الجمـهورية الجزائرية الديمقراطية الشعبيـة

وزارة التعليم العالي والبحث العلمي

جـامعـة سعـيـدة د. مـولاي الطـاهـر

كـلـيـة العلوم

قسم: الإعلام الآلي

Mémoire de Master

Spécialité : Réseaux Informatiques et Systèmes Répartis

Thème

La sélection automatique de clé de blocage lors de la mise en correspondance (matching) pour le Record Linkage

Présenté par :

LARID Redouane

MOKEDDEM Djaballah

Dirigé par :

Mr BENYAHIA Miloud

Promotion 2023 - 2024

Dedication

Je dédie ce projet de fin d'année à mes parents, dont le soutien inébranlable et l'amour constant m'ont inspiré et motivé tout au long de ce parcours académique. Leur croyance en mes capacités et leurs encouragements m'ont permis de surmonter les défis et d'atteindre mes objectifs.

À mes professeurs et mentors, qui ont partagé leur sagesse et leur expertise avec patience et générosité. Leur guidance précieuse a été essentielle à la réalisation de ce travail.

Enfin, à mes amis et collègues, pour leur camaraderie et leur soutien moral inestimable. Votre présence a rendu cette aventure plus enrichissante et mémorable.

Merci à tous pour votre soutien indéfectible.

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué à la réalisation de ce projet de fin d'année sur le "Record Linkage".

Tout d'abord, je remercie sincèrement mon directeur de mémoire, Dr. Benyahia Miloud , pour sa supervision, ses conseils avisés, et son soutien constant tout au long de ce travail. Ses compétences et son expertise ont été indispensables à l'avancement et à la finalisation de ce projet.

Je souhaite également remercier l'ensemble du corps professoral de l'université, pour l'enseignement de qualité et les précieux savoirs transmis durant toutes ces années. Leur dévouement et leur passion pour l'enseignement ont grandement enrichi mon parcours académique.

Mes remerciements vont aussi à mes camarades de promotion, pour les moments de partage, de collaboration et de soutien mutuel. Leur amitié et leur solidarité ont été une source de motivation inestimable.

Je n'oublie pas ma famille et mes proches, dont le soutien moral et affectif a été une source de force et de réconfort durant les moments difficiles.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont apporté leur aide, leur encouragement, et leur inspiration dans la réalisation de ce projet.

Introduction Générale..... 5
Organisation du Mémoire..... 7
1. La Qualité Des Données.....10
1.1. Introduction.....10
1.2. Définition.....10
1.3. Les Critères De La Qualité Des Données.....10
1.3.1. Les Critères Intrinsèques.....10
1.3.2. Les Critères de Services.....11
1.4. Dimensions De Qualité Des Données.....12
1.5. l'importance De La Qualité De Données.....14
1.5.1. Prise de décision précise.....14
1.5.2. Efficacité opérationnelle.....14
1.5.3. Conformité.....15
1.5.4. Opérations financières.....15
1.5.5. Personnalisation et fidélisation des clients.....15
1.5.6. Avantage concurrentiel.....15
1.5.7. Numérisation.....16
1.6. Principaux Problèmes De La Non-Qualité Des Données.....16
1.6.1. Création Des Données.....16
1.6.2. Collecte/Import Des Données.....17
1.6.3. Stockage Des Données.....17
1.6.4. Intégration Des Données.....17
1.6.5. Recherche et Analyse Des Données.....17
1.7. Approches Générales Pour Détecter et Corriger Les Problèmes De Qualité Des Données.....18
1.8. Les Objectives De La Qualité Des Données.....19
1.9. Conclusion.....20
2. Record Linkage..... 23
2.1. Introduction.....23
2.2. Définition RL.....23
2.3. Méthodologie du Record Linkage.....24
2.3.1. Déterministe.....24
2.3.2. Probabiliste.....24
2.4. Les étapes de Record Linkage.....25
2.4.1. Nettoyage et normalisation.....25
2.4.2. L'indexation.....25
Le Blocage.....25
2.4.3. Matching.....26
Codage phonétique.....26

Recherche de motifs..... 27
2.5. Conclusion..... 30
3. Implémentation et Expérimentation..... 33
3.1. Introduction..... 33
3.2. Environnement de Développement..... 33
3.3. La sélection automatique des clés Algorithme de BES..... 34
Génération de clés de candidat..... 36
BES Pseudo Code..... 37
3.4. Présentation de l'Application..... 40
3.4.1. Fonctionnalités de l'Application..... 40
Chargement et Prétraitement des Données :..... 40
Gestion des Doublons :..... 40
Rapports et Exportation :..... 40
3.5. Algorithmes Utilisés..... 41
3.5.1. Algorithme d'Optimisation BES (Bald Eagle Search)..... 41
3.6. Captures d'Écran..... 42
3.7. Analyse des résultats..... 48
3.8. Conclusion..... 51
4. Conclusion Générale..... 52
5. Bibliographie..... 54

Introduction Générale

Ces dernières années, le monde a été témoin d'une explosion massive du volume de données. En particulier, après l'adoption des smartphones et des médias sociaux qui génèrent une énorme

quantité de données au quotidien. Les organisations du monde entier se sont retrouvées dans la nécessité d'intégrer leurs propres données provenant de diverses sources et sous différents formats. Ces données doivent être intégrées afin de faciliter le processus d'analyse des données et d'en extraire des informations utiles. Cependant, l'intégration des données peut devenir un processus très long en raison des problèmes de qualité des données, tels que les doublons, les valeurs manquantes et les problèmes d'intégrité référentielle. Les parties prenantes sont désormais plus conscientes de l'importance de la qualité des données. Beaucoup d'argent est investi pour améliorer la qualité des données stockées.

