Université de la Manouba
Ecole Nationale des Sciences de L'Informatique

Rapport de stage d'immersion en entreprise

SDC : Collecteur de données intelligent

Réalisé par :
Abir Bel Haj Youssef

Organisme : SFM TECHNOLOGIES
Encadré par : Mr. Ali Ben Amor
Adresse : 81, Avenue Hédi Chaker, 1002 Tunis, Tunisie
TEL : +216 98 377 887 *FAX* : +216 71 284 314
eMail : info@sfmtelecom.com
Site web : www.sfmtechnologies.com

Année Universitaire :
2019-2020

Supervisor's signature

Remerciements

Ce projet est rendu possible grâce à la patience et aux conseils de mon encadrant, Mr Ali Ben Amor. Je lui suis très reconnaissante pour son soutien constant et ses encouragements qui ont amené mon travail dans la bonne direction.

Je tiens également à remercier ma chère famille pour ses encouragements, son soutien et son aide à chaque étape que je fais. Elle me donne la force dont j'ai besoin pour aller vers l'avant.

Mes salutations distinguées et mes meilleurs remerciements aux membres du jury d'avoir pris le temps d'évaluer mon travail et me donner de précieux commentaires.

Enfin, merci beaucoup à toutes les autres personnes non mentionnées qui ont contribué directement ou indirectement à ce travail. Votre aide et votre soutien sont très appréciés.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le volume de données à traiter a explosé à des niveaux inimaginables au cours de ces dernières années et, parallèlement, le prix des outils de stockage de données a systématiquement diminué. Les entreprises privées et les instituts de recherche saisissent des téraoctets de données relatives aux interactions de leurs utilisateurs, aux entreprises, aux médias sociaux ainsi qu'aux capteurs fournis par des appareils tels que les téléphones portables et les automobiles.

Le défi de cette époque est de donner un sens à cette mer de données. Ce qui donne naissance à un nouveau domaine de l'informatique appelé le "Big Data". Le Big Data implique en grande partie la collecte de données de différentes sources et structures, de manière à ce qu'elles deviennent disponibles pour être utilisées par les analystes et enfin à fournir des données utiles aux activités de l'organisation.

La collecte de données joue le rôle le plus important dans le cycle du Big Data car elles sont utilisées dans les modèles d'apprentissage automatique. Internet fournit des sources de données presque illimitées pour une variété de sujets. Par exemple, supposons que nous voulons construire un système qui recommande les restaurants. La première étape consiste à recueillir des données, dans ce cas, des critiques de restaurants de différents sites Web et à les stocker dans une base de données.

Dans ce projet, les architectures utilisant l'exécution parallèle et le partage des tâches sont des outils puissants pour améliorer les résultats, accélérer le temps d'exécution et éviter les problèmes de mémoire.

Ce projet s'articule autour de trois chapitres :

Le premier chapitre, intitulé "Étude préliminaire", comprend présentation de l'organisme d'accueil, une analyse théorique des concepts sur lesquels notre projet est basé, une étude de l'existant, et une solution à notre problème.

Le deuxième chapitre, intitulé "Spécification des besoins et conception", décrit la spécification des besoins fonctionnels et techniques de l'application ainsi que sa modélisation conceptuelle.

Le dernier chapitre, intitulé "Réalisation", comprend une présentation de l'environnement de travail avec une description de certaines interfaces de l'application et de l'exécution du résultat obtenu.

Enfin, nous terminons ce travail par une conclusion générale dans laquelle nous résumons notre solution et exposons quelques perspectives futures.

Chapitre 1

Étude préliminaire

Dans ce chapitre, nous procédons à travers trois sections : la première section est consacrée à la présentation de l'organisme d'accueil. La seconde définit le contexte théorique de ce projet pour résoudre sa problématique. La troisième couvre une étude des technologies existantes et établit une analyse comparative afin de choisir un examen technique à la suite.

1.1 Organisme d'accueil

SFM Technologies est un cabinet d'ingénierie et d'expertise en technologies de l'information et réseaux de télécommunications fixes, mobiles, et fibre optique, créé depuis 1995 et opérant en Afrique, Asie, Europe, Océanie et États Unis. Au cours de la dernière décennie, il a construit et maintenu une solide réputation de développeur de solutions et d'outils, d'expert technique, de conseiller stratégique et de formateur de haut niveau. SFM Technologies propose des solutions hautement évolutives et fiables permettant la collecte des données en temps réel à partir de plusieurs systèmes d'installation et de divers réseaux d'entreprise.

1.2 Etat de l'art

Dans cette section, nous définissons les principaux concepts de Big Data. Ensuite, nous nous expliquons le passage du modèle 3-V au modèle multi-V.

1.2.1 Concepts de Big Data

Le Big Data est un ensemble de données dont l'échelle, la diversité et la complexité nécessitent une architecture, des techniques, des algorithmes et des analyses pour le gérer et extraire de la valeur et des connaissances cachées. L'analyse des bases de données traditionnelles indique non seulement ce qui s'est passé et ce qui se passe, mais aussi l'analyse prédictive de ce qui est susceptible de se produire à l'avenir. Les besoins d'infrastructure du Big Data sont l'acquisition, l'organisation et l'analyse de données.

Le terme "Big Data" décrit des techniques et des technologies innovantes pour capturer, stocker, distribuer, gérer et analyser des données de pétaoctets ou de plus grande taille ayant des différentes structures. Les méga-données peuvent être structurées, non structurées ou semi-structurées, ce qui en résulte l'inaptitude des méthodes classiques de gestion des données. Le Big Data fait référence à l'explosion de la quantité et parfois de la qualité des ressources disponibles et des données potentiellement pertinentes, résultant en grande partie des progrès récents et des technologies de stockage de données. Dans ce nouveau et passionnant monde, la taille des échantillons n'est plus mesuré en nombre d'observations, mais plutôt en mégaoctets.

La définition la plus populaire de ces dernières années est celle des "Trois V" : Volume, Vitesse et Variété. Le concept a été soulevé pour la première fois par Doug Laney (2001) dans sa note de recherche du groupe META [1] qui décrit les caractéristiques des données qui ne peuvent pas être traitées par des outils de gestion traditionnels. Avec l'intérêt croissant pour le Big Data, les "Trois V" ont été étendus à "Cinq V" : Volume, Vitesse, Variété, Véracité et Valeur, et maintenant il s'agit du modèle multi-V.

1.2.2 Le modèle 3-V

Il est généralement admis que les méga-données peuvent être expliquées selon trois V : Vélocité, Variété et Volume.

Vélocité :
Elle représente la vitesse de génération des données. Le flux de données dans les sites internet et les réseaux sociaux est massif et continu. Ces données en temps réel peuvent aider les utilisateurs à prendre des décisions utiles. Ainsi, la vitesse signifie l'analyse des données en continu.

Variété :
Elle fait référence aux nombreuses sources et types de données, structurées et non structurées. Les données sont stockées dans des feuilles de calcul et des bases de données provenant de différentes sources. Maintenant, les données se présentent sous la forme de fichiers csv, de photos, de vidéos, PDF, audio, etc. Cette variété de données non structurées pose des problèmes de stockage, d'extraction et d'analyse de données. Afin de résoudre ce problème, nous devons définir les systèmes de stockage qui peuvent analyser une variété de données.

Volume :
Le Big Data implique d'énormes volumes de données. Ces données sont générées et créées à des fins différentes par les machines et les réseaux sociaux. Le volume de données à analyser est une énorme quantité de données. Récemment, les sources de génération de données sont augmentées et cela provoque une diversité de données telles que le texte, la vidéo, l'audio et les bases de données. Afin de traiter l'énorme quantité de données, le traitement des données conventionnelles doit être amélioré.

1.2.3 Du modèle 3-V au modèle multi-V

Les données volumineuses signifient initialement le volume, la vitesse et la variété des données qui deviennent difficiles à analyser en utilisant des techniques et des plates-formes de traitement de données conventionnelles. De nos jours, les sources de production des données sont améliorées rapidement, telles que les réseaux de capteurs, les instruments à haut débit et les machines de streaming, ces environnements génèrent une quantité massive de données.

De nos jours, les méga-données jouent un rôle crucial dans divers environnements tels que l'industrie, la recherche scientifique, la gestion des ressources naturelles et les réseaux sociaux. Le modèle 3-V ne convient plus pour montrer les caractéristiques du Big Data. Par conséquent, un modèle multi-V est adopté.

Vélocité, Variété et Volume : Déjà définis dans la section 1.1.2.

Valeur : La valeur des données représente bien le Big Data. Ayant des quantités continues de données n'est utile que lorsqu'elles peuvent être transformées en valeur. Il est essentiel de comprendre que cela ne signifie pas que le Big Data a toujours de la valeur.

Véracité : Elle représente la compréhensibilité des données non sa qualité.

Validité : Il est essentiel de s'assurer que les données sont précises et exactes pour l'utilisation d'avenir. Afin de prendre les bonnes décisions à l'avenir, les organisations doivent valider les données sensiblement.

Variabilité : Elle fait référence à la cohérence et à la valeur des données.

