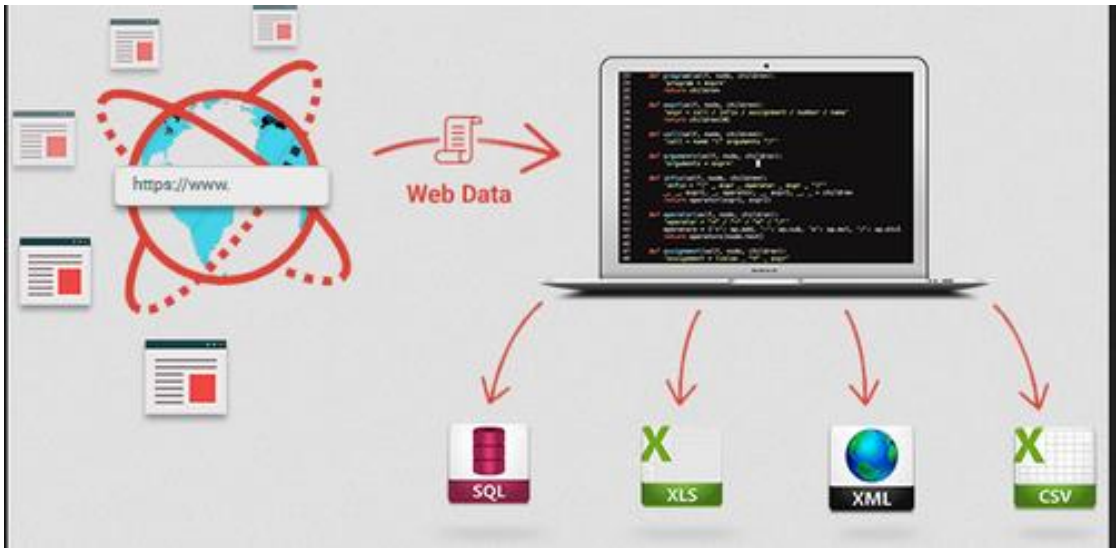


Figure 2. Web Scraping [43]

Les API et les infrastructures de Scraping web traitent les tâches les plus courantes. Les scrapeurs de données web interviennent pour atteindre des objectifs d'extraction particuliers.

L'accès au site: le scrapeur de données web établit la communication avec le site web cible via le protocole HTTP, un protocole internet basé sur du texte, qui coordonne les transactions demande-réponse entre un client, généralement un navigateur web, et un serveur web.

Dans HTTP, les 'méthodes' les plus fréquentes sont GET, utilisé dans les demandes de ressources, et POST, utilisé dans la soumission de formulaires et le téléchargement de fichiers.

L'en-tête «User-Agent» est également un en-tête de requête important, car le serveur le recherche pour savoir quel type de programme accède à son contenu le navigateur par rapport à un robot et éventuellement différencier les réponses des utilisateurs.

En outre, comme tout autre robot web, les scrapeurs de données web doivent être conformes aux «conditions d'utilisation» du site décrites dans son fichier «robots.txt» qui est un fichier hébergé sur le serveur, qui indique quelles ressources ne doivent pas être utilisées. Accessible par des procédures automatiques), et doit planifier les tâches de récupération avec soin pour éviter la surcharge du serveur [17].

L'analyse HTML et l'extraction du contenu: une fois le document HTML récupéré, le scraper web peut extraire le contenu qui vous intéresse. À cette fin, la correspondance d'expression régulière, seule ou en combinaison avec une logique supplémentaire, est largement adoptée.

En guise d'alternative, il existe des bibliothèques d'analyse HTML qui travaillant sur la structure de modèle d'objet de document des pages web et les langages basés sur les sélecteurs, tels que XPath et la syntaxe du sélecteur CSS.

En règle générale, il est recommandé de garder les expressions de correspondance aussi générales que possible afin de rendre les robots moins vulnérables lors des modifications du document HTML. Le scrapeur est plus robuste lorsque le site implémente un balisage sémantique [17].

La création de la sortie: L'objectif principal est de transformer le contenu extrait en une représentation structurée appropriée pour une analyse et un stockage ultérieurs. Bien que cette dernière étape soit marginale par rapport au scraping web, certains outils sont conscients du post-traitement des résultats, fournissant des structures de données en mémoire et des solutions textuelles, telles que les chaînes ou les fichiers généralement les fichiers XML ou CSV.

Les approches existantes pour mettre en œuvre un scrapeur de données web peuvent être structurées en trois catégories principales: les bibliothèques pour les langages de programmation à usage général, les Framework et les environnements de bureau [17].

Les Bibliothèques

Habituellement, les bibliothèques tierces accordent l'accès au site en implémentant le côté client du protocole HTTP, tandis que le contenu récupéré est analysé à l'aide de fonctions de chaîne intégrées, telles que la correspondance d'expression régulière, la génération de jetons et le découpage.

Les packages tiers peuvent également fournir une analyse plus sophistiquée, telle que la construction d'arborescence HTML et la correspondance XPath. Il prend en charge les principales fonctionnalités du protocole HTTP, notamment les certificats SSL, HTTP POST, HTTP PUT, le téléchargement FTP, le téléchargement basé sur un formulaire HTTP, les procurations, les cookies et l'authentification HTTP.

En plus, il possède les liaisons utiles avec de nombreux langages de programmation [17]. En Java, le package Apache HTTP Client émule les principales fonctionnalités HTTP, à savoir toutes les méthodes de requête, les cookies, l'authentification SSL et HTTP, et peut être combiné avec des bibliothèques d'analyse HTML telles que JSOUP.

Java prend également en charge XPath et fournit plusieurs bibliothèques de nettoyage HTML, telles que HtmlCleaner. De même, BeautifulSoup est une bibliothèque d'analyse HTML Python, qui peut être combinée à la prise en charge native de la langue pour les connexions HTTP [17]. En plus, dans les environnements de type Unix, en programmant simplement les programmes de ligne de commande du système d'exploitation à l'aide de scripts Shell, les programmeurs peuvent créer des scrapeurs de données web en une ou plusieurs lignes de code.

Les programmes tels que Curl (lib url) et Wget implémentent la couche client HTTP, tandis que des utilitaires tels que Grep, Awk Sed et couper-coller peuvent être utilisés pour analyser et transformer facilement le contenu. Dans le cas de robots côté serveur, fonctionnant généralement dans les applications web, une technologie 100% compatible avec le langage de programmation généralement PHP ou Java est recommandée [17].

Les Frameworks

L'utilisation d'un langage généraliste pour créer des robots présente certains inconvénients. Plusieurs bibliothèques doivent souvent être intégrées, notamment une pour l'accès web et les autres pour analyser et extraire le contenu des documents HTML. De plus, les robots sont connus pour être des logiciels faibles, qui sont considérablement affectés par les modifications du code HTML des ressources consultées, et qui nécessitent donc une maintenance continue.

Dans les langages compilés, tels que Java, tout changement dans la mise en œuvre du robot oblige à la recompilation, voire au redéploiement de l'ensemble de l'application [17]. Les cadres de Scraping constituent une solution plus intégrative. Par exemple, Scrapy est un puissant Framework de Scraping web pour Python, dans lequel les robots sont définis comme

des classes héritant de la classe base Spider, qui définit un ensemble d'URL de démarrage et une fonction d'analyse appelée à chaque itération web. Les pages web sont automatiquement analysées et le contenu Web extrait à l'aide d'expressions XPath [17].

D'autres cadres présentent des langages spécifiques à un domaine (DSL), qui est les langages de programmation spécifiques conçus pour un domaine particulier et, par conséquent, les robots sont traités comme des artefacts indépendants et externes. Le Web-Harvest est un framework de récupération de données web pour Java, en est un exemple.

Ici, les processus d'extraction web sont décrits en XML à l'aide d'un environnement visuel et sont composés de plusieurs "pipelines", qui peuvent inclure des instructions de procédure, telles que les définitions de variables et les boucles, ainsi que de nombreuses primitives.

Telles que "http pour récupérer le contenu web,' html-to-XML pour nettoyer le HTML et' XPath pour extraire le contenu .Jarvest est un autre exemple d'infrastructure de récupération de données web Java qui définit également un DSL, mais utilise JRuby pour une implémentation de robot plus compacte [17].

Nous avons sélectionné plusieurs packages de Scraping web disponibles destinés aux programmeurs. Il existe six bibliothèques implémentant un client HTTP (C) et / ou une analyse syntaxique HTML (P) et trois Framework (F). Les Framework Web-Harvest et Jarvest présentent un langage spécifique à un domaine pour la définition de robots, basé sur XML et Ruby, respectivement.

Pour toutes les alternatives analysées, nous rapportons leurs fonctions d'extraction, notamment les expressions régulières (R), l'arbre analysé en HTML (H), les expressions XPath (X) et les sélecteurs CSS (C) [17].

	Type	Langue spécifique au domaine	API / autonome	La langue	Installations d'extraction
	C: client HTTP				R: expressions régulières
	P: analyse				H: arbre analysé HTML
	F: cadre				X: XPath
					C: sélecteurs CSS
Unix shell (curl / wget, grep, sed, couper, coller, awk)	CP	Non	SA	frapper	R
Curl / libcurl	C	Non	Tous les deux	Fixations C +	
Web-Harvest	F	Oui	Tous les deux	Java	RX
Jsoup	CP	Non	API	Java	HC
HttpClient	C	Non	API	Java	
JARVEST	F	Oui	Tous les deux	JRuby / Java	RXC
WWW :: Mécaniser	CP	Non	API	Perl	RX
Scrapy	F	Non	Tous les deux	Python	RX
BeautifulSoup	P	Non	Non	Python	H

Figure 3. Bibliothèques et Framework de Scraping Web Open-source [17]

Les Environnements de bureau

Les applications de bureau répondent aux besoins des programmeurs profanes. Ce type d'outils est doté d'environnements de conception graphique facilitant la création et la

maintenance de robots. Habituellement, le logiciel comprend un navigateur intégré dans lequel l'utilisateur peut naviguer vers le web cible et sélectionner de manière interactive les éléments de la page à extraire, en évitant toute spécification d'expressions régulières, de requêtes XPath ou d'autres aspects techniques.

En outre, les modules sont disponibles pour créer plusieurs types de sortie, tels que les fichiers au format CSV, Excel et XML et les insertions dans des bases de données [17].

Les principaux inconvénients des solutions de bureau sont la distribution commerciale et l'accès limité aux API, qui rendent difficile l'intégration de ces Scraping dans d'autres programmes ce qui est souvent une exigence. Nous comparons sept outils de récupération web courants basés sur des postes de travail.

Comparaison des fonctionnalités de différentes solutions de Scraping sur le bureau. Nous avons sélectionné plusieurs fonctionnalités pour évaluer les outils, notamment les licences logicielles, les plates-formes prises en charge, les capacités d'accès au site, les aspects liés à l'exécution et les possibilités de conception de robots [17].

	IrobotSoft [a]	Visual Web Ripper [b]	Débutant [c]	Mozenda [d]	Grattoir [e]	WebSu [f]
Type de logiciel						
Licence	Gratuiciel	Commercial	Commercial	Commercial	Freeware (édition de base)	Comm
Open source	Non	Non	Non	Non	Non	Non
Plateformes	Gagner	Gagner	Gagner	Gagner	Gagner Linux Mac	Gagne

Acces au site						
Formulaire POST	Oui	Oui	Oui	Oui	Oui	Non
Session	Oui	Oui	Oui	Oui	Oui	Non
Conf. agent utilisateur	C'EST À DIRE	IE et 2 UA internes	C'EST À DIRE Firefox	C'EST À DIRE	Non	Firefox UA int
Itération sur les pages	Oui	Oui	Oui	Oui	Oui	Oui
Proxys d'Anonymizer	Oui	Oui	N / A	Non	Non	N / A
Les formats						
Formats d'entrée	.irb	.déchirure	Liste des URL des pages Web	.xml	.sss	.zws
Formats de sortie	Texte	CSV	Texte	CSV	Texte	CSV
	CSV	XML	Exceller	TSV	CSV	XML
	XML	DB	DB	XML	DB	Excelle
	DB	Exceller		Exceller		
Format de fichier robot	.irb	.déchirure	scripts .nbs	.xml	.sss	.zws

Runtime						
Multi-threading	Oui	Oui	Oui	Non	Oui	Oui
Résultats progressifs	Oui	Non	Oui	Oui	Oui	Oui
Environnement de conception						
Concepteur graphique	Oui	Oui	Oui	Oui	Oui	Oui
Accès API	Non	Oui	Oui	Oui	Oui	Oui
Scriptable	Oui	Limité	Oui	Non	Oui	N / A

Figure 4. Solutions de raclage Web sur le bureau [17]

1.1.4. Python

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. C'est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet.

Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications son typage est dynamique et il est interprété, dans de nombreux domaines et sur la plupart des plateformes. L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le site Internet et peuvent être librement redistribués.

Ce même site distribue et pointe vers des modules, des programmes et des outils tiers. Enfin, il constitue une source de documentation. L'interpréteur Python peut être facilement étendu par de nouvelles fonctions et types de données implémentés en C ou C++. Python est également adapté comme langage d'extension pour personnaliser des applications.

1.1.5. JAVA

Java est un langage de programmation et une plate-forme informatique rapide, sécurisé et fiable. Il est polyvalent, simultané, basé sur les classes, orienté objet [11] et spécifiquement conçu pour avoir le moins possible de dépendances de mise en œuvre. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts.

Avec une syntaxe similaire à C et C ++, Il permet aux développeurs d'applications "d'écrire une fois, d'exécuter n'importe où" (WORA) [12], ce qui signifie que le code Java compilé peut être exécuté sur toutes les plates-formes prenant en charge Java sans nécessiter de recompilation [*13*]. Les applications Java sont généralement compilées en "code intermédiaire" pouvant s'exécuter sur n'importe quelle machine virtuelle Java (JVM), quelle que soit l'architecture informatique sous-jacente.

1.1.6. Le format JSON

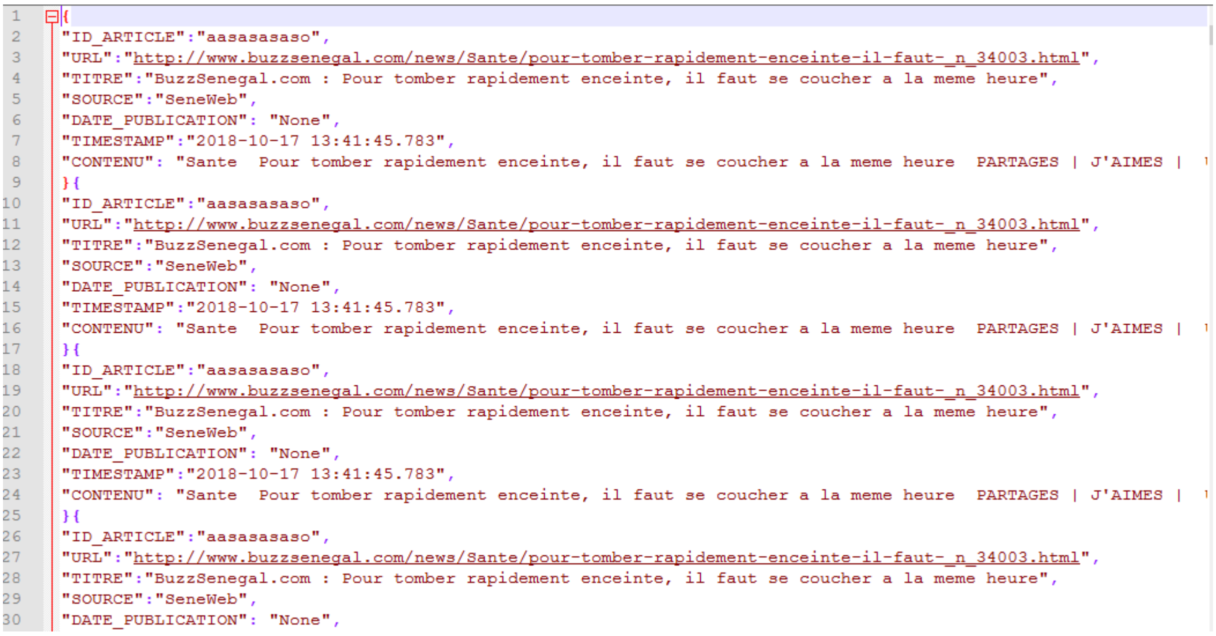
JavaScript Object Notation (JSON) est un format de données textuelles dérivé de la notation des objets du langage JavaScript. JSON (JavaScript Object Notation) est un format d'échange de données léger. Il est facile pour les humains de lire et d'écrire. Il est facile pour les machines d'analyser et de générer.

C'est format de texte totalement indépendant du langage mais utilisant des conventions bien connues des programmeurs de la famille de langages C, y compris le C., C ++, C #, Java, JavaScript, Perl, Python et bien d'autres. Ces propriétés font de JSON un langage d'échange de données idéal. Ainsi, des bibliothèques pour le format JSON existent dans la plupart des langages de programmation, un document JSON. JSON est construit sur deux structures:

- Une collection de paires nom / valeur : Dans différentes langues, ceci est réalisé sous forme d'objet, d'enregistrement, de structure, de dictionnaire, de table de hachage, de liste à clé ou de tableau associatif ;
- Une liste ordonnée de valeurs : Dans la plupart des langues, cela est réalisé sous forme de tableau, de vecteur, de liste ou de séquence.

Ce sont des structures de données universelles. Pratiquement tous les langages de programmation modernes les prennent en charge sous une forme ou une autre. Il est logique qu'un format de données interchangeable avec les langages de programmation soit également basé sur ces structures. Voici les formes que nous retrouvons en JSON :

- Un objet : C'est un ensemble non ordonné de paires nom / valeur. Un objet commence par {(accolade gauche) et se termine par} (accolade droite). Chaque nom est suivi de: (deux points) et les pairs noms / valeur sont séparées par, (virgule) ;
- Un tableau : C'est une collection ordonnée de valeurs. Un tableau commence par [(crochet gauche) et se termine par] (crochet droit). Les valeurs sont séparées par, (virgule) ;
- Une valeur : Elle peut être une chaîne entre guillemets, un nombre, ou true, false ou NULL, un objet ou un tableau. Ces structures peuvent être imbriquées. Une chaîne est une séquence de zéro ou plusieurs caractères Unicode, entourés de guillemets doubles, à l'aide d'échappements de barre oblique inversée. Un caractère est représenté par une chaîne de caractères unique. Une chaîne ressemble beaucoup à une chaîne C ou Java.



```
{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la meme heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"TIMESTAMP":"2018-10-17 13:41:45.783",
"CONTENU": "Sante  Pour tomber rapidement enceinte, il faut se coucher a la meme heure  PARTAGES | J'AIMES |
}{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la meme heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"TIMESTAMP":"2018-10-17 13:41:45.783",
"CONTENU": "Sante  Pour tomber rapidement enceinte, il faut se coucher a la meme heure  PARTAGES | J'AIMES |
}{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la meme heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"TIMESTAMP":"2018-10-17 13:41:45.783",
"CONTENU": "Sante  Pour tomber rapidement enceinte, il faut se coucher a la meme heure  PARTAGES | J'AIMES |
}{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la meme heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
```


Figure 5. Exemple de fichier JSON

1.1.7. Le format CSV

CSV est un format de fichier simple utilisé pour stocker des données tabulaires, telles que des feuilles de calcul ou des bases de données. Les fichiers au format CSV peuvent être importés et exportés à partir de programmes stockant des données dans des tableaux

Un fichier CSV est un fichier de valeurs, séparé par des séparateurs (virgules ou points virgules). Il contient des ensembles de données en texte brut séparés par des séparateurs, chaque nouvelle ligne du fichier CSV représentant une nouvelle information ou une nouvelle

ligne de la base de données en cas d'export vers un SGBDR (Système de gestion de base de données relationnelle). Les fichiers CSV sont couramment utilisés pour transférer des données d'une base de données ou d'un format de tableur à un autre. Cela permet aux personnes qui n'exécutent pas les mêmes applications de base de données de partager des fichiers de base de données entre elles.

1.2. Etat de l'art sur le web Scraping

Plusieurs outils de web Scraping sont proposés dans la littérature [7, 5, 6]. Nous distinguons cependant deux grandes catégories : les logiciels prêts à l'emploi et les librairies permettant de développer des scrapers [7].

1.2.1. Modules des langages de programmation

Cette catégorie regroupe les librairies des langages de programmation Java [13], PHP [14] ou Python [15]. Dans les deux cas, ces solutions sont destinées aux spécialistes en extraction d'information. Parmi ces modules, nous nous attardons sur quelques modules Python et JAVA [16].

1.2.1.1. NewsPaper

Newspaper est une librairie Python qui permet d'extraire automatiquement des articles de presse en ligne, et de sélectionner ou d'en trier les contenus les plus pertinents, sans avoir à parcourir de très nombreuses sources d'informations. Il utilise des algorithmes avancés de web scraping afin d'extraite du texte à partir d'un site web.

Il fonctionne très bien au niveau des sites de presse en ligne. [7] Et constitue récemment l'un des plus performants extracteurs d'articles. [8]. Il fonctionne sur plusieurs langues notamment : l'anglais, le français, le chinois, l'arabe, l'italien, le grec …Parmi ses principales fonctionnalités nous pouvons citer :

- Le modèle multithread d'extraction de plusieurs articles ;
- L'identification d'Url ;
- L'extraction de texte à partir de html ;
- L'extraction de l'image principale à partir de html ;
- L'extraction d'images à partir de html ;
- L'extraction de mots- clés à partir du texte ;
- L'extraction du résumé à partir du texte ;
- L'extraction des auteurs à partir du texte ;

- Le NLP (Traitement automatique de la langue).

1.2.1.2. Beautiful Soup

Beautiful Soup est une bibliothèque Python permettant d'extraire des données de fichiers HTML et XML. Il travaille avec votre analyseur préféré pour fournir des moyens idiomatiques de naviguer, de rechercher et de modifier l'arbre d'analyse. Cela permet généralement aux programmeurs d'économiser des heures ou des jours de travail.

1.2.1.3. Jaunt

Jaunt est une bibliothèque Java pour le web scraping, l'automatisation web et les requêtes JSON. La bibliothèque fournit un navigateur rapide, ultraléger sans tête sans interface graphique. Le navigateur fournit une fonctionnalité de récupération web, un accès au DOM et un contrôle sur chaque demande de réponse HTTP. Jaunt permet à un programme Java de:

- Peform web-scraping et extraction de données JSON,
- DE travailler avec des formulaires et des tables, contrôler, traiter les requêtes et les réponses HTTP individuelles,
- D'interfacer avec des API REST ou des applications Web (JSON, HTML, XHTML ou XML).

1.2.1.4. Jsoup

Jsoup est une bibliothèque de Java open source distribué sous la licence libérale MIT qui permet de travailler avec du HTML réel. Il fournit une API très pratique pour extraire et manipuler des données, en utilisant le meilleur des méthodes DOM, CSS et jQuery. Jsoup implémente la spécification HTML et analyse HTML dans le même DOM que les navigateurs modernes.

Il scrape et analyse le code HTML d'une URL, d'un fichier ou d'une chaîne rechercher pour en extraire des données à l'aide d'un sélecteurs CSS ou de parcours le DOM pour manipuler les éléments HTML. Les attributs et le texte nettoient le contenu soumis par l'utilisateur par rapport à une liste blanche sécurisée, afin de prévenir les attaques XSS sortie ordonnée HTML. Il est possible de récupérer la page d'accueil de Wikipédia, analysez-la dans un DOM et sélectionnez les titres de la section dans les nouvelles liste.

1.2.2. Les Outils prêt à l'emploi

Cette catégorie regroupe les outils automatiques préconfigurés et prêts à l'emploi. Ces outils sont plus destinés aux profanes qui chercheraient à extraire des données d'une ou plusieurs pages web sans aucune connaissance préalable en RI (Recherche d'informations). Les outils de cette catégorie sont divisés en trois groupes.

1.2.2.1. Les navigateurs web

Lorsqu'on visualise une page sur un navigateur Google Chrome ou Mozilla, il suffit juste d'effectuer un simple « copier-coller » pour extraire le contenu de la page. Cependant, cette méthode simpliste est assez fastidieuse et peu efficace lorsque nous souhaitons effectuer une extraction de données sur plusieurs pages.

1.2.2.2. Les extensions des navigateurs web

Elles se présentent sous la forme d'Addons (modules supplémentaires) pour les navigateurs. Elles ont pour rôle de faciliter et d'automatiser le processus d'extraction. Nous avons par exemple Scraper et Web Scraper pour le navigateur Google Chrome, Data Scraper et Cloump U-Scraper Plugin pour le navigateur Firefox etc.