L'un des processus les plus importants dans le domaine de la qualité des données est le processus de couplage d'enregistrements (RL). L'objectif du couplage d'enregistrements est d'identifier les tuples qui se réfèrent à la même entité du monde réel et de les fusionner en un seul, lorsque le processus de couplage d'enregistrements est appliqué à une seule base de données, il s'agit d'un processus de déduplication.

Au cours des dernières années, le couplage d'enregistrements a été utilisé dans plusieurs domaines pour des objectifs multiples. Par exemple, des organisations du monde entier utilisent le processus de couplage d'enregistrements pour trouver des produits en double dans leurs bases de données et les intégrer en un seul tuple. Un autre domaine important où le couplage d'enregistrements est un outil puissant et important est la suppression des doublons dans les citations bibliographiques, ainsi que d'autres domaines tels que les soins médicaux, l'analyse des données de recensement et bien d'autres encore.

Organisation du Mémoire

Cette étude est structurée en III chapitres et organisée comme suite :

Chapitre 1 : la qualité des données

Nous allons présenter la qualité des données, nous donnerons un aperçu sur la qualité et nous parlerons sur les critères qui définissent la qualité des données, les problèmes de la non-qualité des données.

Chapitre 2 : Record Linkage

Nous allons présenter le Recorde Linkage, L'indexation et le blocage et l'implémentation, évaluons Algorithme BES qui résolut le problème de la sélection automatique des clés de blocage, et la mise en correspondance des paires d'enregistrements indexés (Matching).

Chapitre 3 :

Nous allons présenter l'environnement de travail et l'explication de notre application.

Chapitre 1

La Qualité Des Données

1. La Qualité Des Données

1.1. Introduction

La qualité des données est une mesure de l'état d'un ensemble de données basée sur des facteurs tels que l'exactitude, l'exhaustivité, la cohérence, la fiabilité et la validité. La mesure de la qualité des données peut aider les organisations à identifier les erreurs et les incohérences dans leurs données et à évaluer si les données correspondent à l'objectif visé.

1.2. Définition

Le terme général désignant les caractéristiques des données et le processus garantissant ces caractéristiques est la qualité des données. Cette notion vise à assurer que les données sont complètes, fiables, pertinentes, opportunes et cohérentes. L'objectif est de disposer de données sans doublons, sans fautes d'orthographe, sans omissions, sans modifications inutiles et conformes à une structure définie. Les données sont considérées de bonne qualité si elles répondent aux attentes des utilisateurs. En d'autres termes, la qualité des données est déterminée par leur utilité et leur valeur pour les utilisateurs. Pour répondre aux besoins prévus, les données doivent être précises, disponibles en temps voulu, pertinentes, complètes, faciles à comprendre et dignes de confiance.

1.3. Les Critères De La Qualité Des Données

1.3.1. Les Critères Intrinsèques

- **Unicité :** L'unicité signifie qu'une entité du monde réel est représentée par un seul et unique objet métier dans l'entreprise, identifiable par un identifiant unique. Cela permet d'avoir une description unique pour chaque produit, contribuant ainsi à améliorer la qualité des données sur les produits.

- **Exactitude :** Une donnée est considérée comme "exacte" lorsque les valeurs de ses attributs correspondent précisément aux valeurs qu'elles sont censées représenter dans le monde réel. L'exactitude englobe deux aspects essentiels : la précision et la validité.

- **Complétude :** La complétude se réfère à la présence de valeurs significatives pour les attributs des objets. Autrement dit, les données doivent être complètes, sans omission des informations nécessaires.

- **Cohérence :** La cohérence implique l'absence de contradictions au sein d'un même objet. Par exemple, une incohérence serait détectée si le "prix actuel" d'un produit était supérieur à son "prix maximum". Cette notion s'applique aussi au niveau des services : les valeurs d'une instance d'un objet métier ne doivent pas être en conflit avec celles d'une autre instance ou d'un autre objet.

- **Intégrité :** L'intégrité concerne les relations entre les objets. Toutes les relations importantes entre les objets doivent être présentes. Par exemple, chaque facture doit être associée à une commande. Si une facture n'a pas de référence à une commande, cela constitue un problème d'intégrité.

1.3.2. Les Critères de Services

- **Actualité :** Une donnée est considérée comme à jour si elle reste correcte malgré de possibles variations par rapport à sa valeur exacte due à des changements dans le temps. Une donnée devient obsolète si elle est incorrecte à une date donnée, bien qu'elle ait été correcte auparavant. Le degré d'actualisation mesure dans quelle mesure une donnée est à jour, par exemple, l'âge devient obsolète à la date d'anniversaire.

- **Accessibilité :** Ce critère concerne la facilité d'accès aux données. Les services de données doivent être adaptés à leur utilisation, disponibles en mode événement (déclenché à chaque mise à jour), en mode requête (à la demande d'un utilisateur ou d'un processus) ou en mode batch pour des synchronisations en masse, notamment pour les besoins décisionnels.

- **Pertinence :** La pertinence évalue l'utilité des données. Une donnée peut être accessible, mais inutilement détaillée, rendant certains attributs superflus pour les utilisateurs finaux.

Les données doivent être adaptées à leur usage spécifique, avec une granularité de l'information correspondant aux besoins des utilisateurs.