Viscosité : C'est un élément de la vitesse, il représente la latence ou le temps de latence dans le transmission de données entre la source et la destination.

Viralité : Elle représente la vitesse d'envoi des données.

Visualisation : Elle est utilisée pour symboliser le Big Data dans une vue complète et pour déterminer les valeurs cachées. La visualisation est une clé essentielle pour faire du Big Data un partie intégrante de la prise de décision.

1.3 Étude de l'existant

Dans cette section, nous établissons une étude comparative entre les outils existants et déduisons à la fin les meilleurs outils qui peuvent être adoptés dans notre projet.

1.3.1 Outils existants

Nous présentons, dans cette sous-section, les avantages et les inconvénients des outils existants pour enfin choisir le meilleur outil à adopter.

1.3.1.1 Systèmes de gestion de base de données relationnelle

Drizzle : C'est une fourche MySQL avec un micro-noyau enfichable et plus performant. Comme MySQL, Drizzle a une architecture client / serveur et utilise SQL comme langage de commande principal.

Avantages	Inconvénients
+ Meilleure performance dans les machines modernes + Peut exécuter plus de 4 threads matériels à la fois + Peut utiliser du SQL dynamique à instructions multiples. Via le mot clé CONCURRENT, il peut exploiter ces déclarations en parallèle	- Les bases de données relationnelles ne sont pas une solution appropriée pour la gestion et l'analyse de données à grande échelle

TABLE 1.1 – Avantages et inconvénients de Drizzle

1.3.1.2 Systèmes de gestion de base de données non relationnelle

Cassandra : Apache Cassandra est un logiciel de gestion de base de données NoSQL distribué gratuit et à code source ouvert. C'est un système conçu pour traiter de grandes quantités de données sur de nombreux serveurs, offrant ainsi une haute disponibilité sans point de défaillance unique. Cassandra offre un support robuste pour clusters couvrant plusieurs centres de données, avec une réplication asynchrone sans maître permettant des opérations à faible temps de latence pour tous les clients.

Avantages	Inconvénients
+ Vitesse d'écriture + Cohérence ajustable : Cassandra permet, à base de requête-sur-requête, de décider comment gérer les problèmes potentiels. + Basé sur la JVM, cela signifie que Cassandra peut facilement s'intégrer avec d'autres applications basées sur JVM. + CQL : c'est une manière familière d'interroger Cassandra, en faisant la transition d'un SGBDR basé sur SQL vers Cassandra moins choquante.	- La couche de stockage de données Cassandra est un système de stockage à valeur-clé. Donc, les données doivent être modélisées autour des requêtes, plutôt qu'autour de la structure des données elles-mêmes - Pas d'agrégations - Performances imprévisibles car aucun utilisateur n'a programmé les tâches - La gestion de la mémoire se fait par le langage lui-même, pas l'application.

TABLE 1.2 – Avantages et inconvénients de Cassandra

MongoDB : C'est une base de données NoSQL open source qui utilise un modèle de données orienté document. Au lieu d'utiliser des tables et des lignes comme dans les bases de données relationnelles, MongoDB repose sur une architecture de collection et de documents. Les documents comprennent des ensembles de paires clé-valeur et ils sont les unité de base de données dans MongoDB. La collection contient des ensembles de documents et fonctionne comme équivalent des tables de base de données relationnelles.

Avantages	Inconvénients
+ Evolutivité horizontale + Si le serveur principal tombe en panne, le serveur secondaire devient le serveur principal sans aucune intervention humaine + Il supporte les mécanismes d'authentification communs, tels que LDAP, AD et les certificats. Les utilisateurs peuvent connecter MongoDB sur SSL et les données peuvent être chiffrées + MongoDB peut être une solution rentable car il améliore la flexibilité et réduit les coûts sur le matériel et l'espace de rangement	- Aucune transaction de support - Aucune jointure - Limite de mémoire

TABLE 1.3 – Avantages et inconvénients de MongoDB

1.3.2 Spécifications techniques

Comme nous avons, dans notre cas, une énorme quantité de données non structurées, le cadre le plus approprié pour collecter ces données en temps réel est Apache Kafka, car il fournit une riche bibliothèque et des fonctionnalités qui peuvent être adaptées à notre travail. Apache Kafka est bien effectué avec le langage de programmation Python. En fait, Python fournit une variété d'outils et de nombreuses bibliothèques pour l'exploration de données. Le stockage des données a également besoin d'un outil performant pour bien conserver une grande quantité de données afin de les traiter en toute sécurité, qui est MongoDB et qui correspond parfaitement à Apache Kafka.

Conclusion

Tout au long de ce chapitre, nous présentons brièvement l'organisme d'accueil, expliquons certains concepts théoriques essentiels à la compréhension de ce qui suit et établissons également une étude sur laquelle nous comparons les outils existants pour le Big Data. Sur la base des solutions étudiées, nous sommes capables de définir les exigences de notre projet, c'est le but du chapitre à venir.

Chapitre 2

Spécification des besoins et conception

Définir correctement les exigences que notre travail est invité à satisfaire est l'un des plus importantes étapes du développement de notre projet. Au cours de ce chapitre, nous analysons les besoins fonctionnels et non fonctionnels. Ensuite, nous définissons le comportement dynamique du système, en introduisant l'interaction entre ses acteurs et l'ensemble du système.

À travers les résultats de la première partie couvrant l'analyse et la spécification des besoins, nous pouvons commencer la phase de conception car c'est une étape cruciale et nous visons à entreprendre et à préparer le terrain pour l'étape de réalisation. Dans cette deuxième partie, nous présentons notre approche conceptuelle ainsi que des arguments pour ce choix. Ensuite, nous expliquons la conception détaillée.

2.1 Analyse et spécification des besoins

Dans cette section, nous spécifions les besoins fonctionnels et non fonctionnels du système à travers le diagramme de cas d'utilisation.

2.1.1 Analyse des besoins

L'analyse des besoins consiste à définir les différents besoins fonctionnels qui décrivent les fonctionnalités ou les services que peut offrir le système. Elle contient aussi les besoins non fonctionnels qui sont des contraintes à respecter pour tout le système.

2.1.1.1 Identification des acteurs

Un acteur est un rôle joué par une entité externe qui interagit directement avec le système étudié. Dans notre cas, tout utilisateur peut être un acteur externes directement associé à notre système.

2.1.1.2 Besoins fonctionnels

L'application livrée consiste à concevoir et à mettre en œuvre une application qui doit satisfaire les fonctionnalités spécifiques suivantes :

- Collecter les données pertinentes des personnes ou des produits, à partir des différents sites internet (site de commerce, réseaux sociaux, forum, blog, etc).
- Analyser ces données pour sortir des informations utiles à la prédiction.

2.1.1.3 Besoins non fonctionnels

Tout en répondant aux exigences fonctionnelles, la solution doit satisfaire aux besoins non fonctionnels suivants :

Evolutivité : La modularité et l'extensibilité de l'architecture de l'application en cas de l'ajout de nouvelles fonctionnalités ou de nouveaux sites internet. Dans notre cas, lorsque la taille ou le volume des données deviennent plus grands, le collecteur doit garantir un comportement régulier.

Facilité d'utilisation : Les utilisateurs n'appartenant pratiquement pas à l'informatique, le collecteur doit une interface conviviale et ergonomique.

Sécurité : Le stockage des données devrait être loin d'être accédé par un utilisateur non désiré. Chaque accès doit être précédé d'une authentification.

Documentation : La solution doit être bien documentée afin de fournir la meilleure utilisation pour le client.

2.1.2 Spécification

Pour mieux comprendre les besoins requis, nous les modélisons de manière formelle en utilisant le diagramme de cas d'utilisation qui capture le comportement d'un système logiciel. Ce diagramme décrit les différentes actions pouvant être effectuées par un utilisateur extérieur. La figure 2.1 présente le diagramme de cas d'utilisation du Smart Data Collector montrant que l'utilisateur peut collecter les données pertinentes des différents sites internet soit des personnes ou des produits. A partir de l'analyse de ces données, il peut enfin sortir des informations utiles à la prédiction.