1.2.2.3. Les logiciels et des plateformes

Ceux sont des outils implémentés expressément pour le scraping de données web. Leurs puissances et leurs performances dépendent par contre d'un bon paramétrage. Nous pouvons citer les outils en ligne weboob, Import.io, Scrapy, PhantomJS, OutWitHub et Newspaper.

1.2.2.3.1. Scrapy

Scrapy est un framework open-source développé en Python permettant la création de robots d'indexation. Il dispose d'une forte communauté et offre de nombreux modules supplémentaires. Il est simple, aucune notion avancée en Python n'est nécessaire pour utiliser Scrapy, pour la productif l'empreinte du code à générer est très courte et la plupart des opérations sont gérées par Scrapy.

Le framework est rapide notamment avec une gestion d'actions en parallèle, chaque robot peut être personnalisé via des extensions, modifiant son comportement. Les robots Scrapy sont compatibles avec Linux, Windows, Mac et BSD robuste, grâce à une batterie de tests effectués par les développeurs et par la communauté.

1.2.2.3.2. Import.IO

Import.io est une plate-forme d'intégration de données Web SaaS (WDI). Il permet aux utilisateurs de convertir les données web non structurées en un format structuré. Les données Web sont destinées à la consommation sur les plates-formes analytiques ou utilisées dans des applications commerciales ou marketing.

Import.io fournit un environnement visuel pour automatiser le flux de travail l'extraction et de la transformation des données web. Il est possible de spécifier l'***url*** d'un site web cible le module d'extraction de données web vous fournira un environnement visuel pour la conception de flux de travaux automatisés de collecte de vos données.

Une fois extrait, le logiciel fournit des fonctionnalités complètes de préparation des données utilisées pour harmoniser et nettoyer les données. Une bibliothèque de fonctions de type tableur permettant à l'utilisateur final de créer des formules personnalisées pouvant être utilisées pour enrichir l'ensemble des données.

Import.io fournit plusieurs options pour consommer les résultats. Il possède son propre module de visualisation et de tableau de bord pour aider les analystes commerciaux à obtenir les informations dont ils ont besoin. Il fournit également des API offrant un accès complet à tout ce qui peut être fait sur leur plate-forme, ce qui permet d'intégrer les données web directement dans leurs propres applications.

1.2.2.3.3. OutWit Hub

OutWitHub est un logiciel de scraping web, conçu pour extraire et collecter automatiquement des informations à partir de ressources en ligne ou locales. Le programme reconnaît et récolte des liens, des images, des documents, des contacts, des mots et groupes de mots récurrents et les flux RSS. Il convertit les données structurées ou non en tables formatées qui peuvent être exportées vers des feuilles de calcul ou des bases de données.

Le programme comprend un navigateur web et un panneau latéral qui donne accès à un certain nombre de vues contenant les données provenant d'extracteurs prédéfinis. Les pages web et les documents textuels sont décomposés en différents constituants et présentés sous forme de tables dans ces vues. L'application peut parcourir automatiquement des séries de liens ou des séquences de pages de résultats de moteurs de recherche pour extraire les éléments d'information recherchés, les organiser en tables et les exporter dans différents formats.

1.2.2.3.4. Weboob

Weboob (*WEB Outside Of Browsers*) est un ensemble d'applications (QBooblyrics, QBoobMsg, QCineoob, QCookboob, QFlatBoob , QHandjoob, QHaveDate, QVideoob, QWebContentEdit, weboob-config-Qt) écrit en Python et disponible sur la plupart des distributions GNU/Linux , MacOs et Windows. L'objectif est d'agréger et d'interagir avec des sites web via des interfaces unifiées. Applications graphiques utilisant la boîte à outils Qt.

Les applications natives permettent d'exporter les données dans divers formats, comme le CSV ou JSON, mais aussi dans des formats spécialisés standards comme OFX et QIF pour les sites bancaires, RSS pour les nouveaux messages ou encore *.kreml* pour les recettes de cuisine. Ainsi, les données exportées peuvent ensuite être importées dans d'autres applications agnostiques de Weboob.

1.2.2.3.5. PhantomJS

PhantomJS est un navigateur web sans interface graphique utilisé pour automatiser des interactions avec des pages web. Il fonctionne sous Windows, MacOs, Linux et Free BSD. PhantomJS JavaScript offre une interface de programmation permettant la navigation automatisée, la capture d'écran, de simuler les comportements utilisateurs, et l'utilisation d'assertions.

Ces possibilités en font un outil de choix pour exécuter des tests fonctionnels au sein d'un environnement d'intégration. PhantomJS est basé sur Webkit, ce qui en fait un environnement de navigation similaire à Safari ou à Google Chrome. Il est un programme open source distribué sous licence () BSD5.

CHAPITRE 2 : NOTRE PROPOSITION

2.1. Présentation du moteur

Nous présentons dans ce chapitre un collecteur d'articles. Il est constitué de :

2.1.1. Un module d'extraction d'articles

C'est un extracteur d'articles à partir des Url des sites journalistiques en ligne mis en entrée. Ce module utilise la librairie *Newspaper* [1] de Python pour réaliser l'extraction de tous les articles des sites web sélectionnés.

Mais vu que Newspaper a été développé et testé en Anglais alors que nous travaillons avec un corpus français dans le cadre de notre projet, il a fallu reconfigurer l'outil pour un meilleur traitement de la langue française. Notre amélioration s'est portée à plusieurs niveaux.

D'abord sur ***l'intégration de la langue française*** puis le changement des formats des dates, ***le filtrage des articles pour avoir les plus pertinents (éliminations des annonces par exemple)*** et enfin sur l'implémentation d'une fonction supplémentaire de ***détection automatique de la langue de l'article par l'utilisation des Stopwords***.

Les *StopWords* sont des mots vides dans une langue. Ce sont les mots les fréquents dans un texte mais qui ne permettent pas de le distinguer par rapport à d'autres. Pour le Français, ceux-ci sont composée principalement des déterminants (le, la, des, un ...), des adverbes de toute sorte (très, beaucoup, visiblement, ...).

Ainsi, la présence massive de ces *StopWords* nous informe sur la langue utilisée. Nous avons aussi travaillé sur **la gestion de l'encodage afin de permettre à Newspaper de prendre en compte les caractères accentués de la langue française en utilisant la librairie Python *Unicode*.**

Nous l'avons mis en application sur ***30 sites web*** de journalistique sénégalais soit une augmentation de 10 sites par rapport à la version précédente. Il s'agit entre autres : Seneweb, Jeune Afrique, Rewmi, Walfadjri, Igfm, Le quotidien, ActuSen Senego, APS, WiwsSport, Sen360, SeneNews, DakarActu, SenegalDirect et Enquête,…

Une fois toutes ces améliorations effectuées, nous stockons les articles par groupe selon le média. A la fin, les articles issus de Seneweb par exemple sont stockés dans le dossier Seneweb, ceux de Walfadjri dans un dossier Walfadjri etc.

2.1.2. Un module de fusion d'articles et d'affirmation

Ce module attaque, nettoie, segmente en affirmations, fusionne et stocke tous les articles et affirmations issus du module d'extraction. En effet, malgré sa puissance, ***Newspaper*** n'est pas doté de modules de ***segmentation des articles en propos (affirmations), d'identification des commentaires et des phrases vides (courtes).***

Interconnecté à l'extracteur, ce module parcourt et fusionne les articles issus de la phase d'extraction. C'est une API Java implémentée expressément pour les trois taches suivantes : le prétraitement, la segmentation et la fusion.

- **La tâche de prétraitement :** Elle consiste à effectuer des opérations de nettoyage et d'élimination des textes courts tels que les commentaires et les annonces. Ces articles sont très faciles à détecter dans le texte. Ils contiennent un nombre de caractères très en dessous des autres articles du corpus et ont le plus souvent moins de 400 caractères.

- **La tâche de segmentation :** Elle consiste à parcourir le texte de l'article et à le découper en phrases. Nous faisons alors une opération de *Tokenizer* avec comme délimiteur le point «. ». Mais vu que ce dernier n'est pas le seul séparateur de phrase dans la langue française, nous avons au préalable remplacé les autres séparateurs en point. Nous faisons ici allusion au point d'interrogation « ? » et au point d'exclamation « ! ».

- **La tâche de fusion :** Elle consiste à regrouper les articles des medias dans un seul fichier *Sn_Articles_Datasets.JSON* et les affirmations (phrases) dans le fichier *Sn_Affirmation_Datasets.JSON*.

2.1.3. Module de stockage

La partie de stockage est composée d'un ensemble de fichiers JSON et CSV constitués d'articles et d'affirmations.

2.2. Outils et Technologies utilisés

2.2.1. Installation de JAVA

2.2.2. Installation de Python et ses bibliothèques

2.2.2.1. Installation de Python

Python est téléchargeable au site officiel (Download). L'installation est simple. Nous l'avons effectué sous Windows avec la version 3 64 bits.

Figure 6. Interface d'installation Python 3.5[3]

N.B : N'oubliez pas de cocher ***AddPython3.5*to PATH** afin pouvoir utiliser python via CMD de DOS. Une fois l'installation terminée, nous pouvons la tester Via CMD en tapant : python + Enter

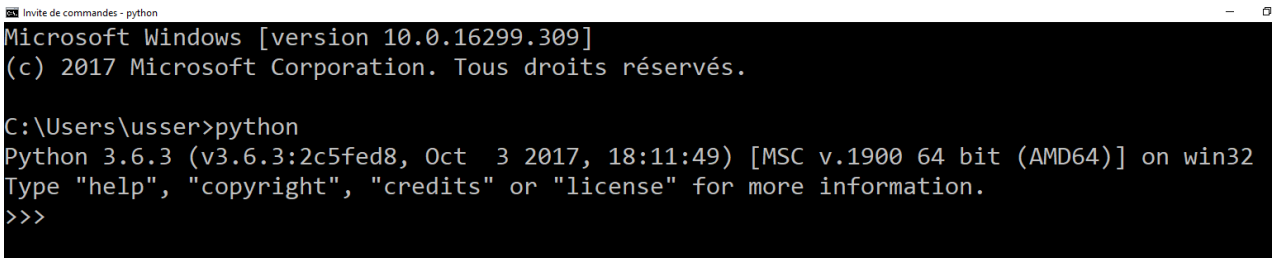


Figure 7. Test de Python

2.2.2.2. Installation de NewsPaper

Il est possible d'installer Newspaper directement en ligne de commande en utilisant pip de python. Pour cela, taper la commande suivante.

Pip3 install newspaper3k

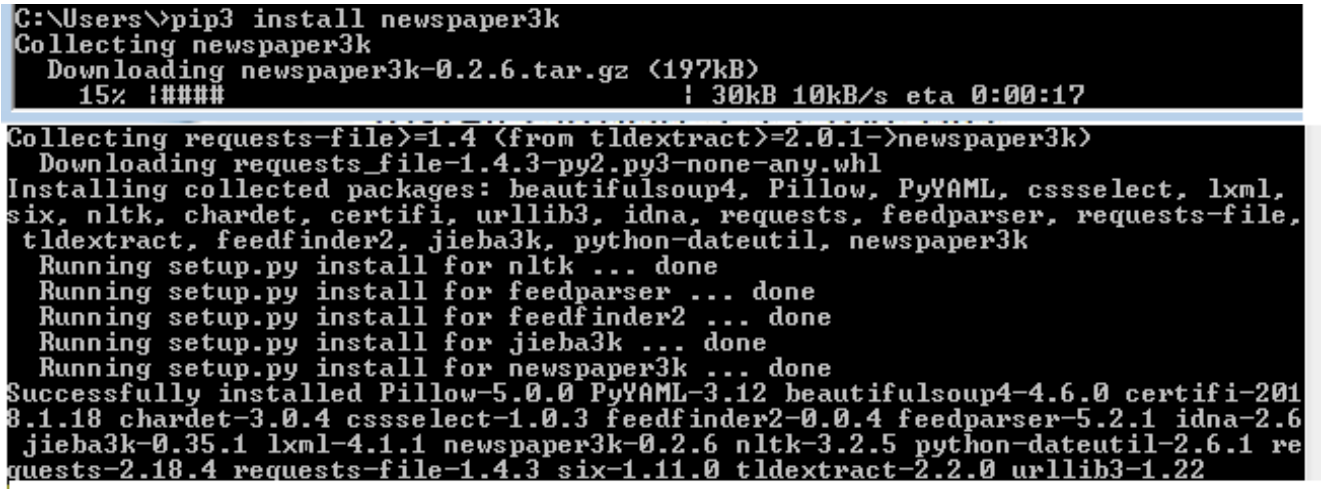


Figure 8. Installation Newspaper pour Python 3

- Installation du module de Traitement de la langue NLTK

```
>>>Import nltk

>>>Nltk. Download ()
```

- Installation du module UNICODE

```
Pip install Unicode
```

2.2.3. Installation du JDK 8 Java

Pour Télécharger le JDK8 Java il faut aller sur le site d'Oracle. Cliquer sur DOWNLOAD JDK. Vous arrivez sur la page représentée à la figure suivante.