- **Compréhensibilité :** Ce critère concerne la clarté et la facilité de compréhension des données. Une donnée est compréhensible si chaque utilisateur, processus ou application peut facilement trouver et comprendre les informations parmi les attributs disponibles. Cela nécessite une définition claire et documentée des concepts pour assurer un alignement sémantique entre tous les utilisateurs, qu'ils soient humains ou informatiques.

1.4. Dimensions De Qualité Des Données

Évoquant les critères de qualité des données, Delphine Barrau et ses co-auteurs attirent l'attention sur ce qui suit : n'importe qui peut publier les données ; les données reflètent la vision du fournisseur ; les données sont liées ; l'utilisation faite des données n'est pas connue a priori ; les données sont ouvertes à tous ; l'accès aux données se fait via le web et les données ouvertes peuvent représenter de gros volumes à traiter De ces considérations, il se dégage les dimensions de qualité des données.

- **Exactitude :**
    - L'exactitude syntaxique définit à quel point les données sont conformes à une règle de format (par exemple, une donnée est-elle bien un numéro de téléphone ?).
    - L'exactitude sémantique définit à quel point les données sont conformes à la réalité décrite (par exemple, le numéro de téléphone est-il bien celui de l'entité décrite ?).
- **Vérifiabilité :** Définit à quel point les données peuvent être contrôlées, certifiées ou garanties par contrat.

- **Traçabilité :** Définit à quel point les données portent l'information de leur provenance (par exemple source d'origine, processus de transformation subi avant publication, identification du producteur).
- **Confiance :** Définit à quel point les données et leur producteur sont fiables.
- **Pertinence :** Définit à quel point les données apportent une valeur ajoutée dans leur utilisation
- **Utilisabilité :**
    - Accessibilité : Définit à quel point les données sont disponibles, récupérables.
    - Compréhensibilité : Définit à quel point les données sont compréhensibles, incluant par exemple l'éventuelle présence d'un support et d'une documentation, ou mesurant la versatilité des données.
    - Licensing : Présence ou non d'une licence indiquant quelle réutilisation peut être faite des données.
- **Visibilité :** Définit à quel point les données sont localisables par les utilisateurs.
- **Fraîcheur :** Définit à quel point les données sont suffisamment récentes.
- **Complétude :** Définit le niveau de couverture avec lequel le phénomène observé est représenté dans l'assemblage des données. Se décline en complétude de la population et des informations présentes au niveau des individus.
- **Cohérence :** Définit à quel point les données satisfont un ensemble de contraintes syntaxiques et sémantiques.
- **Unicité :** Définit à quel point les données évitent les redondances (des données sont redondantes si elles décrivent un même objet du monde réel).
- **Consistance :** Définit à quel point les données évitent les contradictions et incohérences, peut concerner la détection de données redondantes (décrivant un même objet du monde réel) mais contradictoires ou la vérification de règles de cohérence métier.

- **Sécurité de l'accès :** Définit à quel point l'accès aux données est contrôlé (peut être vue comme une sous-dimension de celle d'accessibilité).
- **Confidentialité :** Définit à quel point la confidentialité des informations personnelles est préservée.
- **Interconnexion :** Définit à quel point les données sont riches et précises en termes de lien vers des sources externes complémentaires.

1.5. l'importance De La Qualité De Données

Le maintien de la propreté des données doit être un effort collectif entre les utilisateurs professionnels, le personnel informatique et les professionnels des données. Mais souvent, elle est simplement perçue comme un problème informatique, c'est-à-dire que les données deviennent sales lorsque certains processus techniques de capture, de stockage et de transfert des données ne fonctionnent pas correctement. Bien que cela puisse être le cas, les données nécessitent l'attention de toutes les bonnes parties prenantes pour maintenir sa qualité dans le temps. Pour cette raison, il devient impératif d'établir un argumentaire en faveur de la qualité des données devant les décideurs nécessaires, afin qu'ils puissent contribuer à sa mise en œuvre dans tous les services et à tous les niveaux.

Nous avons répertorié ci-dessous les avantages les plus courants de la qualité des données.

1.5.1. Prise de décision précise

Les chefs d'entreprise ne s'appuient plus sur des hypothèses, mais utilisent plutôt des techniques de business intelligence pour prendre de meilleures décisions. C'est… où une bonne qualité des données peut permettre une prise de décision précise, tandis qu'une mauvaise qualité des données peut fausser les résultats de l'analyse des données et conduire les entreprises à fonder des décisions cruciales sur des prévisions erronées.

1.5.2. Efficacité opérationnelle

Les données font partie de toutes les opérations, petites et grandes, d'une entreprise. Qu'il s'agisse du produit, du marketing, des ventes ou des finances – exploiter efficacement les

données dans tous les domaines est la clé. L'utilisation de données de qualité dans ces services peut amener votre équipe à éliminer les efforts redondants, à obtenir rapidement des résultats précis et à être productive tout au long de la journée.

1.5.3. Conformité

Conformité des données, normes (telles que le GDPR, l'HIPAA et le CCPA) exigent des entreprises qu'elles suivent les principes de minimisation des données, de limitation de la finalité, de transparence, d'exactitude, de sécurité, de limitation du stockage et de responsabilité. La conformité à ces normes de qualité des données n'est possible qu'avec de propres données fiables.

1.5.4. Opérations financières

Les entreprises encourent d'énormes coûts financiers dus à la mauvaise qualité des données . Des opérations telles que le versement des paiements en temps voulu, la prévention des incidents de sous-paiement et sûr paiement, l'élimination des transactions incorrectes et l'élimination des risques d'erreur. La fraude due à la duplication des données ne sont possibles qu'avec des données propres et de qualité.