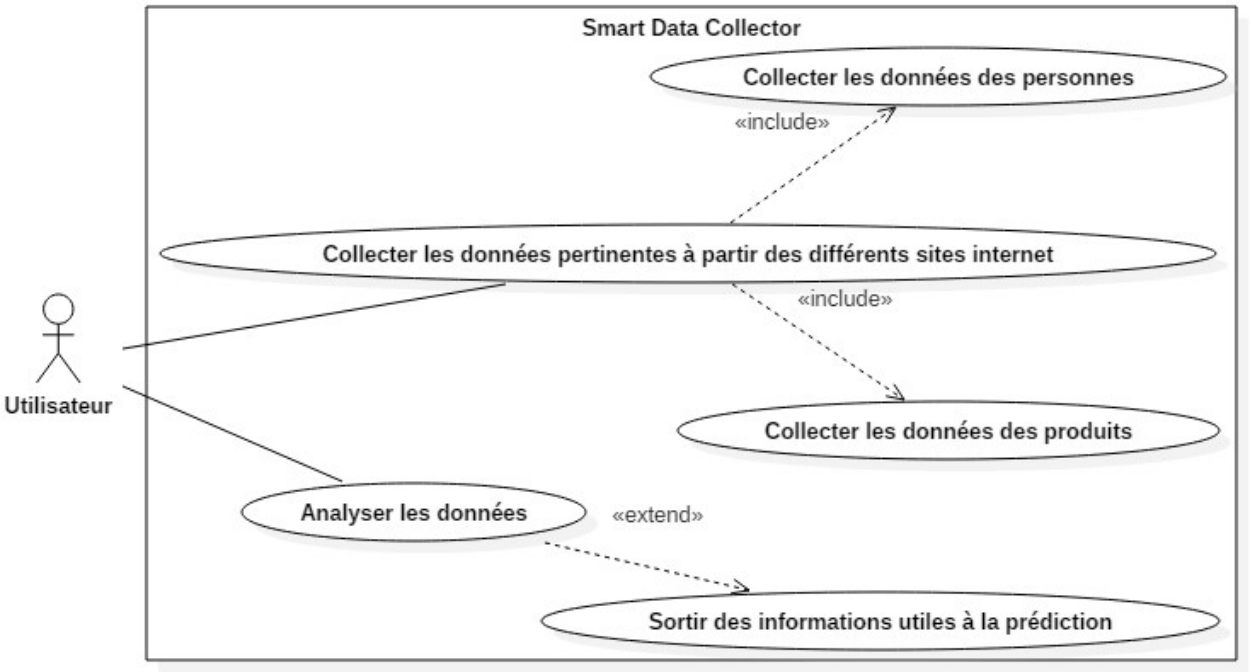


FIGURE 2.1 – Diagramme de cas d'utilisation

2.2 Conception

Dans cette section, nous présentons la conception globale et détaillée du Smart Data Collector.

2.2.1 Conception globale

Nous nous intéressons maintenant aux architectures matérielle et logicielle que nous avons choisies pour notre système.

2.2.1.1 Architecture physique pipeline

La figure 2.2 montre l'architecture physique du système. Il comprend les différents composants matériels qui construisent notre infrastructure d'applications.

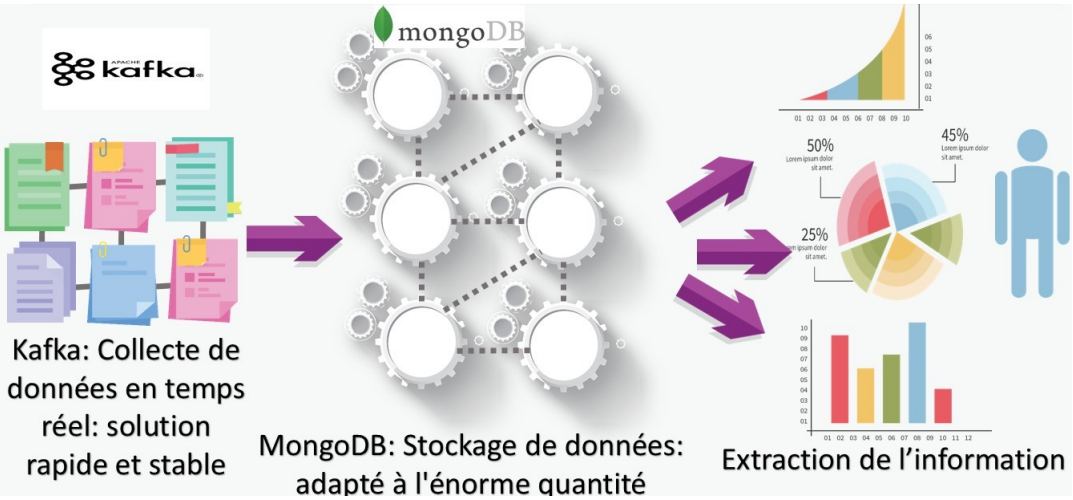


FIGURE 2.2 – Architecture pipeline

L'architecture physique du Smart Data Collector est une pipeline car la sortie d'une phase est l'entrée de l'autre. Tout d'abord, le système collecte les données du LinkedIn, et à partir du champ des noms, il collecte les données de Facebook et les tweets de Twitter. Ensuite, il les fusionne pour les intégrer dans Apache Kafka qui collecte l'énorme quantité de données en temps réel pour les stocker dans MongoDB.

2.2.1.2 Architecture logique 2-couches

L'architecture logique interne du collecteur est structurée autour de 2 couches principales comme le montre la figure 2.3.

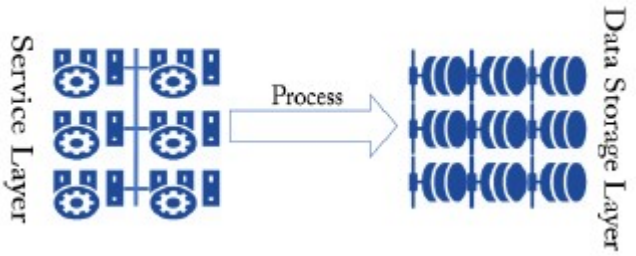


FIGURE 2.3 – Architecture 2-couches

Couche Service : C'est le module qui exécute les différents composants de notre noyau d'application.

Couche de stockage de données : C'est la couche de base de données contenant les informations où les fichiers sont gérés via MongoDB.

2.2.2 Conception détaillée

La figure 2.4 montre le diagramme de classes de notre solution proposée.

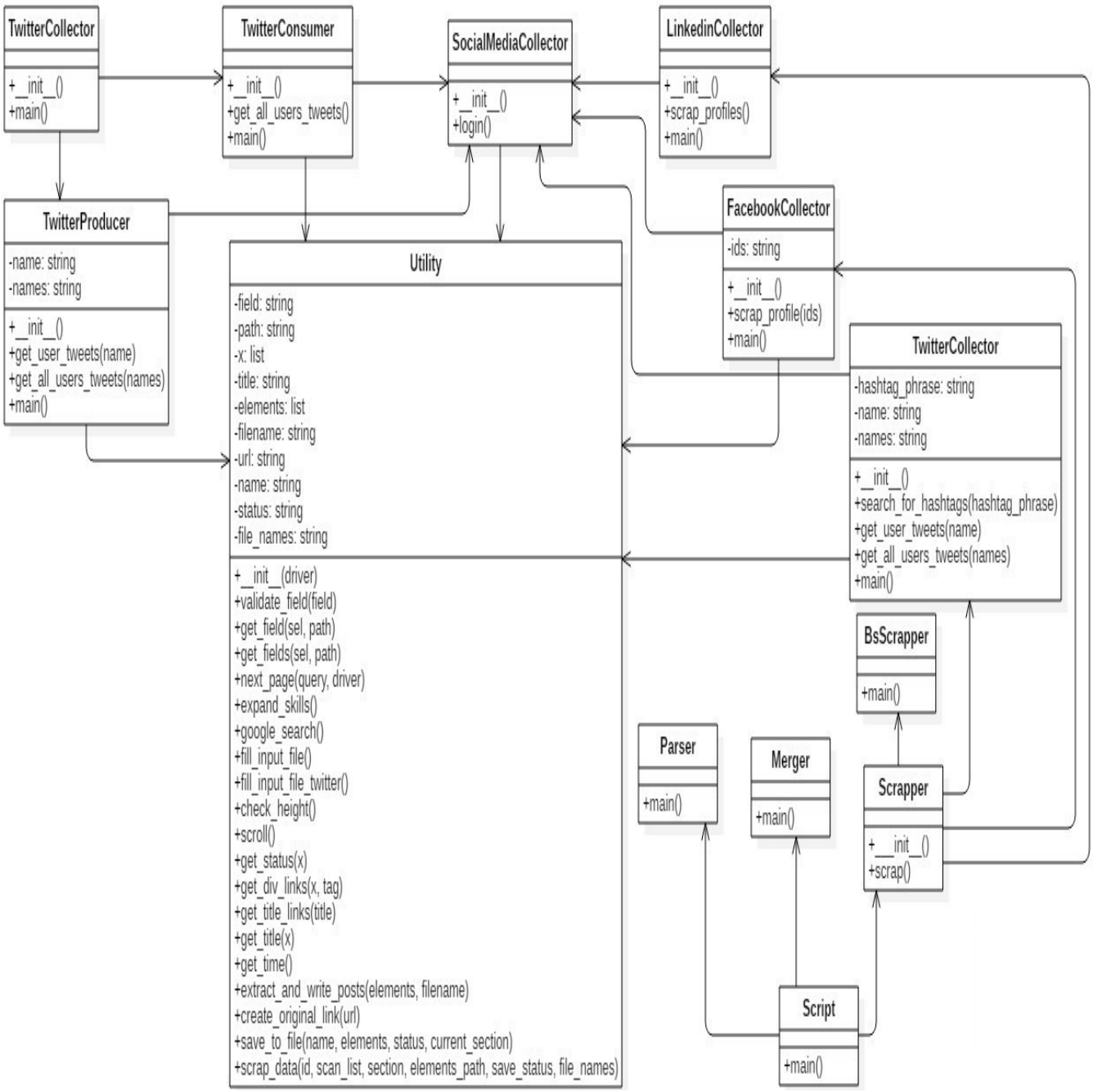


FIGURE 2.4 – Diagramme de classes

Couche Service : Cette couche contient essentiellement les 2 classes : "SocialMeediaCollector" qui permet à l'utilisateur de se connecter pour collecter les données, et "Utility" qui contient les différentes fonctions utiles à la collecte.