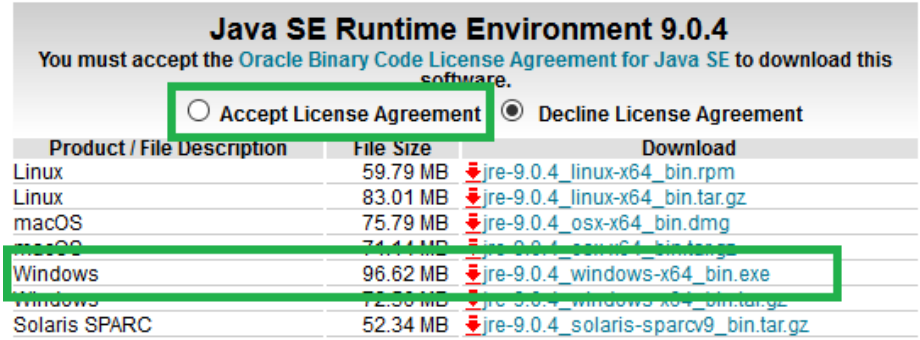


Figure 9. Interface 2 installation environnement JAVA [5]

Cochez la case : ***Accept License Agreement***, puis cliquez sur le lien correspondant à votre système d'exploitation (x86 pour un système 32 bits et x64 pour un système 64 bits). Une fenêtre pop-up de téléchargement doit alors apparaître.

2.3. Mode d'utilisation

Le moteur d'extraction combine deux programmes codés en deux langages de programmation distincts : Java et Python. Il se présente sous la forme de fichiers *jar* Java. Une fois à votre disposition, le dossier exécution est constitué de :

- Le fichier NewspaperJSON.py : C'est l'extracteur basé sur Newspaper de Python (Voir Annexe 1)
- Auto.bat : un fichier exécutable dans laquelle la commande d'exécution est édité.

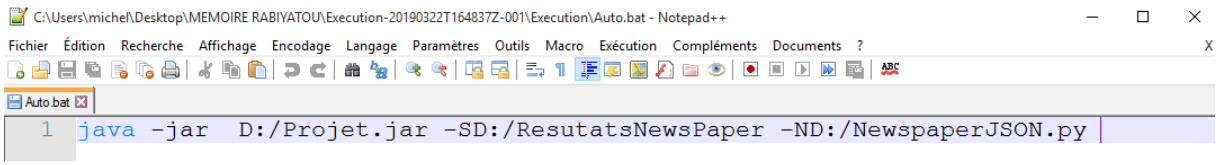


Figure 10. Fichier Auto.bat

- Projet.jar : C'est le fusionner. Le code JAVA est disponible en ANNEXE 2, ANNEXE 3 et ANNEXE 4. C'est un JAR Java qui prend deux paramètres :
  - **-SD:/Résultats** : la destination des résultats
  - **-ND: /NewspaperJSON.py** : le chemin du fichier python (extracteur)

La commande d'exécution donne :

```
Java -jar  D:/Projet.jar -SD:/ResutatsNewsPaper -ND:/NewspaperJSON.py
```

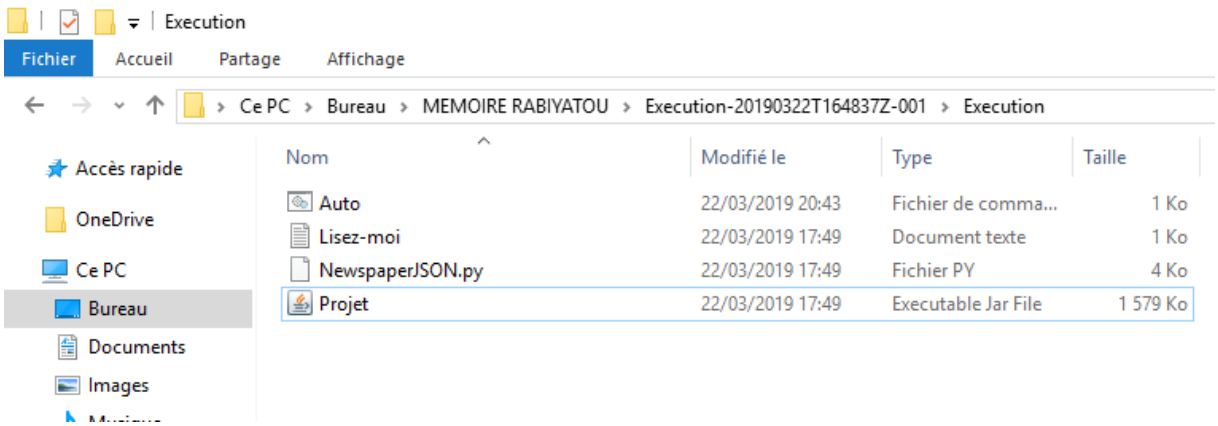


Figure 11. Dossier exécution

2.4. Full extraction et Incrémental Extraction

```
def FactExtract(src,cheminTexte):
      s = newspaper. Build(src,memoize_articles=True)
      tabURL=[]
      LesFaits=''
```

Tableau 1. Mémorisation de l'historique des url

2.5. Automatisation du processus

Le serveur étant sous Windows, nous avons alors automatisé l'exécution sous forme de deux tâches planifiées, une à 00h et l'autre à 12h. Les étapes de la configuration sont détaillées dans la suite de captures suivante.

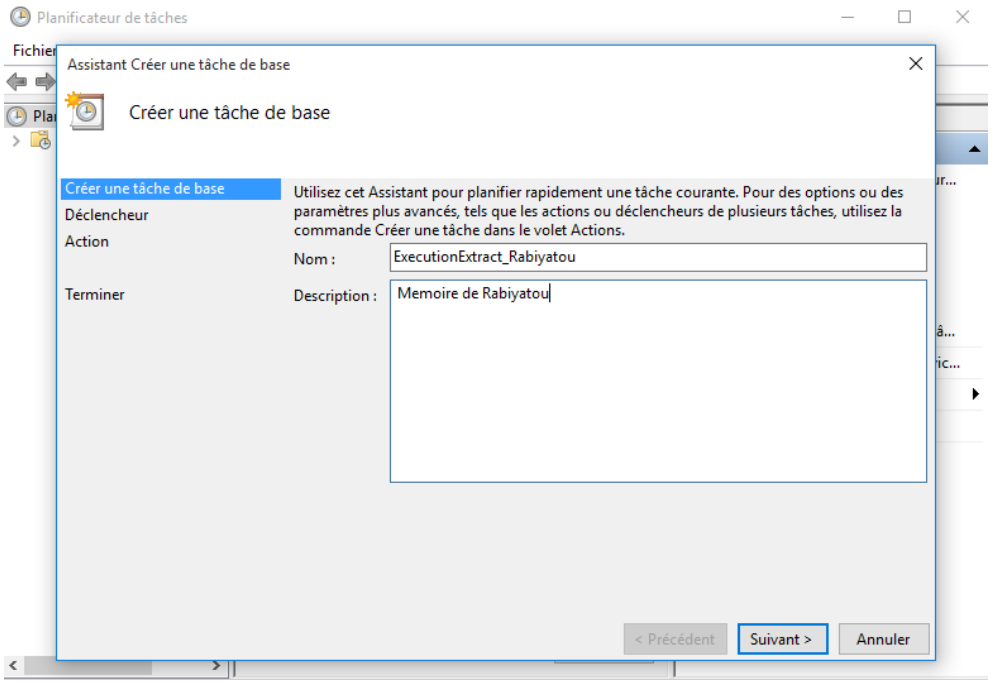


Figure 12. Création de la tâche planifiée de 00h.