1.5.5. Personnalisation et fidélisation des clients

Offrir des expériences personnalisées aux clients est le seul moyen de les convaincre d'acheter auprès de votre marque plutôt que d'un concurrent. Les entreprises utilisent une tonne de données pour comprendre le comportement et les préférences des clients. Grâce à des données précises, vous pouvez découvrir des acheteurs pertinents et leur offrir exactement ce qu'ils recherchent – ce qui garantit la fidélité des clients à long terme tout en leur donnant l'impression que votre marque les comprend comme personne d'autre.

1.5.6. Avantage concurrentiel

Presque tous les acteurs du marché ont utilisé les données pour comprendre la croissance future du marché et les éventuelles possibilités de vente incitative et croisée. L'alimentation de cette analyse en données de qualité provenant du passé vous aidera à créer un avantage concurrentiel sur le marché, convertir plus de clients et augmenter votre part de marché.

1.5.7. Numérisation

Numérisation des processus cruciaux peut vous aider à éliminer le travail manuel, à accélérer le temps de traitement et à réduire les erreurs humaines. Mais avec des données de mauvaise qualité, ces attentes ne peuvent être satisfaites. Au contraire, une mauvaise qualité des données vous obligera à vous retrouver dans un désastre numérique où la migration et l'intégration des données semblent impossibles en raison de structures de bases de données variables et de formats incohérents.

1.6. Principaux Problèmes De La Non-Qualité Des Données

Les problèmes des données ne naissent pas de nulle part, les causes de la non-qualité des données sont connues : On trouve les problèmes techniques ou les problèmes humains. Ces problèmes s'accumulent avec le temps, depuis la création, durant la manipulation et jusqu'à l'exploitation et l'analyse.

1.6.1. Création Des Données

- Entrée manuelle : absence de vérifications systématiques des formulaires de saisie
- Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données : attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données
- Entrée de doublons
- Approximations
- Contraintes matérielles ou logicielles
- Erreurs de mesure
- Corruption des données : faille de sécurité physique et logique des données

1.6.2. Collecte/Import Des Données

- Destruction ou mutilation d'information par des prétraitements inappropriés.
- Perte de données : buffer over flows, problèmes de transmission.
- Absence de vérification dans les procédures d'import massif.
- Introduction d'erreurs par les programmes de conversion de données.

1.6.3. Stockage Des Données

- Absence de méta-données.
- Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées.
- Modifications ad-hoc.
- Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système.
- Contraintes matérielles ou logicielles.

1.6.4. Intégration Des Données

- Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et de diverse agrégation.
- Problèmes de synchronisation temporelle.
- Systèmes de données non conventionnels.
- Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données.

1.6.5. Recherche et Analyse Des Données

- Erreur humaine.
- Contraintes liées à la complexité de calcul.

- Contraintes logicielles, incompatibilité.
- Problèmes de passage à l'échelle, de performances et de confiance dans les résultats.
- Approximations dues aux techniques de réduction des grandes dimensions

1.7. Approches Générales Pour Détecter et Corriger Les Problèmes De Qualité Des Données

Comme le représente la figure 1.1, on peut classer la plupart des travaux abordant la problématique de la qualité des données selon quatre grands types d'approches complémentaires.

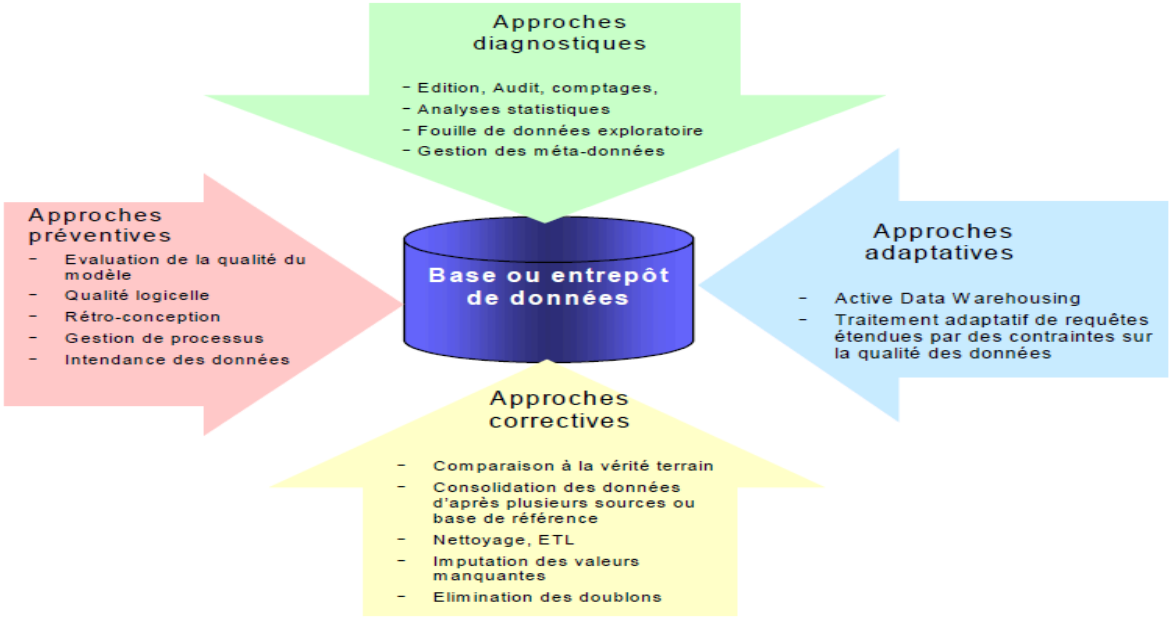

Figure 1. – Panorama des approches pour l'évaluation et le contrôle de la qualité des données.