Couche stockage de données : Cette couche contient les classes de collecte de données à partir des sites internet à l'aide de la bibliothèque BeautifulSoup et des réseaux sociaux : LinkedIn, Facebook et Twitter.

Conclusion

Dans ce chapitre, nous présentons les exigences et les besoins du collecteur. D'abord, nous identifions les acteurs de notre système et expliquons les besoins fonctionnels et non fonctionnels. Ensuite, nous détaillons ces besoins en spécifiant plus de détails via le diagramme de cas d'utilisation.

Ainsi, nous commençons la phase de conception où nous donnons une vue d'ensemble de notre solution avec son architecture physique et logique. Ensuite, nous fournissons le diagramme de classes pour donner plus de détails sur la solution et l'interaction entre ses différents éléments. Le chapitre suivant porte sur la mise en œuvre des résultats obtenus.

Chapitre 3

Réalisation du SDC

Ce chapitre nous permet de présenter le travail réalisé. Nous commençons par énumérer les technologies et l'environnement logiciel utilisés tout au long du cycle de vie du projet. Après, nous présentons un aperçu des résultats obtenus et des captures d'écran de différentes interfaces de notre collecteur. Nous terminons ce chapitre en montrant les tests et la validation.

3.1 Environnement de travail

Dans cette partie nous présentons les différents outils matériel et logiciel nécessaires pour le développement de notre système.

3.1.1 Environnement matériel

Pour réaliser ce travail, nous avons utilisé un ordinateur personnel présentant les caractéristiques suivantes :

Marque : Ordinateur portable Asus
Processeur : processeur Intel Core i7
Mémoire : 8,00Go de RAM
Système d'exploitation : Ubuntu 18.04

3.1.2 Environnement logiciel

Pour réaliser ce travail, nous avons utilisé les logiciels suivants :
StarUML : C'est un logiciel de modélisation UML.

Latex : C'est un logiciel de composition de documents.

Spyder : C'est un environnement de développement intégré (IDE) multi plateformes à source ouverte, utilisé pour la programmation en langage Python. Spyder intègre NumPy, SciPy, Matplotlib et IPython, ainsi que d'autres bibliothèques open source. Spyder est extensible avec les plugins, inclut le support d'outils interactifs pour la visualisation de données. Il est disponible sur plusieurs plates-formes via Anaconda.

Anaconda : C'est une distribution freemium open source de la programmation Python et R, deux langages pour le traitement de données à grande échelle et l'analyse prédictive, qui vise à simplifier la gestion et le déploiement des paquets dont les versions sont gérées par le système de gestion de paquets Conda.

3.1.3 Choix techniques

Python :
C'est un puissant langage de programmation orienté objet de haut niveau. Python est un langage interprété, il a une philosophie de conception qui met l'accent sur la lisibilité du code par une indentation d'espace et une syntaxe permettant aux programmeurs d'exprimer des concepts en quelques lignes de code.

Sélénium WebDriver :
Le changement le plus important intervenu récemment dans Selenium est l'inclusion de l'API WebDriver. C'est un outil qui marque un progrès décisif en termes d'automatisation du navigateur Web c'est-à-dire utiliser un navigateur de manière native, comme un utilisateur le fait localement ou sur une machine distante à l'aide du serveur Selenium. Le classe d'implémentation choisi dans ce projet est ChromeDriver.

TwitterAPI :
Il nous permet d'accéder aux fonctionnalités de Twitter sans passer par l'interface du site Web. C'est utile pour poster des tweets ou envoyer des messages dirigés de manière automatisée avec des scripts, ou bien pour trouver des tweets contenant un mot.

3.2 Manuel d'utilisation de l'application réalisée

Notre SDC est simple à utiliser, il suffit d'exécuter le fichier script.py pour que l'application commence à collecter les données à partir du LinkedIn selon les noms des personnes précisés préalablement dans un fichier .txt. A partir de ce champ Nom, elle commence à collecter les données de Facebook et les tweets de Twitter. Ensuite, le fichier merger.py fusionne toutes ces données dans un seul fichier output.csv.

3.2.1 Données extraites du LinkedIn

Les données extraites à partir du LinkedIn sont celles illustrées par les figures allant de 3.1 jusqu'à 3.4.

	A	B	C
1	Name	Job Title	Company
2	Chtioui Malek	Senior developper Team lead chez Alphalyr	Alphalyr
3	Miled OTHMEN	.NET Research & Development Engineer at Linedata	Linedata
4	Athil Belhadj	Full-stack Developer(Angular & .NET) at Poulina Group Holding	Poulina Group Holding
5	Nada Ghedamsi	Business Developer at Valomnia	Valomnia
6	Mohamed Ouederni	Software Engineer	Self-Employed
7	Rim Hmani	iOS Developer	USERADGENTS
8	Malek Boubakri	CTO at Optimal Solution ltd. & Full-Stack Developer	Optimal solution s.a.r.l
9	Houssem Abid	Python & Odoo developer	CliniSys
10	Elyes Labidi	Java/J2EE Junior Developer	Linedata
11	Hichem HAMDAOUI	Frontend Developer at Codix	Codix
12	Tarek Hammami	UX/UI Designer & Developer	think tank Business Solutions
13	kais chebbi	Senior Software Development Engineer chez Crédit Agricole CIB	Crédit Agricole CIB
14	Karim Gharbi	Responsable technique à Orange developer center	Orange Tunisie
15	Mohamed Boujnah	Full-Stack JS developer	Mobelite
16	walid sassi	Assistant professor chez faculté des sciences de Bizerte	Mobile Powered
17	Chaker Ben Hammouda	PHP Backend Developer - Freelance/Remote	BCM Energy
18	Hamza Dhahri	★ Full-Stack JAVA-JEE/Angular Developer ✨	VERMEG for Banking & Insurance Software
19	Amir Ben Haj Khaled	Software Engineer, Game Developer, Unity C# Developer.	Idealump Tunisia
20	Mohamed Masmoudi	Typo3 Web Developer Freelancer :: Intégrateur Typo3 en Freelance	Offshore web development.
21	Haithem Hamzaoui	Software Developer chez VERMEG for Banking & Insurance Software	VERMEG for Banking & Insurance Software
22	Fouèd MOUSSI	PHP Developer at 8 WAYS MEDIA TUNISIA & Founder of monostore.tn	8 WAYS MEDIA TUNISIA
23	Adam Bedoui	Fullstack Java EE/Angular Developer chez AYMAX	AYMAX
24	Omar THAMRI	iOS Developer at Mobile Powered	Mobile Powered
25	Ahmed Amine Mejri	Fullstack Developer	IRIS TUNIS
26	Ines Trabelsi	Developer Analyst COBOL at IID- International Information Development	IID- International Information Development (Euro
27	Mohamed MEJRI	Analyst Developer chez Vermeg	VERMEG for Banking & Insurance Software
28	Adnen rebai	Senior frontend developer	No result
29	Mohamed Amine Karoui	Sharepoint Developer	Tenstep-EPM-Tunisia