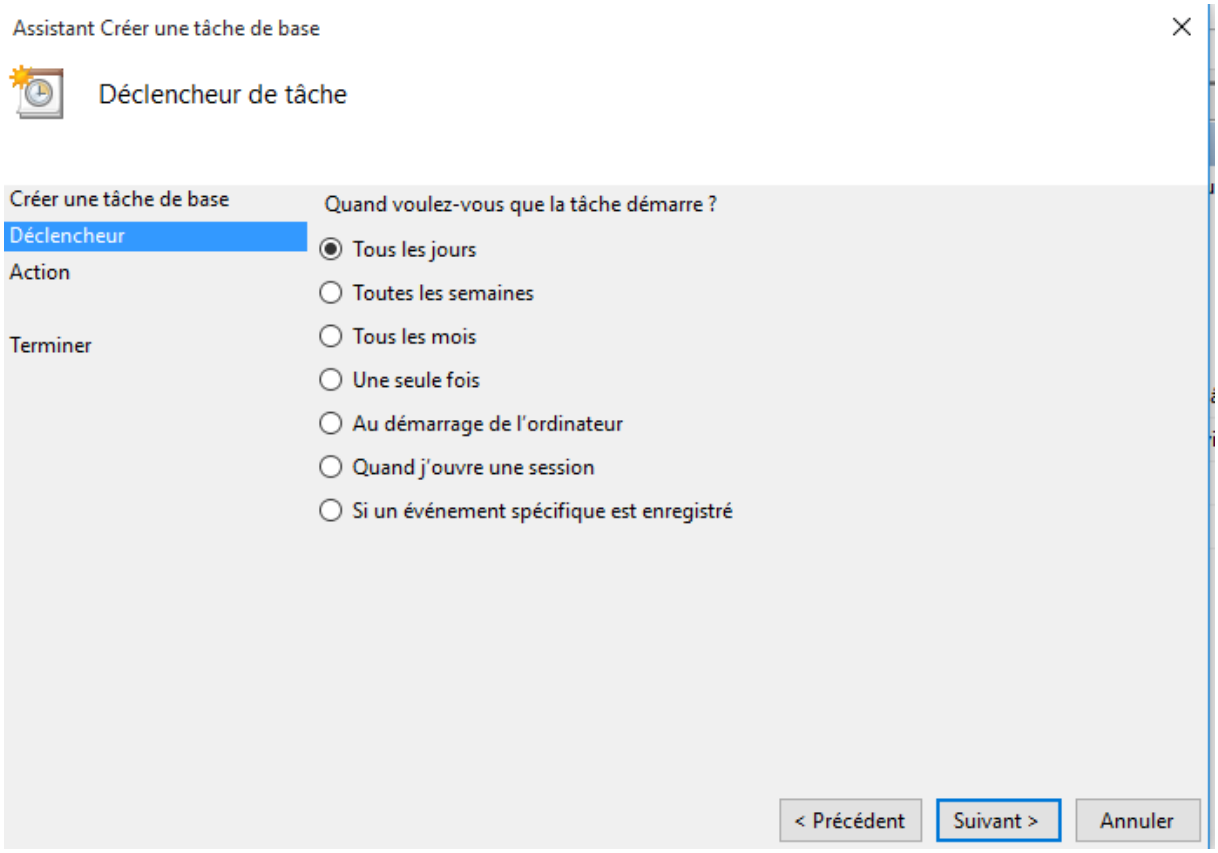


Figure 13. Etape 2 tâche planifiée

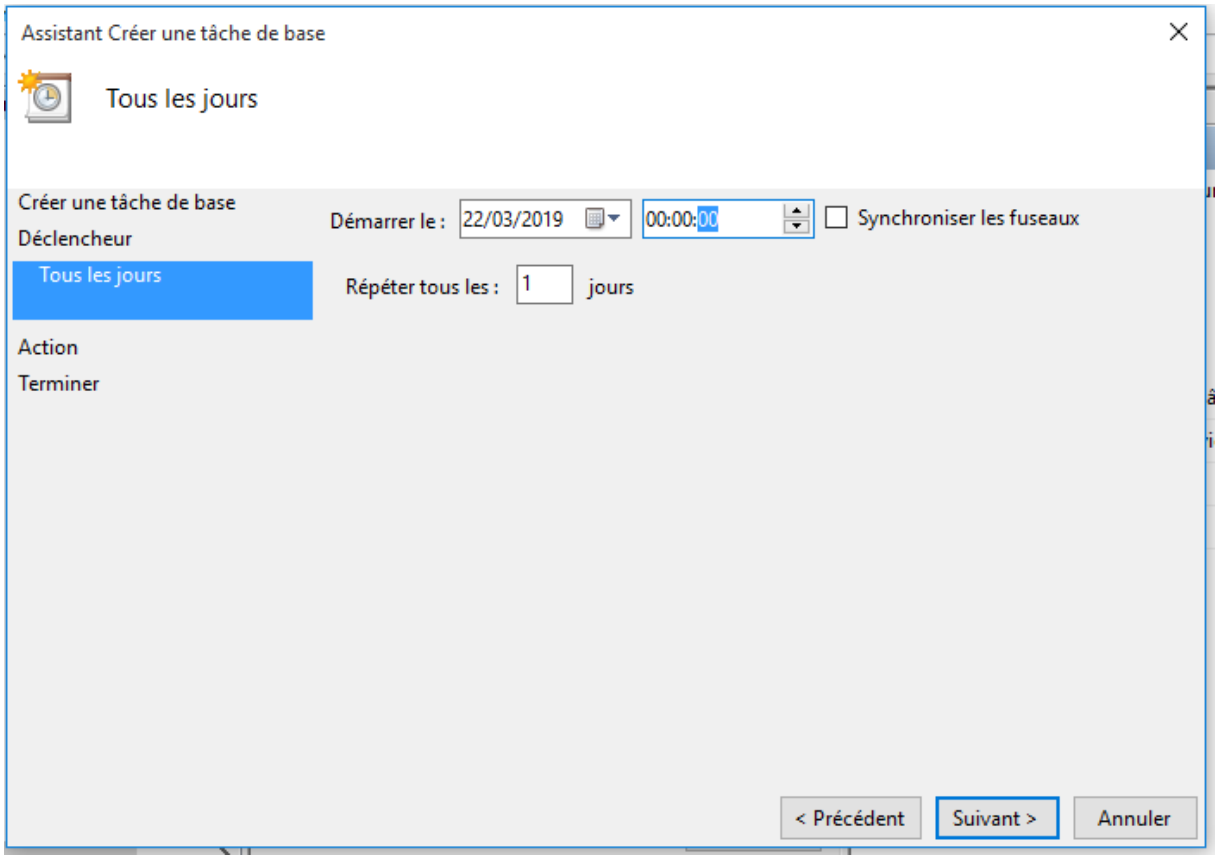


Figure 14. Etape 3 : Tache planifiée

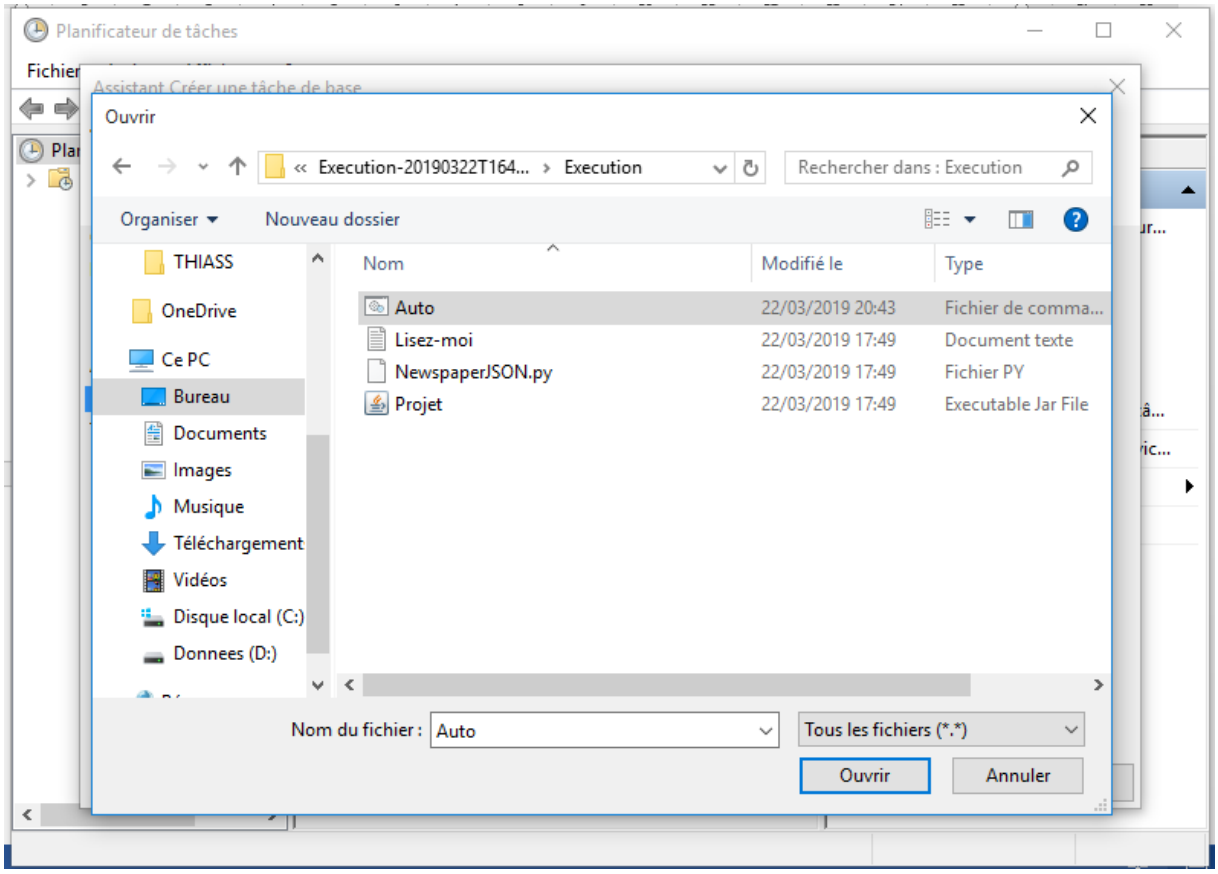


Figure 15. Etape 4 : Tâche planifiée

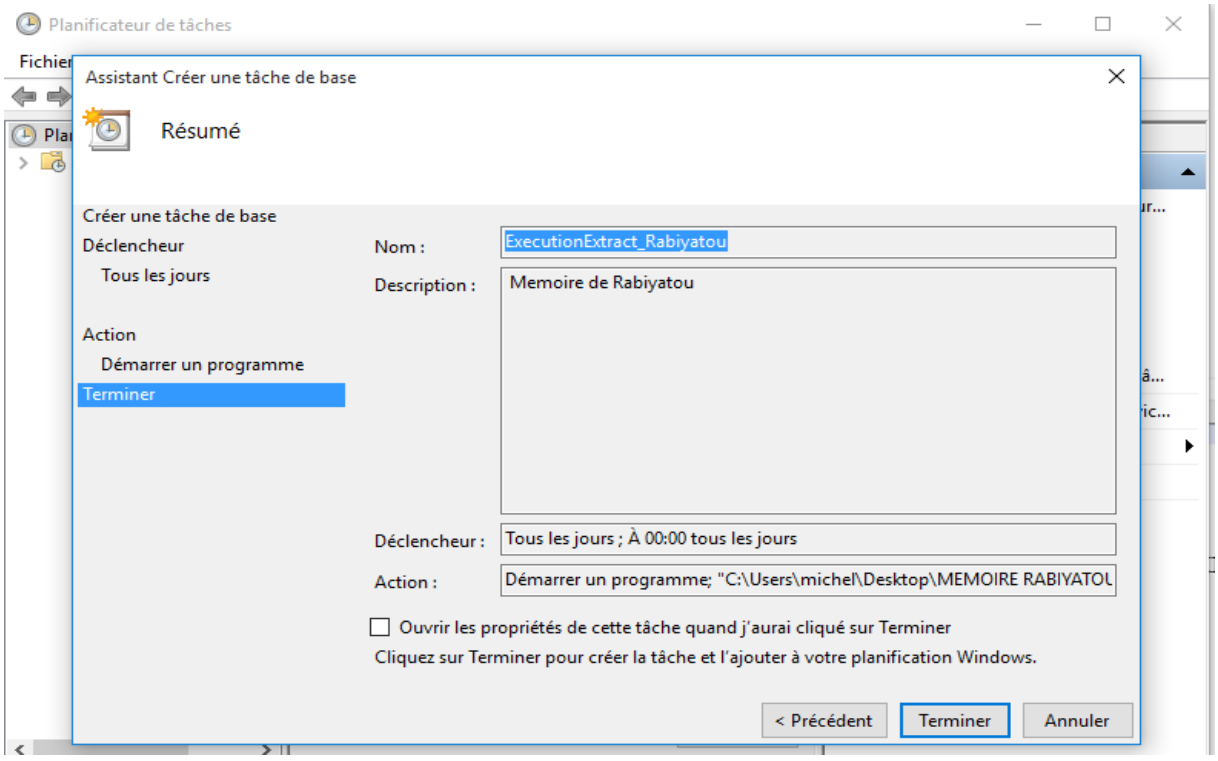


Figure 16. Etape 5 planification des tâches

2.5. Format des données en sortie

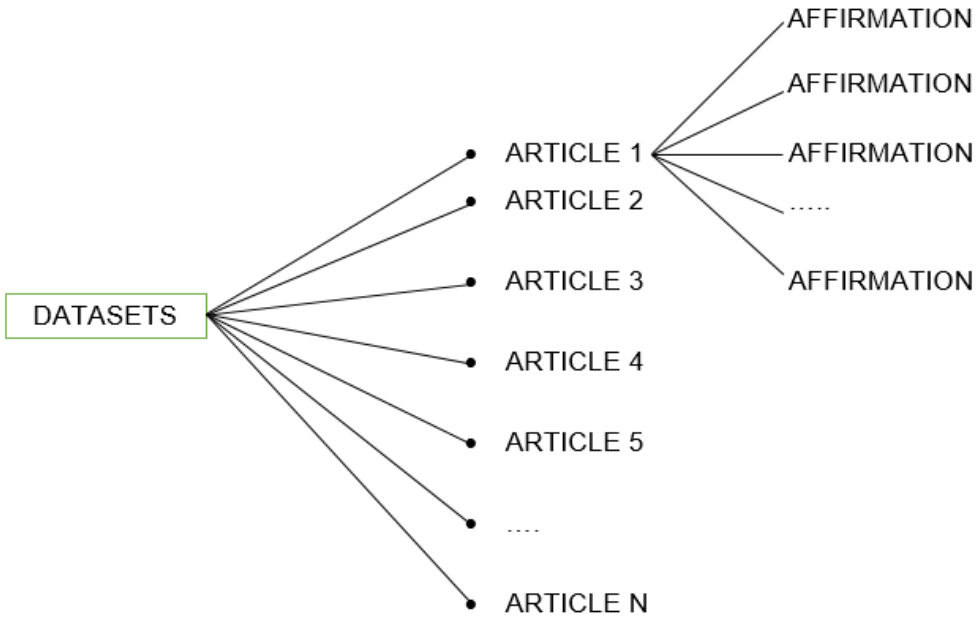


Figure 17. Architecture des données

Un article de presse en ligne est une publication sur un sujet faite par une source. Il est modélisé suivant les informations suivantes :

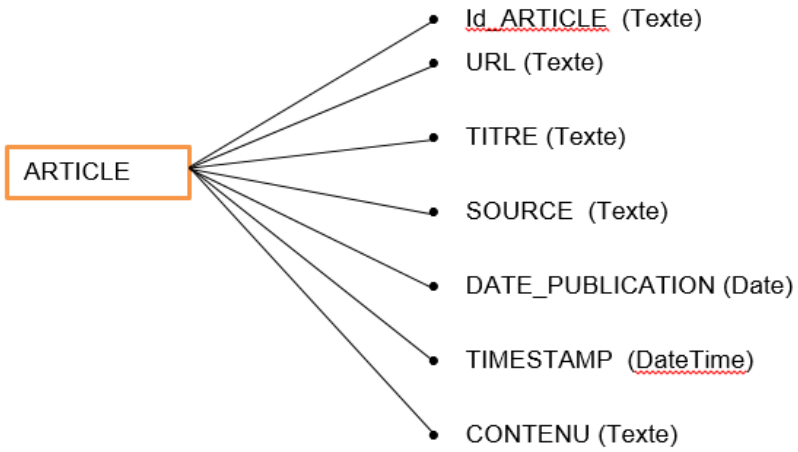


Figure 18. Format des articles

- **ID_ARTICLE**: Est un identifiant unique généré et affecté à chaque article. Il est de format TEXTE
- **URL** : Représente l'adresse web où l'article a été extrait. Il est de format TEXTE
- **TITRE** : c'est le titre principal de l'article. Lui aussi est un TEXTE
- **SOURCE** : C'est le journal (le media) ou le nom du site internet ayant publié l'article.