- Les approches préventives centrées sur l'ingénierie des systèmes d'information et le contrôle des processus avec des techniques permettant d'évaluer la qualité des modèles conceptuels, la qualité des développements logiciels et celle des processus employés pour le traitement des données.

- Les approches diagnostiques centrées sur des méthodes statistiques, d'analyse et de fouille de données exploratoire permettant de détecter des anomalies sur les données.
- Les approches correctives centrées sur des techniques de nettoyage et de consolidation de données et utilisant des langages de manipulation des données étendus et des outils d'extraction et de transformation de données (ETL Extraction-Transformation-Loading).
- Les approches adaptatives ou actives appliquées généralement lors de la médiation ou de l'intégration des données : elles sont centrées sur l'adaptation des traitements (requêtes ou opérations de nettoyage sur les données) de telle façon que ceux-ci incluent à l'exécution en temps-réel la vérification des contraintes sur la qualité des données.

1.8. Les Objectives De La Qualité Des Données

- La qualité des données implique la collecte d'informations précises et fiables à travers un système de gestion des données, de suivi et d'évaluation (S&E).
- La qualité des données est cruciale pour les programmes de lutte contre le VIH/sida, qui sont généralement axés sur les résultats.
- Des données de haute qualité sont essentielles pour surveiller et évaluer les progrès accomplis dans la réalisation de ces objectifs.

1.9. Conclusion

Ce chapitre s'est concentré sur la revue de la littérature concernant la qualité des données. Nous avons d'abord défini la qualité des données et souligné son importance. Ensuite, nous avons présenté les différentes dimensions de la qualité des données ainsi que les problèmes engendrés par une mauvaise qualité des données. Ces problèmes peuvent avoir des impacts négatifs considérables sur l'efficacité d'une organisation.

Dans le chapitre suivant, le Couplage d'enregistrement qui est L'un des principaux processus dans le domaine de la qualité des données.

Chapitre 2

Record Linkage

2. Record Linkage

2.1. Introduction

La **liaison de données** (ou **record linkage** en anglais) est le processus de rapprochement et de combinaison de différentes sources de données pour identifier des enregistrements qui se rapportent à la même entité, qu'il s'agisse de personnes, d'entreprises, ou d'autres objets. Ce processus implique souvent des techniques de comparaison de chaînes de caractères, d'appariement de probabilités, et d'autres méthodes statistiques pour résoudre les incohérences et les variations dans les données afin de garantir une correspondance précise.

2.2. Définition RL

Le couplage d'enregistrements est le processus qui consiste à combiner des informations provenant de deux ou plusieurs enregistrements censés représenter la même entité. Cette technique est utilisée pour intégrer des données provenant de diverses sources ou pour identifier et éliminer les doublons au sein d'un même ensemble de données. Dans le domaine de l'informatique, le couplage d'enregistrements est également connu sous le nom d'appariement ou de dédoublonnage des données, en particulier lorsque l'objectif est de trouver des enregistrements en double dans un seul fichier.

Pour réaliser le couplage d'enregistrements, les attributs d'une entité au sein d'un enregistrement sont utilisés pour relier plusieurs enregistrements. Ces attributs peuvent être des identifiants uniques, tels que le numéro de sécurité sociale ou le numéro de plaque d'immatriculation, ou d'autres caractéristiques telles que le nom, la date de naissance et les détails de la voiture (par exemple, le modèle et la couleur). Le processus comporte plusieurs étapes : nettoyage, indexation, comparaison, classification et évaluation des données. Si nécessaire, les paires d'enregistrements classées sont utilisées pour affiner les étapes précédentes. La boîte à outils Python de couplage d'enregistrements est conçue pour suivre ce flux de travail.

2.3. Méthodologie du Record Linkage

2.3.1. Déterministe

Les méthodes de couplage déterministes utilisent des algorithmes pour vérifier si les paires d'enregistrements s'alignent ou divergent sur la base d'identifiants spécifiques. L'accord ou le désaccord sur chaque identifiant est traité comme un résultat binaire, où il y a soit une concordance complète, soit une discordance. La détermination de la concordance peut se faire en une ou plusieurs étapes.

Dans l'approche en une seule étape, tous les enregistrements sont comparés simultanément à l'aide de l'ensemble des identifiants. Une paire d'enregistrements n'est qualifiée de concordante que si les deux enregistrements présentent un alignement précis sur tous les identifiants et si la paire est identifiable de manière unique, ce qui signifie qu'aucune autre paire ne partage des valeurs identiques. À l'inverse, toute discordance entre les identifiants ou l'absence d'identification unique entraîne la classification de la paire en tant que non-concordance.

Dans la stratégie en plusieurs étapes, également connue sous le nom de méthode itérative ou par étapes, les enregistrements sont appariés par le biais d'une série d'étapes de moins en moins strictes. Les paires d'enregistrements qui ne satisfont pas aux critères lors de la première phase d'appariement sont soumises à d'autres comparaisons lors des phases suivantes. Une concordance est établie si la paire satisfaite aux critères à n'importe quelle étape du processus. Dans le cas contraire, la paire est classée comme non correspondante. Ces méthodes de couplage déterministes peuvent également être qualifiées de "déterministes exactes", nécessitant une concordance précise pour tous les identifiants, ou de "déterministes approximatives ou itératives", nécessitant une concordance précise lors d'un des nombreux cycles de concordance, mais pas pour tous les identifiants possibles.