output

FIGURE 3.1 – Nom, profession et entreprise

	D	E	F
1	College	Location	skills
2	Faculté Droit et Science Economique De Sousse	Gouvernorat de Tunis, Tunisia	C#,ASP.NET,ASP.NET MVC
3	TEK-UP	Gouvernorat de l'Ariana, Tunisia	HTML5,C#,C
4	ISETN	Kelibia, Gouvernorat de Nabeul, Tunisia	Android Development,Programmation orientée objet,Java
5	Faculté des Lettres et Sciences Humaines de Sousse	Gouvernorat de Sousse, Tunisia	Management,Microsoft Excel,Microsoft Office
6	No result	Gouvernorat de Tunis, Tunisia	
7	ENSI - Ecole Nationale des Sciences de l'Informatique	Tunisie	iOS,Objective C,Swift
8	Higher Institute of Computer Science and Communication Technologies of Hammam Sousse	Sousse Jawhara, Gouvernorat de Sousse, Tunisia	
9	Institut Privé Polytechnique des Sciences Avancées de Sfax (IPSAS)	Gouvernorat de Sfax, Tunisia	Crystal Reports,Entity Framework,Microsoft SQL Server
10	ISI	Gouvernorat de la Manouba, Tunisia	
11	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Gouvernorat de Tunis, Tunisia	
12	Institut Superieur des Arts Multimedia de la Manouba	Gouvernorat de Tunis, Tunisia	UI/UX Design,HTML5,Design
13	Institut Supérieur d'Informatique ARIANA-Tunisia	Gouvernorat de l'Ariana, Tunisia	
14	--	Gouvernorat de Tunis, Tunisia	Web Project Management,Mobile Applications,Gestion de proje
15	No result	Gouvernorat de Sousse, Tunisia	JavaScript,React.js,angular
16	Faculty of sciences Bizerte	Menzel Djemil, Gouvernorat de Bizerte, Tunisia	
17	Ecole Supérieure de Communications de Tunis (Sup'Com)	Gouvernorat de Médenine, Tunisia	Linux,MySQL,Matlab
18	ENSI - Ecole Nationale des Sciences de l'Informatique	Mannouba, Gouvernorat de la Manouba, Tunisia	
19	Higher Institute of Applied Sciences and Technology of Sousse	Ariana Médina, Gouvernorat de l'Ariana, Tunisia	
20	ISIMS	Gouvernorat de Sfax, Tunisia	
21	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Kelibia, Gouvernorat de Nabeul, Tunisia	Framework Symfony,Applications web,Applications mobiles
22	Institut Supérieure Des Etudes Technologiques De Charguia	Gouvernorat de Tunis, Tunisia	
23	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Gouvernorat de Sfax, Tunisia	Java Enterprise Edition,Angular,Android
24	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Tunisie	iOS,Android,PHP
25	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Gouvernorat de Tunis, Tunisia	JavaEE,Primefaces,Full Stack
26	ENSIT	Gouvernorat de Sousse, Tunisia	
27	Ecole Supérieure Privée d'Ingénierie et de Technologies - ESPRIT	Gouvernorat de Tunis, Tunisia	UML,Java,JEE
28	No result	Gouvernorat de Tunis, Tunisia	
29	Sciences faculty of Bizerte	Gouvernorat de Bizerte, Tunisia	

output

FIGURE 3.2 – Université, emplacement et compétences

	G
1	WorkPlaces and Organization
2	Alphalyr,HADRUM,Monoprog,Xcess Company,.....,Stack Overflow,New Web Technologies (NWT
3	Linedata,TEK-UP,ASSURIA,TEK-UP,IntilaQ Tunisia,.....
4	Poulina Group Holding,AthilTech,PROTOL,STT,searching for new opportunity,.....
5	Valomnia,BeanAir,The Carter Center,AMIDEAST,TAV Airports,.....,Positifs
6	Self-Employed,IP-TECH,.....
7	USERADGENTS,NORRIQ,Bluewave Tunisie,,,
8	Free,SBC,.....
9	CliniSys,Faculté des Sciences Economiques et de Gestion de Sfax (FSEGS),.....
10	Linedata,SFM Telecom-SFM Technologies,FormaLab,tn,Elmenus,Zitouna Bank,.....,Smart Kids,AIES
11	Codix,Octopouce Digital Ltd,Pi2R,Digit-U,Skopeo,.....,JCI (Junior Chamber Internatio
12	think tank Business Solutions,appsnsites,Favizone,Free - lance,.....
13	Crédit Agricole CIB,Web Agency Tozeur,.....,Microsoft
14	Jet Multimedia,ANORIA,Karoui & Karoui Interactive (Nessma TV),ADHARA / NET CONCEPT
15	Freelance,Ecole Pluridisciplinaire Internationale EPI Sousse,TUNINK,CMS,.....
16	Mobile Powered,faculté des sciences de Bizerte,Ministry of education,Applicaphone,Softs4mobiles,,,
17	BCM Energy,Baby Sittor,CodinTek,.....
18	axefinance,Ooredoo Tunisie,.....,Netlinks,AIESEC Local Committee University, Tunisia
19	Idealump Tunisia,.....
20	Offshore web development,,Web Development,SOFT-TO-DO - Sfax TUNISIA,ImTech Communicatior
21	VERMEG for Banking & Insurance Software,.....,لغة و أحة
22	8 WAYS MEDIA TUNISIA,Tchesy,Bulldozer,progres formation,eMobile Technologies,.....
23	AYMAX,Freelance (à mon compte),ESPRIT (Ecole Supérieure Privée d'Ingénierie et de Technologies
24	Mobile Powered,ESPRIT,CHIFCO,déparement informatique CNSS,cni
25	IRIS TUNIS,APP4MOB,HP France,Octasoft,SFR,.....
26	IID- International Information Development (Euro Information),Sagemcom Tunisia,.....,Netlinks
27	VERMEG for Banking & Insurance Software,axefinance,Société Générale,Setecstra,Iprecision,.....
28	ALLOcloud,Novisdev,Military Technical Training Center,Sagemcom,.....
29	Tenstep-EPM-Tunisia,.....

output

FIGURE 3.3 – Lieux de travail et organisation

	H
1	URL
2	https://www.linkedin.com/in/chtiwi/
3	https://www.linkedin.com/in/miled-othmen-004902ba/
4	https://www.linkedin.com/in/athil-belhadj-a98691ab/
5	https://www.linkedin.com/in/nadagh/
6	https://www.linkedin.com/in/medouederni/?locale=en_US
7	https://www.linkedin.com/in/rimhmani/
8	https://www.linkedin.com/in/malekboubakri/?locale=en_US
9	https://www.linkedin.com/in/houssem-abid/?locale=en_US
10	https://www.linkedin.com/in/elyeslabidi/?originalSubdomain=tn
11	https://www.linkedin.com/in/hichemhamdaoui/?originalSubdomain=tn
12	https://www.linkedin.com/in/tarek-hammami-35878312b/?originalSubdomain=tn
13	https://www.linkedin.com/in/chebbikais/?originalSubdomain=tn
14	https://www.linkedin.com/in/karim-gharbi-3b543310/?originalSubdomain=tn
15	https://www.linkedin.com/in/mohamed-boujnah-599b1a10b/?originalSubdomain=tn
16	https://www.linkedin.com/in/sassi-walid/?originalSubdomain=tn
17	https://www.linkedin.com/in/benhammoudachaker/?originalSubdomain=tn
18	https://www.linkedin.com/in/hamza-dhahri/?originalSubdomain=tn
19	https://www.linkedin.com/in/amirbhk/?locale=en_US
20	https://www.linkedin.com/in/mohamedmasmoudi/?originalSubdomain=tn
21	https://www.linkedin.com/in/haithem-hamzaoui-580187117/?originalSubdomain=tn
22	https://www.linkedin.com/in/foued-moussi/?originalSubdomain=tn
23	https://www.linkedin.com/in/adam-bedoui-337b3a42/?originalSubdomain=tn
24	https://www.linkedin.com/in/omar-thamri-462039b6/?originalSubdomain=tn
25	https://www.linkedin.com/in/ahmedaminemejri/?originalSubdomain=tn
26	https://www.linkedin.com/in/ines-trabelsi-94b66587/?originalSubdomain=tn
27	https://www.linkedin.com/in/mohamed-mejri-b81580101/?originalSubdomain=tn
28	https://www.linkedin.com/in/adnen-rebai-65843263/?originalSubdomain=tn
29	https://www.linkedin.com/in/mohamed-amine-karoui/?originalSubdomain=tn

output

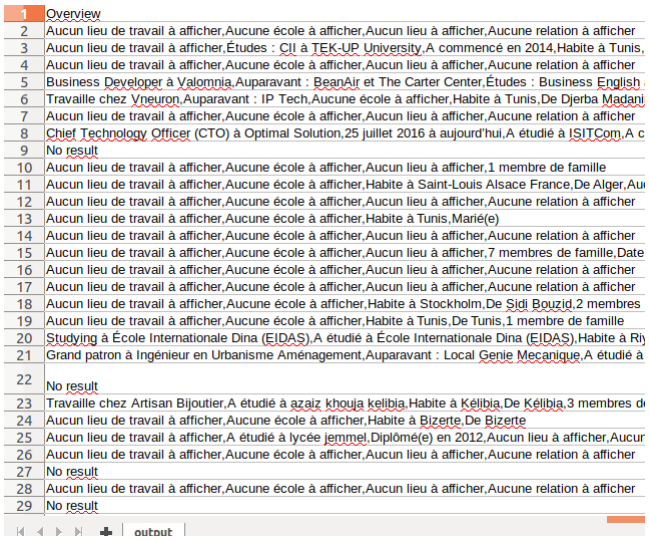
FIGURE 3.4 – Url

3.2.2 Données extraites du Facebook

Les données extraites à partir du Facebook à l'aide de Selenium WebDriver sont celles illustrées par les figures allant de 3.5 jusqu'à 3.12.