- **DATE_PUBLICATION**: C'est la date à laquelle l'article a été publié. Il est de format DATE
- **TIMESTAMP** : C'est l'instant précis pour l'extraction s'est faite. Il est de format DATETIME.
- **CONTENU**: C'est le texte intégral de l'article. C'est un TEXT

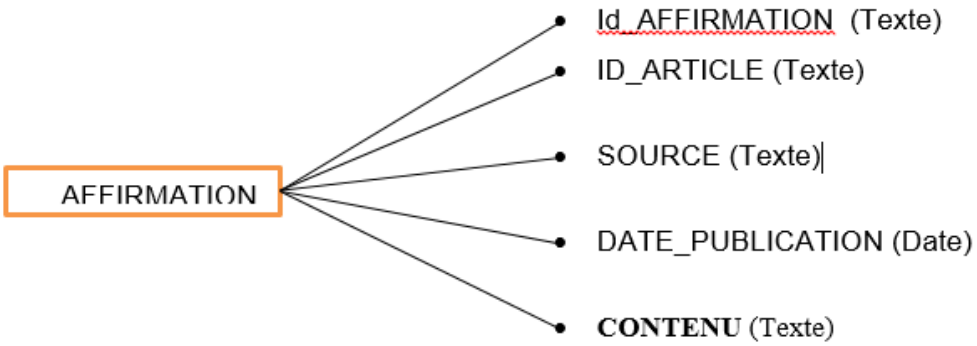


Figure 19. Format des affirmations

- **ID_AFFIRMATION**: Est un identifiant unique généré et affecté à chaque affirmation. Il est de format TEXTE
- **ID_ARTICLE**: Est un identifiant unique généré et affecté à chaque article. Il est de de format TEXTE
- **SOURCE** : C'est le journal (le media) ou le nom du site internet ayant publié l'article.
- **DATE_PUBLICATION**: C'est la date à laquelle l'article a été publié. Il est de format DATE
- **CONTENU**: C'est le texte de l'affirmation : une phrase. C'est un TEXTE

Un exemple de fichier d'Article. JSON et Affirmation. JSON sont respectivement présentés en Annexe 5 et Annexe 6.

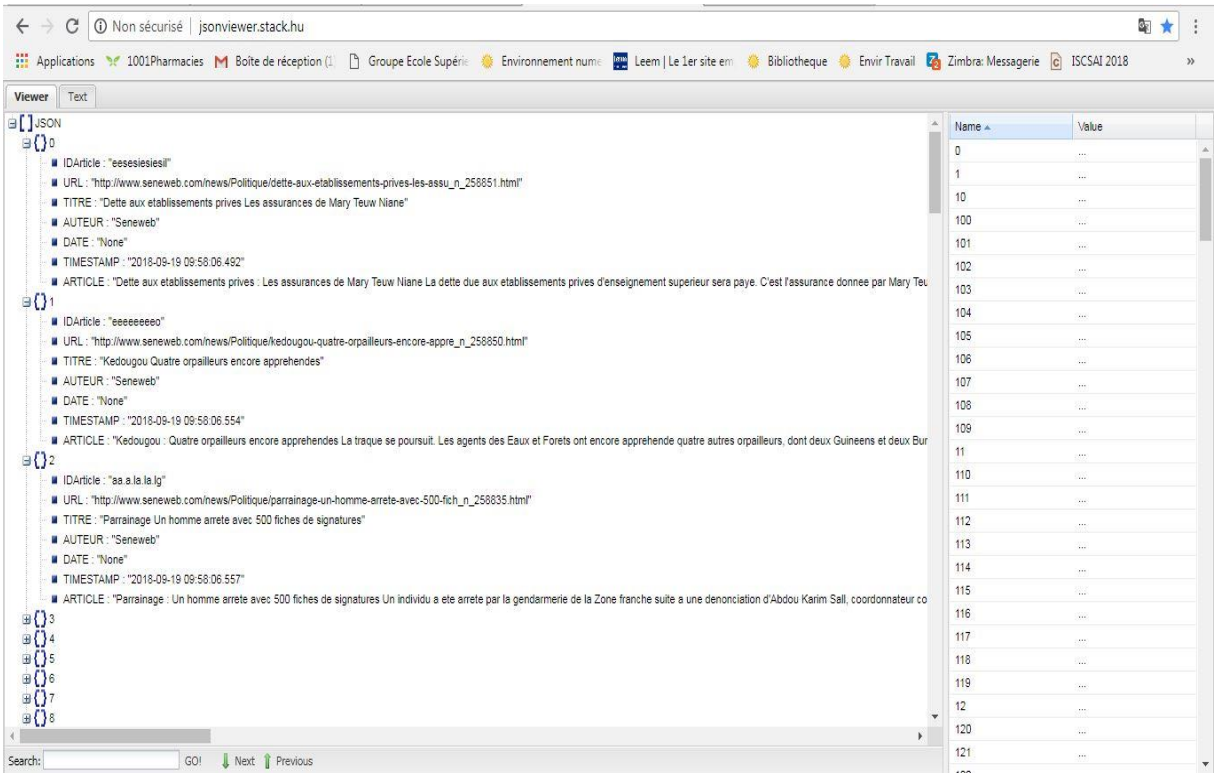


Figure 20. Article. JSON

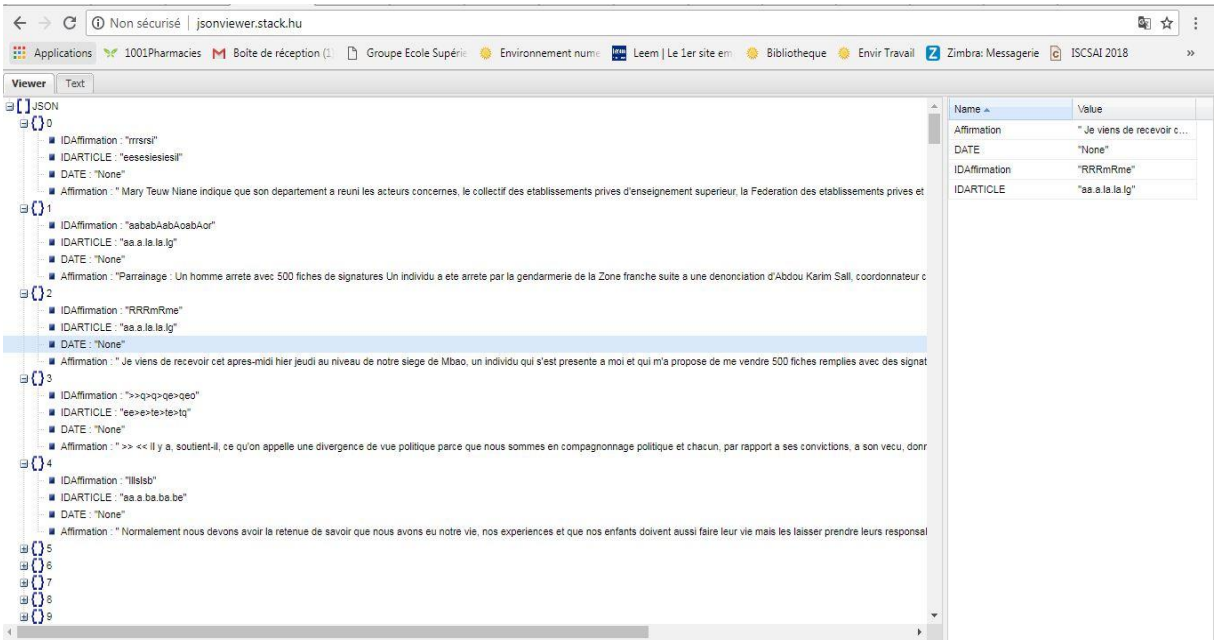


Figure 21. Affirmations. JSON

CONCLUSION ET PERSPECTIVES

De nos jours, il existe une diversité de ressources web contenant du contenu qui complète une variété de besoins. Le défi consiste à gérer ces volumes d'informations, en permettant un filtrage les informations valables et l'intégration des données. Les chercheurs doivent souvent consulter plusieurs ressources indépendantes et hétérogènes pour résoudre des problèmes spécifiques.

Cependant, le filtrage manuel prend beaucoup de temps, nécessite une expertise sur le terrain et reste néanmoins susceptible de rater les détails précieux. Les informations sont généralement disséminées sur les sites web d'institutions et peuvent uniquement être mises en évidence dans des articles de journaux ou des comptes rendus de conférences.

De plus, l'hétérogénéité des formats de données et des types de données impliqués est importante, Les ressources de données sont en général liées aux mouvements de données scientifiques ouverts, qui favorisent la diffusion publique des résultats et des données scientifiques. La valeur des données augmente donc avec une plus grande ouverture.

De nombreux sites incorporent progressivement des standards de marquage sémantique émergents et faciles à utiliser. Parallèlement, les services web constituent le moyen standard et recommandé d'activer l'accès externe aux bases de données. Néanmoins, les services Web ne suffisent pas pour assurer une interopérabilité et une intégration complètes des données.

En raison de diverses contraintes de développement telles que l'expertise technique, les coûts, l'évaluation des fonctionnalités attendues et l'établissement de la qualité de service souhaitée, il n'est pas courant de mettre à la disposition des utilisateurs des API publiques pour des bases de données et des serveurs web dès leur lancement.

En règle générale, les créateurs de sites web se concentrent sur la fourniture de contenu de haute qualité par le biais de procédures de sélection manuelle et sur le déploiement de fonctions de recherche en ligne ciblant les intérêts des praticiens. En fait, les services web standard ne sont généralement développés que pour des bases de données et des serveurs matures,

En outre, les coûts associés à l'apprentissage d'une interface programmatique plusieurs, selon toute vraisemblance, pour faire face aux problèmes d'intégration ne doivent pas être écartés et doivent être évalués en fonction du temps de réponse souhaité c'est-à-dire du temps

nécessaire au traitement des données, nature de l'application par exemple, peut ne pas être intéressé par la publication de données mises au début, mais plutôt par son utilisation dans des processus internes; par conséquent, il peut être souhaitable de garder le déploiement aussi simple que possible et la longévité de la tâche d'extraction de données tâche ponctuelle par rapport aux tâches récurrentes.

En conséquence, il est juste de reconnaître que le Scraping des données Web peut encore aider à de nombreuses tâches d'extraction d'informations, qu'il soit quotidien, ponctuel ou privé, ainsi que d'intervenir dans des projets plus vastes, Le moyen le plus courant de créer des robots web consiste à utiliser les bibliothèques tierces, souvent un tandem d'une bibliothèque d'accès au site et une bibliothèque d'analyse HTML, ce qui représente une petite courbe d'apprentissage.

Bien qu'ils impliquent une courbe d'apprentissage plus prononcée, les cadres de Scraping fournissent une couverture complète du cycle de vie du Scrapeur et, dans certains cas, des DSL facilitant la maintenance des robots. En outre, il existe également des environnements de bureau graphiques commerciaux, adaptés aux utilisateurs moins expérimentés et aux déploiements rapides et simples.

Indépendamment de la mise en œuvre, les développeurs de web Scraping doivent prendre en compte les problèmes juridiques et politiques, ainsi que les scrapeurs web du programme, dans le respect des règles. Bien que les implications juridiques ne soient pas totalement claires dans tous les cas et dans tous les pays, les développeurs doivent prendre en compte les conditions d'utilisation, à savoir, empêcher la violation du droit d'auteur, équilibrer le nombre et la fréquence des demandes et ignorer les ressources marquées.

Le moteur d'Extraction nous a permis de faire une extraction d'articles à partir des Url au niveau de 30sites de la presse en ligne sénégalaise. Le module d'extraction utilise la librairie *Newspaper* de Python pour réaliser l'extraction de tous les articles des sites web sélectionnés.

Le stockage est composée d'un ensemble de fichiers JSON et CSV constitués d'articles et d'affirmations. Le moteur combine deux programmes codés en deux langages de programmation distincts : Java et Python. Nous avons attaqué, nettoyé, segmenté en affirmations, puis fusionner et stocker tous les articles et affirmations issus du module d'extraction.

Cet outil à ce stade peut aider les journalistes dans leur travail du scraping de données en passant par le traitement en données plus structurées. Offrant ainsi une diminution considérable du travail d'extraction de données et donc un gain de temps considérable.