2.3.2. Probabiliste

L'approche déterministe ne tient pas compte du fait que certains identifiants ou certaines valeurs ont un pouvoir discriminatoire plus important que d'autres. Des stratégies probabilistes ont été développées pour évaluer (1) le pouvoir discriminant de chaque identifiant et (2) la probabilité

que deux enregistrements soient réellement concordants en fonction de leur concordance ou de leur non-concordance avec les différents identifiants.

2.4. Les étapes de Record Linkage

2.4.1. Nettoyage et normalisation

Le nettoyage des données est un processus qui vise à identifier et à corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale du traitement des données améliore la cohérence, fiabilité et valeur des données. La normalisation des données est le processus de normalisation des attributs de données qui peuvent représenter les mêmes informations, mais avec une identification différente dans chacun. Par exemple, l'attribut sexe peut être trouvé dans un jeu de données en tant que valeur binaire (0/1) et dans un autre comme (M/F).

2.4.2. L'indexation

Ceci est considéré comme l'étape la plus importante de ce processus. L'objectif d'indexation est de trouver les mots qui représentent le mieux contenu d'un document, la technique d'indexation la plus courante est "le blocage".

Le Blocage

Le blocage est la technique la plus utilisée dans l'étape de l'indexation. Le blocage est le processus qui divise le dataset en un ensemble de blocs. Tous les tuples affectés au même bloc partagent une valeur commune appelée valeur de clé de blocage (BKV). La clé de verrouillage peut être sélectionnée comme un attribut unique.

La figure 1 représente un exemple de clés de blocage générées à partir du jeu de données de restaurant où la valeur de clé de blocage est formée par la concaténation de la ville du restaurant et du numéro de téléphone.

BK1	BK2	Name	Address	City	Phone	Type
A6553102461501	LASANG435	arnie morton's of Chicago	435 s. la cienega blv.	Los Angeles	310/246-1501	American
H3413104721211	STADAC12224	art's deli	12224 ventura bold	Studio city	818-762-1221	delis

Tableau 1. Exemple de clés de blocage à partir du jeu de données du restaurant.

2.4.3. Matching

La troisième et dernière étape du processus de couplage d'enregistrement consiste à faire correspondre les enregistrements indexés qui se trouvent dans le même bloc et à décider s'ils représentent la même entité du monde réel ou pas. La valeur de correspondance est normalisée entre 0 et A ou 1 représente une correspondance exacte et 0 une non-correspondance totale. La correspondance peut être effectuée à l'aide d'un ensemble de fonctions de similarité de chaines qui existent dans la littérature [Levenshtein, 1966] ou en utilisant un algorithme d'apprentissage automatique pour classer l'enregistrement comme correspondant ou non correspondant [11].

Codage phonétique

La première famille de techniques de Matching est là l'encodage phonétique Une variété d'algorithmes de Matching existe dans la littérature.

- **Soundex**

Soundex est considéré comme l'une des fonctions d'encodage phonétique les plus efficaces. Il transforme les chaines selon leur prononciation afin qu'elles puissent être comparées les unes aux autres sans tenir compte des fautes d'orthographe. En utilisant Soundex, des noms comme ALLAN et ALLEN sont tous deux représentés avec le même code "A450", ce qui facilite la correspondance entre les deux noms.

Les principales étapes de Soundex sont :

- Conservez la première lettre de la chaine.

- Remplacez toutes les consonnes en respectant les règles suivantes : (0 pour les caractères A, E, H, I, O, U, W, Y. 1 pour les caractères B, F, P, V. 2 pour C, G, J, K, Q, S, X, Z. 3 pour D, T. 4 pour L et 5 remplace M, N. 6 remplace le caractère R.

- Dans le cas où la chaîne est trop courte, l'algorithme complète les trois nombres après le premier caractère par des zéros.

- **NYSIIS**

NYSIIS a la même idée et le même objectif que l'algorithme Soundex. La différence est que NYSIIS renvoie un code composé de lettres, ce qui n'est pas le cas de Soundex. L'algorithme NYSIIS augmente la précision de 2, 7 % par rapport à Soundex. Les règles de base de l'algorithme NYSIIS sont la transformation des premiers caractères ou : (MAC est remplacé par MCC et KN devient NN, K en C, PH-PF en FF, SCH en SSS) et les derniers caractères (EE-IE en Y, DT-RT RD-NT-ND a D).

Recherche de motifs

La deuxième famille de techniques de Matching est la recherche de modèles. Une variété d'algorithmes de Matching existe dans la littérature.

- **Distance d'édition**

La distance d'édition est également connue sous le nom de distance de Levenshtein. Il a été proposé en 1965 par Vladimir Levenshtein . Il est considéré comme l'une des métriques les plus utilisées pour mesurer la similarité entre deux chaînes. Généralement, il est défini comme le nombre d'insertions, de suppressions et de mises à jour afin de transformer une chaîne en une autre. Pour mieux comprendre, l'exemple suivant montre une démonstration de la façon de calculer les coûts de passage d'un mot à l'autre.

Word 1	Word 2	Operation	Cost
I		Delete (I)	1
N	E	Substitute (E)	1
T	X	Substitute (X)	1
E	E	Comparison	0
	C	Insert (C)	1
N	U	Substitute (U)	1
T	T	Comparison	0
I	I	Comparison	0
O	O	Comparison	0
N	N	Comparison	0
		Sum	5

Tableau 3 – Edit distance example

Dans l'exemple présenté dans la figure 2.4, on voit que la distance d'édition entre les deux mots (Intention, Exécution) est la somme des coûts pour transformer la première chaîne en la seconde qui est égale à Cinq. Les étapes décrites dans l'exemple ne sont pas la seule solution pour transformer la première chaîne en la seconde, mais sont celles qui coûtent le moins cher.