3.2.3 Données extraites du Twitter

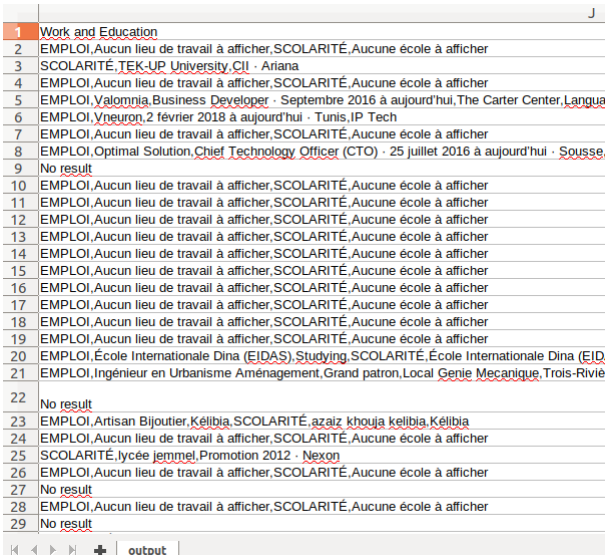
Les tweets extraits à partir du Twitter en utilisant TwitterAPI sont ceux illustrées par la figure 3.13. Ensuite, nous collectons des tweets en temps réel en intégrant Apache Kafka et nous les stockons dans la base de données de MongoDB à l'aide du résultat final qui est sous forme json comme le montre la figure 3.14.



	Overview
2	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
3	Aucun lieu de travail à afficher,Études : CII à TEK-UP University,A commencé en 2014,Habite à Tunis,
4	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
5	Business Developer à Valomnia,Auparavant : BeanAir et The Carter Center,Études : Business English
6	Travaille chez Vneuron,Auparavant : IP Tech,Aucune école à afficher,Habite à Tunis,De Djerba Madani
7	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
8	Chief Technology Officer (CTO) à Optimal Solution,25 juillet 2016 à aujourd'hui,A étudié à ISITCom,A c
9	No result
10	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,1 membre de famille
11	Aucun lieu de travail à afficher,Aucune école à afficher,Habite à Saint-Louis Alsace France,De Alger,Au
12	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
13	Aucun lieu de travail à afficher,Aucune école à afficher,Habite à Tunis,Marié(e)
14	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
15	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,7 membres de famille,Date
16	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
17	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
18	Aucun lieu de travail à afficher,Aucune école à afficher,Habite à Stockholm,De Sidi Bouzid,2 membres
19	Aucun lieu de travail à afficher,Aucune école à afficher,Habite à Tunis,De Tunis,1 membre de famille
20	Studying à École Internationale Dina (EIDAS),A étudié à École Internationale Dina (EIDAS),Habite à Ri
21	Grand patron à Ingénieur en Urbanisme Aménagement,Auparavant : Local Genie Mecanique,A étudié à
22	No result
23	Travaille chez Artisan Bijoutier,A étudié à azaiz khouja kelibia,Habite à Kélibia,De Kélibia,3 membres d
24	Aucun lieu de travail à afficher,Aucune école à afficher,Habite à Bizerte,De Bizerte
25	Aucun lieu de travail à afficher,A étudié à lycée jemmel,Diplômé(e) en 2012,Aucun lieu à afficher,Aucur
26	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
27	No result
28	Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à afficher,Aucune relation à afficher
29	No result

output

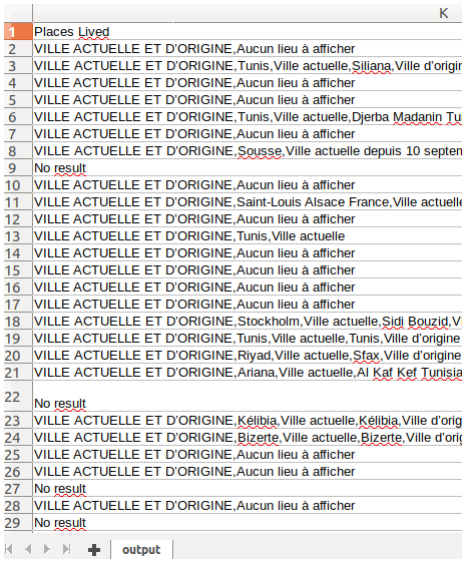
FIGURE 3.5 – Vue d'ensemble



	J
1	Work and Education
2	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
3	SCOLARITÉ,TEK-UP University,CII · Ariana
4	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
5	EMPLOI,Valomnia,Business Developer · Septembre 2016 à aujourd'hui,The Carter Center,Langua
6	EMPLOI,Vneuron,2 février 2018 à aujourd'hui · Tunis,IP Tech
7	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
8	EMPLOI,Optimal Solution,Chief Technology Officer (CTO) · 25 juillet 2016 à aujourd'hui · Sousse,
9	No result
10	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
11	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
12	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
13	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
14	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
15	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
16	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
17	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
18	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
19	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
20	EMPLOI,École Internationale Dina (EIDAS),Studying,SCOLARITÉ,École Internationale Dina (EID
21	EMPLOI,Ingénieur en Urbanisme Aménagement,Grand patron,Local Genie Mecanique,Trois-Rivi
22	No result
23	EMPLOI,Artisan Bijoutier,Kélibia,SCOLARITÉ,azaiz khouja kelibia,Kélibia
24	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
25	SCOLARITÉ,lycée jemmel,Promotion 2012 · Nexon
26	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
27	No result
28	EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à afficher
29	No result

output

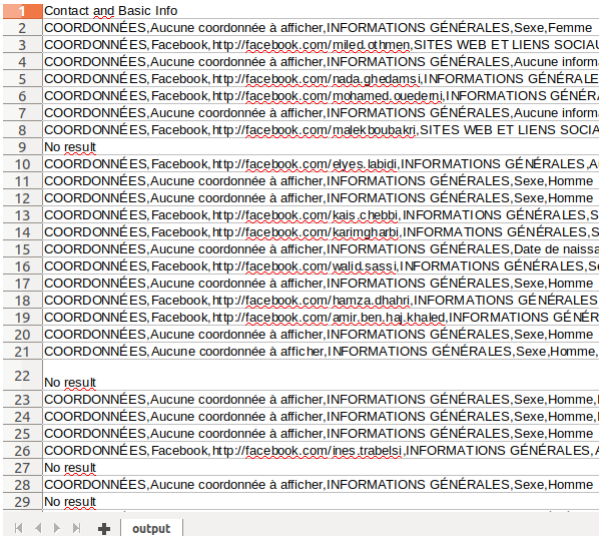
FIGURE 3.6 – Travail et éducation



	K
1	Places Lived
2	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
3	VILLE ACTUELLE ET D'ORIGINE,Tunis,Ville actuelle,Siliana,Ville d'origir
4	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
5	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
6	VILLE ACTUELLE ET D'ORIGINE,Tunis,Ville actuelle,Djerba Madanin Tu
7	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
8	VILLE ACTUELLE ET D'ORIGINE,Sousse,Ville actuelle depuis 10 septen
9	No result
10	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
11	VILLE ACTUELLE ET D'ORIGINE,Saint-Louis Alsace France,Ville actuell
12	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
13	VILLE ACTUELLE ET D'ORIGINE,Tunis,Ville actuelle
14	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
15	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
16	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
17	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
18	VILLE ACTUELLE ET D'ORIGINE,Stockholm,Ville actuelle,Sidi Bouzid,V
19	VILLE ACTUELLE ET D'ORIGINE,Tunis,Ville actuelle,Tunis,Ville d'origine
20	VILLE ACTUELLE ET D'ORIGINE,Riyad,Ville actuelle,Sfax,Ville d'origine
21	VILLE ACTUELLE ET D'ORIGINE,Ariana,Ville actuelle,Al Kaf Kef Tunisia
22	No result
23	VILLE ACTUELLE ET D'ORIGINE,Kélibia,Ville actuelle,Kélibia,Ville d'orig
24	VILLE ACTUELLE ET D'ORIGINE,Bizerte,Ville actuelle,Bizerte,Ville d'ori
25	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
26	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
27	No result
28	VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher
29	No result

output

FIGURE 3.7 – Lieux habités



	Contact and Basic Info
2	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Femme
3	COORDONNÉES,Facebook,http://facebook.com/miled.othmen,SITES WEB ET LIENS SOCIAU
4	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Aucune inform
5	COORDONNÉES,Facebook,http://facebook.com/nada.ghedamsi,INFORMATIONS GÉNÉRALE
6	COORDONNÉES,Facebook,http://facebook.com/mohamed.ouedemi,INFORMATIONS GÉNÉR
7	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Aucune inform
8	COORDONNÉES,Facebook,http://facebook.com/malek.boubakri,SITES WEB ET LIENS SOCIA
9	No result
10	COORDONNÉES,Facebook,http://facebook.com/elyes.labidi,INFORMATIONS GÉNÉRALES,A
11	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
12	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
13	COORDONNÉES,Facebook,http://facebook.com/kais.chebbi,INFORMATIONS GÉNÉRALES,S
14	COORDONNÉES,Facebook,http://facebook.com/karimgharbi,INFORMATIONS GÉNÉRALES,S
15	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Date de naissa
16	COORDONNÉES,Facebook,http://facebook.com/walid.sassi,INFORMATIONS GÉNÉRALES,S
17	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
18	COORDONNÉES,Facebook,http://facebook.com/hamza.dhahri,INFORMATIONS GÉNÉRALES
19	COORDONNÉES,Facebook,http://facebook.com/amir.ben.haj.khaled,INFORMATIONS GÉNÉR
20	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
21	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme,
22	No result
23	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme,I
24	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme,I
25	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
26	COORDONNÉES,Facebook,http://facebook.com/ines.trabelsi,INFORMATIONS GÉNÉRALES,A
27	No result
28	COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES,Sexe,Homme
29	No result

output

FIGURE 3.8 – Contact et informations de base

	O
1	Life Events
2	ÉVÈNEMENTS MARQUANTS
3	ÉVÈNEMENTS MARQUANTS,2014,A commencé à étudier à TEK-UP University,2013,Déménagement à Ariana
4	ÉVÈNEMENTS MARQUANTS
5	ÉVÈNEMENTS MARQUANTS,2016,A commencé à travailler chez Valomnia,A quitté son poste à BeanAir,2014,A travaillé à The Carter Center,A travaillé à American Corner Sousse,A commenc
6	ÉVÈNEMENTS MARQUANTS,2018,A commencé à travailler chez Vneuron
7	ÉVÈNEMENTS MARQUANTS
8	ÉVÈNEMENTS MARQUANTS,2016,A commencé à travailler chez Optimal Solution,2012,A commencé à étudier à ISITCom,Voyage à Sousse,Déménagement à Sousse,En couple avec May'sa
9	No result