Une bonne perspective peut aller dans le sens de la mise en place d'un méta moteur de recherche de faits journalistiques ou la constitution d'un Dataset d'article de presse sur un sujet bien défini.

BIBLIOGRAPHIE

1. Lucas Ou-Yang. Newspaper. Consulté le 12/12/2018
2. Tragha, H. Benlahmar, S. Lassri (2016). Revue de littérature des approches d'extraction D'informations à partir du web
3. Lucas OuYang. Newspaper. Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (2017). News-please: A Generic News Crawler and Extractor. In *Proceedings of the 15th International Symposium of Information Science*. 2017.
4. Adrien Lachaize qu'est-ce-que-le-web-Scraping ?
5. Tasim, T. (2016). A general framework for scraping newspaper websites.
6. Turland, M. (2010). Php architec'ts Guide to Web Scraping. *Victoria (Canadá): Nanobooks*, *1*.
7. Beretta, V., Harispe, S., Ranwez, S., & Mougenot, I. (2016). Utilisation d'ontologies pour la quête de vérité: une étude expérimentale. In IC2016: Ingénierie des Connaissances.
8. Pons, C. (2015). L'émergence de la vérification des faits ou factchecking, et son expérimentation du futur.
9. Christophe, BRASSEUR. (2013). Enjeux et usages du Big Data: Technologies, méthodes et mise en œuvre. Lavoisier.
10. Benabdeslem, K., Biernacki, C., & Lebbah, M. (2015). Les trois défis du Big Data Éléments de réflexion. Statistique et Société, 3(1), 19-22.
11. Malik, S. K., & Rizvi, S. A. M. (2011, October). Information extraction using web usage mining, web scrapping and semantic annotation. In *2011 International Conference on Computational Intelligence and Communication Networks* (pp. 465-469). IEEE.
12. Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.
13. Nimbalkar, P. P., Chigle, M. S. R., Handoo, M. R., & Padghan, M. S. A. A Survey on Data Extraction Using Java Application and Visual Basics Macros.
14. Rouby, Aurélien, and Thibaut Tournier. "Scraping & Crawling."
15. Antheaume, A. Le Journalisme numérique, Presses de Sciences Po, 2e éd. entièrement mise à jour, 2016.
16. Hanretty, C. Scraping the web for arts and humanities (2013).

17. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in bioinformatics*, *15*(5), 788-797.

WEBOGRAPHIE

1. https://fr.wikipedia.org/wiki/Journalisme visité le 32/03/2019
2. https://www.studyrama.com/formations/fiches-metiers/internet-web/journaliste-web-1085 visité le 23/03/2019
3. https://www.python.org/ consulté le 12/02/2019
4. https://www.python.org/ Visité le 10/01/2019
5. https://www.java.com/fr/ Visité le 10/01/2019
6. https://www.java.com/fr/download/faq/whatis_java.xmlVisité le 10/01/2019
7. https://www.json.org/ Visité le 10/01/2019
8. https://www.computerhope.com/issues/ch001356.htm visité le 22/03/2019
9. https://pypi.org/project/newspaper/ visité le 22/03/2019
10. https://www.crummy.com/software/BeautifulSoup/bs4/doc/ visité le 22/03/2019
11. https://jaunt-api.com/index.htm visité le 22/03/2019
12. https://jsoup.org/ visité le 22/03/2019
13. https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd visité le 21/10 /2018
14. https://chrome.google.com/webstore/detail/webscraper/jnhgnonknehpejjnehehllkliplmbmhn?hl=fr visité le 12/10/2018
15. https://addons.mozilla.org/en-US/firefox/addon/datascraper/?src=search visité le 10/10/2018
16. https://addons.mozilla.org/en-US/firefox/addon/cloump-u-scraper-plugin/?src=search visité le 14/10/2018
17. https://webhose.io/ visité le 10/03/2018
18. https://www.import.io/ visité le 13/03/2018
19. https://scrapy.org/ visité le 10/03/2018
20. http://phantomjs.org/ visité le 22/03/2019
21. https://www.outwit.com/#hub visité le 22/03/2019
22. https://media.readthedocs.org/pdf/newspaper/latest/newspaper.pdf visité le 10/10/2018
23. https://scrapy.org/ visité le 22/03/2019

24. https://www.import.io/product/ visité le 22/03/2019
25. https://www.outwit.com/#hub visité le 22/03/2019
26. http://weboob.org/ visité le 22/03/2019
27. http://phantomjs.org/ visité le 22/03/2019
28. https://www.python.org/downloads/release/python-364https://github.com/codelucas/newspaper
29. https://www.studyrama.com/formations/fiches-metiers/internet-web/journaliste-web-1085
30. https://docs.python.org/fr/3/tutorial/index.html
31. https://www.computerhope.com/issues/ch001356.htm
32. https://www.rgdesign.fr/blog/web-scraping/
33. https://www.scraperapi.com/blog/the-10-best-web-scraping-tools
34. http://htmlcleaner.sourceforge.net
35. http://www.gnu.org/software/wget/
36. http://www.gnu.org/software/ gawk /
37. http://www.gnu.org/software/sed/
38. http://web-harvest.sourceforge.net/
39. http://sing.ei.uvigo.es/jarvest
40. http://developers.facebook.com/docs/opengraph
41. https://lobstr.io/index.php/2018/04/19/les-meilleurs-outils-web-scraping-gratuits-2018/
42. www.seneweb.com
43. https://www.google.com/search?q=web+scraping&rlz=1CDGOYI_enSN797SN799&hl=fr&prmd=vnbi&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiX0PuJvJjiAhVpRxUIHQspCgwQ_AUoBHoECA8QBA&biw=375&bih=551

ANNEXE 1: SCRAPER PYTHON

```
# coding: utf8
# toutes les chaines sont en Unicode (même les docstrings)
# -*- coding: utf-8 -*-

from __future__ import unicode_literals
import sys
import os
import codecs
import newspaper
import nltk
import shutil
from nltk.stem.snowball import FrenchStemmer
from nltk.corpus import stopwords
import re
from collections import Counter
#nltk. Download()
#nltk. Download('punkt')
from newspaper import Article
from nltk import wordpunct_tokenize
import codecs
from Unicode import Unicode
import simplejson as json

#On crée une fonction ScrapingSource que nous allons appeler plus tard
def MoteurExtract (src, cheminTexte, cheminTexteJSON):
        s = newspaper. Build(src,memoize_articles=True)
        tabURL=[]
        LesFaits="
        ContenuArticle="
        for a in articles:
                TabURL. Append(str(a.url))
                print(a.url)

        for i in range(len(tabURL)):
                try:
```

```
                        art= Article(tabURL[i], language = 'fr')
                        art.download()
                        art.parse()
                        art.nlp()
                        url=str(art.url)
                        title=str(art.title)
                        authors=str(art.authors)
                        datep=str(art.publish_date)
                        mots=str(art.keywords)
                        resume=str(art.summary)
                        textes=str (art.text)
                        image =str (art.top_image)
                        texte= textes. Replace ('\r\n', ' ')
                        texte= textes. Replace ('\n', ' ')
                        texte= texte.replace ('#', ' ')
                        texte= texte.replace ("", ' ')
                        texte=texte+' '
                        datep=datep. Replace ('\n', ' ')
                        strTexte= url+'#'+title+'#'+authors+'#'+datep+'#'+texte+'#'+image
                        strTexteJSON =
'[{"URL":"'+url+'","TITRE":"'+title+'","AUTEUR":"'+authors+'","DATEPUB":"'+datep+'","ARTICL
E":"'+texte+'","IMAGETOP":"'+image+'"}]'
                        faJSON = open(cheminTexteJSON+'/Article'+str(i)+'.txt','w')
                        fa = open(cheminTexte+'/texte'+str(i)+'.json','w')
                        strTexteJSON=unidecode (strTexteJSON);
                        strTexte=unidecode (strTexte);
                        print (strTexteJSON)
                        faJSON.write (strTexteJSON)
                        fa.write (strTexte)
                        strTexteJSON="
                        strTexte="
                        faJSON.close ()
                        fa.close ()
                except:
                    print (" ")

def StopWords (sentence):
```

```
        stop=set(stopwords.words('French'))
        return ([i for i in sentence.lower().split() if i not in stop])

# run*********************************
def run(chemin, source, media):
        chemin=chemin+'/'+media
        cheminTexte=chemin+'/Textes'
        cheminTexteJSON=chemin+'/ArticlesJSON'
        if not os.path.exists(chemin):
                os.mkdir(chemin)
        if not os.path.exists(cheminTexte):
                os.mkdir(cheminTexte)
        if not os.path.exists(cheminTexteJSON):
                os.mkdir (cheminTexteJSON)
        MoteurExtract (source, cheminTexte, cheminTexteJSON)

#--------------EXECUTION ----------------------
import os
chemin = sys.argv[1]
if not os.path.exists(chemin):
os.mkdir (chemin)

import subprocess
sourceSeneweb='http://www.seneweb.com/news/politique/'
sourceObs='https://www.igfm.sn/category/politique/'
sourceJeuneAfrique='http://www.jeuneafrique.com/pays/senegal/'
sourceRewmi='http://www.rewmi.com/categorie/politique'
sourceActuSen='https://actusen.sn/category/politique/'
sourceWalf='https://www.walf-groupe.com/category/actualites/politique/'
sourceAps ='http://www.aps.sn/actualites/politique/'
sourceQuotidien='https://www.lequotidien.sn/'
sourceEnquete='http://www.enqueteplus.com/sections/politique'
sourceSenego='https://senego.com/rubrique/politique'
sourceSeneNews='https://www.senenews.com/category/politique'
sourceSen360='https://news.sen360.sn/politique/'
sourceDakarActu='https://www.dakaractu.com/'
sourceDirect='https://www.senegaldirect.net/category/politique/'
```

```
run(chemin, sourceSeneweb, ‘SeneWeb')
print ("Extraction SENEWEB TERMINER")
#------------------------------------------------------------
```

ANNEXE 2: APPEL DE NEWSPAPER DANS JAVA

```
package MoteurExtract_Exemple;
import java.io.*;
public class AppelNewsPaper {
public static void Newspaper (String Newspaper, String pythonScriptPath) throws
IOException {
        //creation de la commande python
        String[] cmd = new String[7];
        cmd[0] = "python";
        cmd[1] = pythonScriptPath;
        cmd[2] = Newspaper;
        System.out.println("Chemin fichier python "+cmd[1]);
        System.out.println ("Chemin destination des résultats "+cmd [2]);
        // création d'un runtime pour une exécution externe
        Runtime rt = Runtime.getRuntime();
        Process pr = rt.exec(cmd);
        // retrieve output from python script
        BufferedReader bfr = new BufferedReader(new InputStreamReader(pr.getInputStream()));
        String line = "";
        while((line = bfr.readLine()) != null) {
        // display each output line form python script
        System.out.println(line);
        }
        }
}
```