- **Jaro-Winkler**

Le Jaro-Winkler est une metrique de similarite de chaine qui a ete proposee par William E. Winkler en 1990 comme une extension de la distance Jaro. Afin de mesurer la similarite

Jaro-Winkler entre deux chaines, la première etape consiste a mesurer la similarite Jaro traditionnelle qui est definie comme suit :

$$Jaro - Sim(s_1, s_2) = \begin{cases} 0, & m = 0 \\ \frac{1}{3}(\frac{m}{s1} + \frac{m}{s2} + \frac{m-t}{m}) & sinon \end{cases}$$

Ou :

– s représente la longueur de la chaîne.

– m représente le nombre de caractères communs entre les séquences comparées avec le même indice.

– t représente le demi-nombre de transpositions.

Afin d'améliorer la métrique précédente, William E. Winkler utilise une échelle de préfixe P afin de mettre en favoris les chaines commençant par le même préfixe L pour une longueur maximale de Quatre. La similarite Jaro-Winkler est définie comme suit :

JaroWinlker −Sim(s1, s2) = Jaro−Sim(s1, s2)+LP(1−Jaro−Sim(s1, s2))

Ou :

– JaroSim (s1,s2) est la similarité Jaro entre les chaines.

– L est la longueur du préfixe.

– P est un facteur d'échelle (une constante qui prend generalement la valeur 0,1).

- **Distance de Jaccard**

La distance de Jaccard est généralement utilisée pour mesurer la similarité entre deux jeux d'échantillons, ce qui peut être le cas de Strings. Pour mesurer la distance de Jaccard, il faut

d'abord calculer le coefficient de Jaccard qui est défini comme :

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B}$$

Une fois cela fait, la distance de Jaccard n'est obtenue que par la soustraction du coefficient de Jaccard a 1.

2.5. Conclusion

Le processus d'identification des paires d'enregistrements représentant la même entité dans différentes bases de données, connu sous le nom de couplage d'enregistrements, constitue une étape initiale cruciale dans de nombreuses applications d'exploration de données. Le traitement de millions d'enregistrements pour ce couplage est une tâche exigeante en termes de calcul.

Chapitre 3

Implémentation et Expérimentation

3. Implémentation et Expérimentation

3.1. Introduction

Dans ce chapitre, nous allons détailler l'environnement de développement ainsi que les technologies employées pour la mise en œuvre de notre application de couplage d'enregistrements. Ensuite, nous présenterons notre contribution spécifique au domaine du couplage d'enregistrements, Enfin, nous conclurons avec quelques captures d'écran illustrant les principales fonctionnalités de notre application.

3.2. Environnement de Développement

Pour le développement de notre application de couplage d'enregistrements, nous avons utilisé les outils et technologies suivants :

Langage de programmation : Python. Python a été choisi pour sa simplicité, sa robustesse, et sa vaste bibliothèque de modules, ce qui facilite le traitement des données et la mise en œuvre des algorithmes complexes nécessaires pour le couplage d'enregistrements.

Environnement de développement intégré (IDE) : Visual Studio Code (VS Code). VS Code a été sélectionné pour ses fonctionnalités avancées, telles que l'autocomplétion, le débogage intégré et les extensions qui augmentent la productivité du développement.

DataSet : Nous avons utilisé des ensembles de données synthétiques et réels pour tester et valider notre application. Ces ensembles de données comprenaient des enregistrements de diverses sources, présentant des problèmes de doublons, de variations orthographiques, et de données manquantes.

3.3. La sélection automatique des clés Algorithme de BES

L'algorithme d'optimisation BES (Bald Eagle Search) a été récemment proposé pour résoudre les problèmes d'optimisation. Il a été construit en utilisant à la fois des techniques d'essaim et d'évolution (Alsattar et al., 2019). Comme mentionné ci-dessus, notre objectif dans cet article est d'adapter l'algorithme BES au problème de sélection de clé de blocage. Ce dernier peut en effet être modélisé comme un problème de sélection de caractéristiques. La population initiale est un groupe de sous-ensembles de fonctionnalités, c'est-à-dire des clés de blocage. Ainsi, chaque membre d'une population représente un sous-ensemble concurrent de clés de blocage. De plus, les sous-ensembles n'ont pas nécessairement les mêmes tailles ou les mêmes éléments. Les populations sont régénérées à l'aide de l'algorithme BES.

L'aptitude de chaque membre d'une population est calculée en utilisant l'approche RL proposée dans (Benkhaled et al., 2019) de manière globale. Dans cette approche, K-Modes est utilisé comme étape d'indexation en regroupant les données en utilisant uniquement les clés de blocage qui sont le sous-ensemble de fonctionnalités actuellement sélectionné.

Les meilleures clés de blocage en termes de fitness sont celles dans lesquelles K-Modes regroupe les enregistrements les plus dupliqués lorsqu'ils sont utilisés comme attributs de clustering. En conséquence, la fonction de fitness est le paramètre de complétude de la paire (PC). Le PC mesure le nombre de doublons détectés par une approche RL en utilisant les touches de blocage sélectionnées.