FIGURE 3.9 – Evènements de la vie

1	Family and Relationships
2	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
3	SITUATION AMOUREUSE,Célibataire,MEMBRES DE LA FAMILLE,Aucun membre de famille à affich
4	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
5	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
6	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
7	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
8	SITUATION AMOUREUSE,May'sa Ahmed,En couple depuis 5 avril 2012,MEMBRES DE LA FAMILLE
9	No result
10	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Labidi Karim,Frère
11	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
12	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
13	SITUATION AMOUREUSE,Marié(e),MEMBRES DE LA FAMILLE,Aucun membre de famille à afficher
14	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
15	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Landolsii Walid,Fil
16	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
17	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
18	SITUATION AMOUREUSE,Célibataire,MEMBRES DE LA FAMILLE,Dhahri Abdelkarim,Frère,Khalil Dh
19	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Youssef Dhraief,F
20	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
21	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Ahmed Khadhraou
22	No result
23	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Feten Bedoui Mes
24	SITUATION AMOUREUSE,Célibataire,MEMBRES DE LA FAMILLE,Aucun membre de famille à affich
25	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
26	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
27	No result
28	SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA FAMILLE,Aucun membre de
29	No result

output

FIGURE 3.10 – Famille et relations

1	Details About
2	À PROPOS DE CHTIOUI,Aucun détail supplémentaire à afficher,CITATIONS FA
3	À PROPOS DE MILED,loading . . . . . . Please wait █████████████] 99%,CITA
4	À PROPOS DE ATHIL,Aucun détail supplémentaire à afficher,CITATIONS FAVO
5	À PROPOS DE NADA,Aucun détail supplémentaire à afficher,CITATIONS FAVO
6	À PROPOS DE MOHAMED,Aucun détail supplémentaire à afficher,CITATIONS F
7	À PROPOS DE RIM,Aucun détail supplémentaire à afficher,CITATIONS FAVORI
8	À PROPOS DE MALEK,Aucun détail supplémentaire à afficher,AUTRES NOMS,
9	No result
10	À PROPOS DE ELYES,Ne fais pas l'impossible pour quelqu'un qui n'a pas voulu
11	À PROPOS DE HICHEM,Aucun détail supplémentaire à afficher,CITATIONS FA
12	À PROPOS DE TAREK,Aucun détail supplémentaire à afficher,CITATIONS FAVO
13	À PROPOS DE KAIS,قلبي كبير وعقلي أكبر: لا أجامل ولا أعادي,CITATIONS FAVORITES,
14	À PROPOS DE GHARBI,Aucun détail supplémentaire à afficher,CITATIONS FA
15	À PROPOS DE MOHAMED,J'aime ma mère et j'adore que mon CRÉATEUR ♥♥♥
16	À PROPOS DE WALID,Aucun détail supplémentaire à afficher,CITATIONS FAVO
17	À PROPOS DE CHAKER,Aucun détail supplémentaire à afficher,CITATIONS FA
18	À PROPOS DE HAMZA,Aucun détail supplémentaire à afficher,CITATIONS FAVO
19	À PROPOS DE AMIR,Aucun détail supplémentaire à afficher,CITATIONS FAVOR
20	À PROPOS DE MOHAMED,Aucun détail supplémentaire à afficher,CITATIONS F
21	À PROPOS DE HAITHEM,Aucun détail supplémentaire à afficher,CITATIONS FA
22	No result
23	À PROPOS DE ADAM,Aucun détail supplémentaire à afficher,CITATIONS FAVOF
24	À PROPOS DE OMAR,Aucun détail supplémentaire à afficher,CITATIONS FAVOI
25	À PROPOS DE AHMED AMINE,Aucun détail supplémentaire à afficher,CITATION
26	À PROPOS DE INES,♥,CITATIONS FAVORITES,Aucune citation favorite à affic
27	No result
28	À PROPOS DE ADNEN,Aucun détail supplémentaire à afficher,CITATIONS FAV
29	No result

FIGURE 3.11 – Détails sur la personne

1	Posts
2	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,17/09/2010 16:24 || status update || Chtioui Malek || عيدكم مبروك عيدكم مبروك عيدكم مبروك عيدكم مبروك || https://www.facebook.com/zaghouan.mc
3	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,11/09/2018 15:21 || others || Miled Othmen a changé sa photo de couverture. || || , 12/05/2018 18:41 || others || Miled Othmen a chang
4	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
5	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,16/10/2018 11:42 || life event || || || https://www.facebook.com/photo.php?fbid=10212793382684804&set=a.2749905302219&type=3&
6	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,28/06/2019 22:26 || life event || Mohamed Ouederni || || https://www.facebook.com/photo.php?fbid=10220753031398260&set=a.18653
7	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
8	akri || || , 29/08/2011 05:10 || others || Malek Boubakri a p
9	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
10	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,22/06/2012 00:16 || life event || Elyes Labidi || || https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv
11	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,08/04/2019 21:36 || others || Hichem Hamdaoui a changé sa photo de profil. || || , 08/04/2019 21:34 || others || || || ,
12	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
13	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,03/07/2019 14:06 || life event || Kais Chebbi || || https://www.facebook.com/photo.php?fbid=2532175800160841&set=a.82016395136
14	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,31/05/2019 10:54 || life event || || || https://www.facebook.com/photo.php?fbid=10157179710529598&set=a.10150957321654598&type
15	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,03/07/2019 19:55 || others || Mohamed Boujnah a partagé une publication. || 13:31 03/07/2019 || تحول إلى عدو اصلا كان حاقدا منذ البداية
16	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,23/06/2019 22:51 || others || Walid Sassi a partagé une publication. || || , 23/06/2019 02:37 || others || || حملة مك المناطق نفذت العديد من الأخيرة.
17	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
18	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,30/11/2018 18:46 || status update || Hamza DH || ان شاء الله || https://www.facebook.com/hamza.dhahri# , 21/10/2018 16:27 || others || H
19	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,07/04/2011 20:32 || others || Amir Khaled aime Gaming Zone ! La zone de jeux gratuit du web. || || , 30/10/2010 19:18 || others || Amir
20	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,07/02/2019 23:56 || others || Mohamed Masmoudi a changé sa photo de couverture. || || , 07/02/2019 23:52 || others || Mohamed Mas
21	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,05/11/2014 02:05 || others || Haithem Hamzaoui a changé sa photo de couverture. || || , 30/10/2014 21:33 || others || || || ,
22	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
23	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,21/07/2011 10:35 || others || Adam Bedoui a partagé une vidéo. || || , 21/07/2011 10:20 || others || || || ,
24	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,
25	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,08/04/2012 11:07 || others || Ahmed Amine Mejri a changé sa photo de profil. || || , 01/01/2012 10:00 || others || || || ,
26	TIME || TYPE || TITLE || STATUS || LINKS(Shared Posts/Shared Links etc) ,22/06/2011 17:49 || others || Ines Trabelsi aime 14:28 08/06/2011, || || . تحميل أغنية يلبنت الله || others || Ines Trabelsi a partagé un lien