ANNEXE 3: APPEL DU SCRAPER PYTHON

```
package MoteurExtract_Exemple;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.nio.charset.Charset;
import java.util.StringTokenizer;
Import java.io.*;
Import java.sql.*;
public class MoteurExtractJSON {
        public static void ModuleMFA(String Newspaper, String media, String user,String  pwd)
throws Exception {
                        int i = 0;
            //création des dossiers Datasets et DatasetsGlobal
                        String dataset=Newspaper+"/Datasets";
                        String datasetGlobal=Newspaper+"/DatasetsGlobals";
                        File fileD = new File(dataset);
                        File fileDG = new File(datasetGlobal);
                        if (!(fileD.exists()))
                              fileD.mkdir();
                        if (!(fileDG.exists()))
                                        fileDG.mkdir();
            try {
                                run(user, pwd,Newspaper, media);
            }catch(Exception e) {
                System.out.println(e.getMessage());
            }
}

        public static void run(String user, String pwd, String chemin, String source) throws Exception {
                String Source= chemin+"/"+source+"/Textes";
                String dest= chemin+"/Datasets/Articles_"+source+".csv";
```

```
                String destAff= chemin+"/Datasets/Affirmations_"+source+".csv";
                String destAffStop= chemin+"/Datasets/Affirmations_StopWord_"+source+".csv";
                String destAffirmationDataset=
chemin+"/DatasetsGlobals/Sn_Affirmations_Dataset.csv";
                String destAffirmationDatasetStop=
chemin+"/DatasetsGlobals/Sn_Affirmations_Dataset_StopWord.csv";
                String destArticlesDataset= chemin+"/DatasetsGlobals/Sn_Articles_Dataset.csv";
String [] TabOldAffirmation= ChargerTab (destAffirmationDataset);

                String auteur=source;
                int nbarticle= countFiles(Source);
            System.out.println ("------- "+ nbarticle +" articles "+source+"  trouvés");

                ChargerFaits (user, pwd, Source, nbarticle, dest, auteur, destAff, destAffStop,
destAffirmationDataset, destArticlesDataset, destAffirmationDatasetStop, TabOldAffirmation, chemin,
source);



         }
         static boolean deleteAll (File dir) {
if (dir.isDirectory()) {
        File[] children = dir.listFiles();
        for (int i=0; i<children.length; i++) {
           boolean success = deleteAll(children[i]);
           if (!success) return false;
        }
     }
     return dir.delete();
}
         //********************** compter le nombre de fichier dans un dossier
         public static int countFiles (String parent) throws Exception {
         File  = new File (parent);

            if (!file.exists ())
               throw new FileNotFoundException ();
     int nb=file.list ().length;
```

```
            return nb;
        }

…



        public static   String CreerID (String Titre, String Texte) throws IOException
   {
                int charLength = Texte.length();
        StringBuilder  pass = new StringBuilder (charLength);
        String ID="";
      ID+=pass.append(Texte.charAt(1));
      ID+=pass.append(Texte.charAt(charLength-3));
      ID+=pass.append(Texte.charAt(charLength-10));
      ID+=pass.append(Texte.charAt(charLength-1));
      ID+=pass.append(Texte.charAt(charLength-8));
      ID=ID.replaceAll(" ", "");
      ID=ID.replaceAll(",", "");
      ID=ID.replaceAll("'", "");
      return ID;
}



        //****** vérifier si existe ********** return 1 si déjà enregistré*********
        public static int Verification Si Déjà Enregistrer(String [] Tab, String ID) {
                int nb= Tab.length;
                int i=0;
                int rep=1;
                for(i=0; i<nb; i++) {
                        if (Tab[i].equals(ID)) {
                                rep=0;
                                break;
                        }
                        else {
                                rep=1;
                        }
```

```
                }
                return rep;
        }


}
```

ANNEXE 4: SETUP.JAVA

```
package MoteurExtract_Exemple;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.nio.charset.Charset;
import java.util.StringTokenizer;
Import java.io.*;
Import java.sql.*;
public class Setup {
        public static void main(String[] args) throws Exception {
        System.out.println ("------ EXTRACTION AUTOMATIQUE DES ARTICLES ET
AFFIRMATIONS JOURNALISTIQUES DE LA PRESSE SENEGALAISE-----");
        System.out.println("-------Edouard SARR, Ousmane SALL & Babiga BIRREGAH   ");
          System.out.println("-------UNIVERSITE DE THIES (THIES-SENEGAL) & UTT (TROYES-
FRANCE) \n");
                AppelNewsPaper NP= new AppelNewsPaper();
                 String Sortie="Resultats Collecte Articles";
                        int i=0;
                        String user="root";
                        String pwd="";
                        String pythonScriptPath= "NewspaperJSON.py";

            while (i < args.length)
            {
                    System.out.println(args[i]);
                if(args[i].startsWith("-S"))  //spécifier la sortie
                        Sortie = args[i].substring(2);
                else if(args[i].startsWith("-N"))  //spécifier le fichier python
                        pythonScriptPath = args[i].substring(2);
                else if(args[i].startsWith("-S"))  //
                        Sortie = args[i].substring(2);
```

```
                i++;
            }
            System.out.println(" &&&&&&&&&&&&&&&&&&&&&&&&& "+Sortie);
            File F =new File(Sortie);
            if (F.exists()) {
        System.out.println ("Le dossier de destination existe déjà : " + F.getAbsolutePath());
            }
            else {
                if (F.mkdir()) {
                   System.out.println ("Ajout du dossier : " + F.getAbsolutePath());
                }
                else {
                   System.out.println ("Echec sur le dossier : " + F.getAbsolutePath());
                }
            }



                   NP.Newspaper (Sortie, pythonScriptPath);



            //** extraction et fusion sur 20 sources dans des datasets
                MoteurExtractJSON omfa=new MoteurExtractJSON ();
        System.out.println ("-------FIN EXTRACTION DES ARTICLES");
        System.out.println ("------- DEBUT DE LA SEGMENTATION \n");

try {
                omfa.ModuleMFA(Sortie, "Seneweb",user, pwd);
            }catch(Exception e) {}

            try {
                omfa.ModuleMFA (Sortie, "LObs", user, pwd);
        } catch (Exception e) {}
            try {
                omfa.ModuleMFA (Sortie, "JeuneAfrique", user, pwd);
        }catch(Exception e) {}
            try {
                omfa.ModuleMFA(Sortie, "ActuSen",user, pwd);
```

```
        }catch(Exception e) {}
        try {
            omfa.ModuleMFA(Sortie, "Rewmi",user, pwd);
        }catch(Exception e) {}
        try {
            omfa.ModuleMFA(Sortie, "Walf",user, pwd);
        }catch(Exception e) {}
        try {
            omfa.ModuleMFA (Sortie, "Aps", user, pwd);
    } catch (Exception e) {}
        try {
            omfa.ModuleMFA (Sortie, "Quotidien ", user, pwd);
    }catch(Exception e) {}
        try {
            omfa.ModuleMFA (Sortie, "Enquete ", user, pwd);
        }catch(Exception e) {}
        try {
            omfa.ModuleMFA(Sortie, "Senego ", user, pwd);
        }catch(Exception e) {}
        try {
            omfa.ModuleMFA(Sortie, "SeneNews ", user, pwd);
        }catch(Exception e) {}

        try {
            omfa.ModuleMFA(Sortie, "Sen360",user, pwd);
        }catch(Exception e) {}

        try {
            omfa.ModuleMFA (Sortie, "DakarActu ", user, pwd);
    }catch(Exception e) {}

        try {
            omfa.ModuleMFA (Sortie, "Direct ", user, pwd);
        }catch(Exception e) {}


                    System.out.println("\n\n----------------------------------------------------------------
```

```
");
                        System.out.println ("----------------- EXTRACTION ET FUSION TERMNIER --
----------------");
                        System.out.println("-----------------------------------------------------------------");


        }
}
```

ANNEXE 5: ARTICLE. JSON

{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la même heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"TIMESTAMP":"2018-10-17 13:41:45.783",
"CONTENU": "Sante Pour tomber rapidement enceinte, il faut se coucher a la même heure PARTAGES | J'AIME | Une étude, réalisée par des chercheurs de l'université Washington, de Saint-Louis, montre l'importance du sommeil pour avoir une chance de concevoir rapidement un bébé. D'après l'Institut national d'études démographiques Ined, il faut compter environ 7 mois pour concevoir un bébé. Au bout d'un an, 97 % des couples en désir d'enfant y arrivent. Se coucher à heure fixe Les chercheurs ont suivi les habitudes de sommeil de 176 femmes pendant un an, à l'aide de montres connectées. Ils se sont aperçus que les femmes qui se couchaient, chaque soir, dans la même fourchette d'une heure, étaient arrivées à concevoir plus rapidement un enfant que celles qui avaient des heures de coucher très différentes d'un jour à l'autre. Un rythme quotidien régulier est important pour le bon fonctionnement de l'organisme, et pas seulement pour lutter contre la fatigue diurne. Etre régulier dans ses heures de coucher a aussi un retentissement sur l'ovulation, ce que ne savent pas beaucoup de couples. "
}{
"ID_ARTICLE":"aasasasaso",
"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",
"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la même heure",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"TIMESTAMP":"2018-10-17 13:41:45.783",
"CONTENU": "Sante Pour tomber rapidement enceinte, il faut se coucher a la même heure PARTAGES | J'AIME | Une étude, réalisée par des chercheurs de l'université Washington, de Saint-Louis, montre l'importance du sommeil pour avoir une chance de concevoir rapidement un bébé. D'après l'Institut national d'études démographiques Ined, il faut compter environ 7 mois pour concevoir un bébé. Au bout d'un an, 97 % des couples en désir d'enfant y arrivent. Se coucher à

heure fixe. Les chercheurs ont suivi les habitudes de sommeil de 176 femmes pendant un an, à l'aide de montres connectées. Ils se sont aperçus que les femmes qui se couchaient, chaque soir, dans la même fourchette d'une heure, étaient arrivées à concevoir plus rapidement un enfant que celles qui avaient des heures de coucher très différentes d'un jour à l'autre. Un rythme quotidien régulier est important pour le bon fonctionnement de l'organisme, et pas seulement pour lutter contre la fatigue diurne. Etre régulier dans ses heures de coucher a aussi un retentissement sur l'ovulation, ce que ne savent pas beaucoup de couples. "

}{

"ID_ARTICLE":"aasasasaso",

"URL":"http://www.buzzsenegal.com/news/Sante/pour-tomber-rapidement-enceinte-il-faut-_n_34003.html",

"TITRE":"BuzzSenegal.com : Pour tomber rapidement enceinte, il faut se coucher a la même heure",

"SOURCE":"SeneWeb",

"DATE_PUBLICATION": "None",

"TIMESTAMP":"2018-10-17 13:41:45.783",

"CONTENU": "Sante Pour tomber rapidement enceinte, il faut se coucher a la même heure PARTAGES | J’AIME | Une étude, réalisée par des chercheurs de l'université Washington, de Saint-Louis, montre l'importance du sommeil pour avoir une chance de concevoir rapidement un bébé. D'après l'Institut national d'études démographiques Ined, il faut compter environ 7 mois pour concevoir un bébé. Au bout d'un an, 97 % des couples en désir d'enfant y arrivent. Se coucher à heure fixe. Les chercheurs ont suivi les habitudes de sommeil de 176 femmes pendant un an, à l'aide de montres connectées. Ils se sont aperçus que les femmes qui se couchaient, chaque soir, dans la même fourchette d'une heure, étaient arrivées à concevoir plus rapidement un enfant que celles qui avaient des heures de coucher très différentes d'un jour à l'autre. Un rythme quotidien régulier est important pour le bon fonctionnement de l'organisme, et pas seulement pour lutter contre la fatigue diurne. Etre régulier dans ses heures de coucher a aussi un retentissement sur l'ovulation, ce que ne savent pas beaucoup de couples. "

..

ANNEXE: AFFIRMATIONS. JSON

{
"ID_AFFIRMATION":"aaeaenaeneaene",
"ID_ARTICLE":"aasasasaso",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"CONTENU": "Sante Pour tomber rapidement enceinte, il faut se coucher a la même heure PARTAGES | J'AIME | Une étude, réalisée par des chercheurs de l'université Washington, de Saint-Louis, montre l'importance du sommeil pour avoir une chance de concevoir rapidement un bébé"
}{
"ID_AFFIRMATION":"IItItIteIte",
"ID_ARTICLE":"aasasasaso",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"CONTENU": " Ils se sont aperçus que les femmes qui se couchaient, chaque soir, dans la même fourchette d'une heure, étaient arrivées à concevoir plus rapidement un enfant que celles qui avaient des heures de coucher très différentes d'un jour à l'autre"
}{
"ID_AFFIRMATION":"LLeLeLeuLeuu",
"ID_ARTICLE":"LLLLLn",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"CONTENU": "OLA Energy : Le nouveau visage de votre distributeur pétrolier panafricain, un des plus apprécies Nouveau logo, nouvelle identité visuelle et encore plus d'attention aux clients ! Coup de neuf sur les stations-services du groupe OiLibya dont la nouvelle image se déclinera progressivement à travers l'ensemble du réseau"
}{
"ID_AFFIRMATION":"iiisisf",
"ID_ARTICLE":"LLLLLn",
"SOURCE":"SeneWeb",
"DATE_PUBLICATION": "None",
"CONTENU": " Dakar, 16 Octobre 2018 - Marquant un tournant de son histoire dans la distribution et la commercialisation des produits pétroliers en Afrique, le Groupe OiLibya vient de dévoiler la nouvelle identité visuelle des stations-service de son réseau de distribution panafricain constitue de plus de 1100 stations-service à travers 17 pays africains"
}{
"ID_AFFIRMATION":"eeeeeel",
"ID_ARTICLE":"LLLLLn",
"SOURCE" : "SeneWeb",
"DATE_PUBLICATION": "None",
"CONTENU": " OLA Energy, a l'instar des autres filiales du Groupe, mettra davantage l'accent sur l'excellence du service client, tout en offrant de nouvelles gammes élargies de produits et services pour mieux répondre à l'évolution des besoins de sa clientèle"
}