FIGURE 3.12 – Statuts

1	tweets
2	['No twitter account']
3	[['815908026649559041', datetime.datetime(2017, 1, 2, 13, 10, 7), 'RT @prw190: #تعلم_الإحتلال والبرتغال : إسبانيا دولة : أين\n وكان عام800 حكموهم الذين والمسلمين العرب حكامهم اليمن دول وتصبح إستغلالها تأخذ لم\n#ذكرى...'], ['81590543913732
4	['No twitter account']
5	[['926036600890552321', datetime.datetime(2017, 11, 2, 10, 41, 45), '@dinatokio #creatorsforchange series is so inspiring! Dina,you are a great hijabi Muslim woman that we are proud of! https://t.co/FCpr4dXXbJ'], ['
6	['No twitter account']
7	['No twitter account']
8	photo sur Facebook http://t.co/ErLAzRzIBC'], ['3298239552779919
9	['No twitter account']
10	[['717075606597275648', datetime.datetime(2016, 4, 4, 19, 45, 21), 'RT @ranahayek3: التوعية خلال حصص الخبراء مالبس تجربة الأطفال يحب بالتجربة\n التعليم\n #الأم ضد _ معاً \n@maglebanon https://t.co/fADsA89nz0'], ['7170275
11	[['364035325795901440', datetime.datetime(2013, 8, 4, 14, 49, 19), '@eahram محمد مرسي الرئيس د, شرعية يرجوع الأمة حل'], ['358897026240491521', datetime.datetime(2013, 7, 21, 10, 31, 33), الأمريكان يلعصاة والحكومة الجيش يد
12	['No twitter account']
13	[['940937601418125314', datetime.datetime(2017, 12, 13, 13, 33), 'Bonjour à tous, ma société recrute ! https://t.co/lW5i8mwRh0 Si une offre t'intéresse, contacte-moi ! N'hésite pas… https://t.co/fxAPgnjvoe'], ['93289
14	['No twitter account']
15	['No twitter account']
16	[['98244476312633345', datetime.datetime(2011, 8, 2, 4, 11, 23), 'hey everyone youve got to check this out I made $530 today! http://t.co/8si4PYv'], ['98000457988980736', datetime.datetime(2011, 8, 1, 12, 1, 44), "I
17	['No twitter account']
18	['No twitter account']
19	['No twitter account']
20	['No twitter account']
21	[['653488581105815552', datetime.datetime(2015, 10, 12, 8, 33, 13), 'salam'], ['644131522509803520', datetime.datetime(2015, 9, 16, 12, 51, 37), 'I i Haythem nice to see you http://t.co/WmZ2Sr0bnI'], ['48919468651
22	['No twitter account']
23	a new photo to Facebook https://t.co/gAriVOYcvW'], ['803680764520853504', datetime.datetime(2016, 11, 29, 19, 23, 20), 'I posted a new photo to Facebook https://t.co/VZDghooEiM'], ['803590179562864643', dateti
24	['No twitter account']
25	[['875154776039858177', datetime.datetime(2017, 6, 15, 0, 55, 32), 'How to install multiple php versions on ubuntu linux https://t.co/EqKqfGRJRy'], ['870960148810010625', datetime.datetime(2017, 6, 3, 11, 7, 35), 'H
26	9, 12, 4), 'Happy 67th independence day Indonesia:)'], ['236004558164357120', datetime.datetime(2012, 8, 16, 7, 40, 46), '@HananSoussi rabi iysabrek 7asa bik raw lee mi ho b kolou'], ['236002207814803456', dateti

FIGURE 3.13 – Tweets

```
{
   "Name":"Athil Belhadj",
   "Job Title":"Full-stack Developer(Angular & .NET) at Poulina Group Holding",
   "Company":"Poulina Group Holding",
   "College":"ISETN",
   "Location":"Kelibia, Gouvernorat de Nabeul, Tunisia",
   "skills":"Android Development,Programmation orientée objet,Java",
   "WorkPlaces and Organization":"Poulina Group Holding,AthilTech,PROTOL,STT,searching for r
   "URL":"https://www.linkedin.com/in/athil-belhadj-a98691ab/",
   "Overview":"Aucun lieu de travail à afficher,Aucune école à afficher,Aucun lieu à affiche
   "Work and Education":"EMPLOI,Aucun lieu de travail à afficher,SCOLARITÉ,Aucune école à af
   "Places Lived":"VILLE ACTUELLE ET D'ORIGINE,Aucun lieu à afficher",
   "Contact and Basic Info":"COORDONNÉES,Aucune coordonnée à afficher,INFORMATIONS GÉNÉRALES
   "Family and Relationships":"SITUATION AMOUREUSE,Aucune relation à afficher,MEMBRES DE LA
   "Details About":"À PROPOS DE ATHIL,Aucun détail supplémentaire à afficher,CITATIONS FAVOR
   "Life Events":"ÉVÈNEMENTS MARQUANTS",
   "Posts":" TIME || TYPE  || TITLE || STATUS  ||   LINKS(Shared Posts/Shared Links etc) ,",
   "tweets":"['No twitter account']"
},
```


FIGURE 3.14 – résultat

3.3 Tests et validation

Comportement vis-à-vis du temps d'exécution :
Notre solution est performante en terme de temps d'exécution. En effet, pour des données de 5.000.000 valeurs, la collecte dure 3 minutes.

Comportement vis-à-vis de la mémoire :
Notre application ne s'intéresse pas seulement au temps d'exécution, elle donne également de grandes performances en termes d'utilisation de la mémoire. Par exemple, les données de 5.000.000 valeurs utilisent 1150 Mo de mémoire pour leur stockage.

Comportement vis-à-vis du besoin d'évolutivité :
Les modules développés de notre collecteur de données intelligent respectent la programmation orientée-objet pour pouvoir ajouter de nouvelles fonctionnalités ou de nouveaux sites internet, ou lorsque la taille ou le volume des données deviennent plus grands.

Comportement vis-à-vis du besoin de facilité d'utilisation :
Les utilisateurs n'appartenant pratiquement pas à l'informatique, peuvent collecter les données publiques des personnes ou des produits à l'aide du SDC.

Conclusion

Ce dernier chapitre décrit la phase de mise en œuvre de notre projet. Tout d'abord, il présente les outils que nous utilisons pendant le cycle de vie du projet. Puis, il illustre quelques captures de notre solution afin de vérifier les résultats obtenus.

Conclusion et perspectives

Comme nous sommes à l'ère des données, l'analyse des données est devenue l'une des sciences informatiques cruciales. Les défis au-delà de l'analyse des données augmentent de plus en plus et fournissent des informations exploitables. Les résultats du traitement d'une grande échelle de données sont devenus une fortune inestimable pour les propriétaires de données.

Ce projet a pour objectif de collecter les données pertinentes sans tomber dans des problèmes de traitement lent ou de mémoire insuffisante afin de concevoir une solution adaptée au traitement de grande quantité de données, et de mettre en œuvre une interface utilisateur permettant aux clients d'analyser les résultats significatifs pour sortir des informations utiles à la prédiction.

Tout au long de ce rapport, nous avons détaillé les différentes phases du cycle de vie de notre projet en commençant par une étude préliminaire dans laquelle nous avons défini certains concepts utiles, puis nous avons examiné la solution existante à notre problème. Après, nous avons présenté les besoins fonctionnels et non fonctionnels du SDC. Ensuite, nous nous sommes concentrés sur la phase de conception en élaborant le diagramme de classes. Enfin, nous sommes passés à la phase de réalisation où nous avons exposé le travail accompli.

Tout au long de ce projet, nous avons confrontés à divers défis liés à la conception et à la recherche de la meilleure façon pour obtenir un collecteur de données volumineuses, qui ne collecte seulement les données publiques des personnes, mais aussi celles des produits, à partir les réseaux sociaux aussi bien que les différents sites de commerce.

Pour la prochaine version, de nombreuses extensions peuvent être ajoutées à notre solution actuelle. En effet, nous pouvons ajouter plus d'interfaces utilisateur graphiques à notre application et utiliser les résultats pour offrir une visualisation plus expressive à nos clients. La rapidité, la sécurité et la suffisance de mémoire sont également des éléments importants pour la performances de notre collecteur. Ainsi, les améliorations peuvent également couvrir ce niveau.

Bibliographie

[1] Eglantine Schmitt. L'ambiguïté épistémologique des big data : le cas de la donnée web en sciences sociales. *Big data, Open data, quelles valeurs ? Quels enjeux*, 2015.

Résumé

Dans le monde de l'analyse de données, de nombreuses fonctionnalités sont proposées pour réaliser des données à grande échelle dans un processus rapide, sûr et intelligent. Un tel champ informatique s'appelle BIG DATA. Dans ce projet, nous proposons un Collecteur de Données Intelligent ayant pour objectif principal de collecter les données pertinentes des personnes ou des produits, à partir des différents sites internet (site de commerce, réseaux sociaux, forum, blog, etc), afin de les analyser pour sortir des informations utiles à la prédiction.

Mots-clés : Python, Scrapy, Beautiful Soup, Big Data

Abstract

In the world of data analysis, many features are proposed to realize large-scale data in a fast, safe and intelligent process. Such a computer field is called BIG DATA. In this project, we propose a Smart Data Collector whose main objective is to collect relevant data of people or products, from different websites (commerce website, social networks, forum, blog, etc.), in order to analyze them to extract information useful for prediction.

Keys Words : Python, Scrapy, Beautiful Soup, Big Data

ملخص

في عالم تحليل البيانات، توجد العديد من الميزات لأداء البيانات على نطاق واسع في عملية سريعة وآمنة وذكية. يسمى هذا المجال حوسبة البيانات الضخمة. في هذا المشروع، نقترح جامع بيانات ذكي هدفه الرئيسي هو جمع البيانات ذات الصلة، العامة أو الخاصة بالأفراد أو بالمنتجات، من مواقع مختلفة (موقع تجاري، شبكات اجتماعية، منتدى، مدونة، وما إلى ذلك) من أجل تحليلها وذلك لغرض استخراج معلومات مفيدة للتنبؤ.