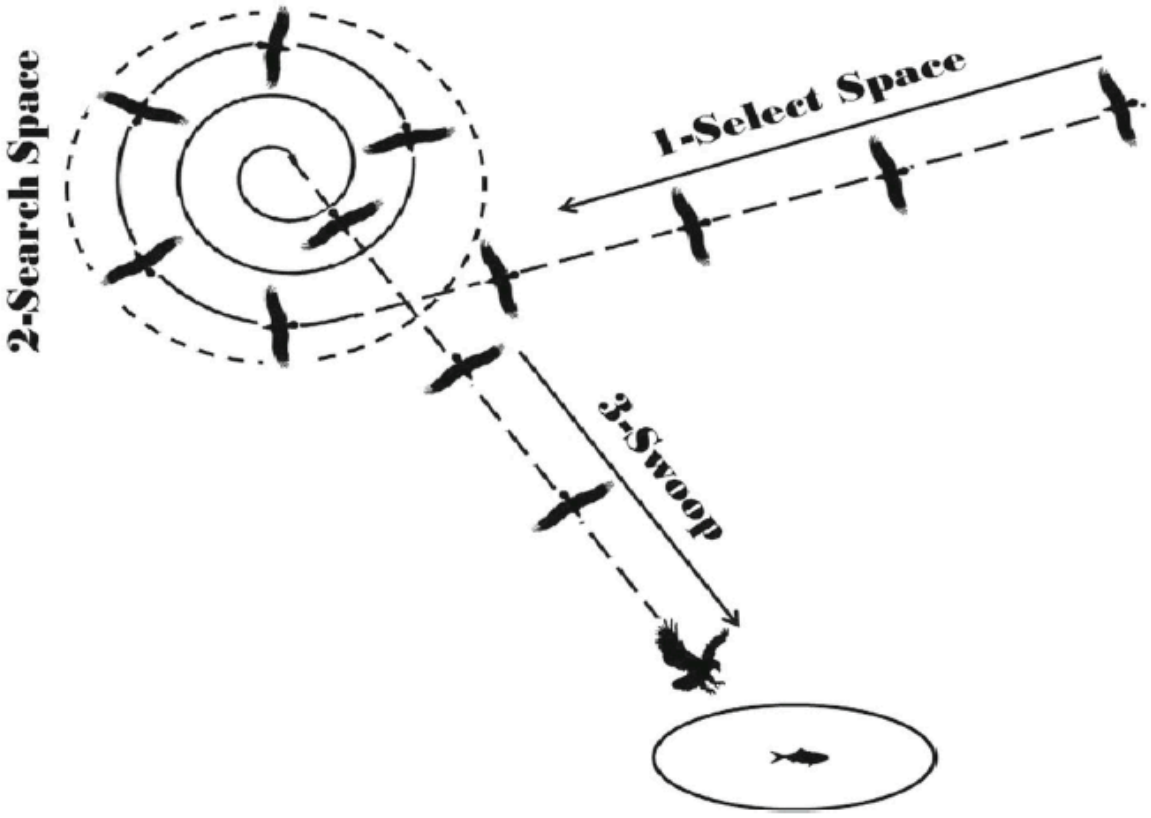

Figure 2 Conséquences pour les trois principales étapes de la chasse par BES

Notre approche proposée peut être résumée par les points suivants :

- Générer toutes les listes de clés de blocage possibles.
- Réinitialisez la première population qui est un sous-ensemble d'entités aléatoires de la liste des clés de blocage précédemment générée.
- Exécutez l'algorithme BES pour les itérations T sur la population précédemment générée dans une méthode wrapper avec l'exhaustivité de la paire comme fonction de fitness.
- Le meilleur membre de la dernière population est sélectionné comme meilleur sous-ensemble d'entités à utiliser comme clés de blocage.

Génération de clés de candidat

La liste des clés candidates est celle à partir de laquelle la population initiale de l'algorithme BES sera sélectionnée au hasard. Avant de générer la liste des clés candidates, une étape essentielle de prétraitement ne peut être négligée. Il s'agit, en fait, de nettoyer l'ensemble A. en d'autres termes, il faut éliminer les attributs de mauvaise qualité de l'ensemble A. Deux paramètres sont utilisés pour calculer la qualité globale d'un attribut. Premièrement, l'exhaustivité représente le pourcentage de valeur nulle concernant les attributs spécifiés (Pipino et al., 2002). Nous avons utilisé la mesure NBC (complétude basée sur zéro) où la complétude est mesurée à l'aide de l'équation (1). En utilisant cette méthode, la valeur 1 représente le meilleur résultat et 0 le pire.

Tous les attributs qui ont une valeur d'exhaustivité inférieure au seuil prédéfini sont éliminés à partir de la génération de la liste des clés de blocage candidates :

$$Completness\ (Attj) = 1 - \frac{number\ of\ null\ values\ in\ Attj}{Number\ of\ instances} \quad (1)$$

Le deuxième paramètre est la cardinalité d'un attribut. La cardinalité représente le nombre de valeurs distinctes pour un attribut spécifié. Dans le processus RL, les attributs à très faible cardinalité ne conviennent pas pour être utilisés comme clés de blocage. Par exemple, l'utilisation de l'attribut sex comme clé de blocage divise les données en seulement 2 blocs (M / F). Par conséquent, dans notre approche, les attributs à très faible cardinalité sont éliminés de la génération de la liste des clés de blocage candidates.

Une fois que les attributs de mauvaise qualité sont éliminés ; pour chaque ensemble de données D, différentes clés de blocage peuvent être générées en fonction du domaine de l'ensemble de données et du type d'attributs. Nous avons utilisé un ensemble de fonctions F (comme expliqué dans la section 3) pour générer des clés candidates telles que First4Chars (Attributes), Concatenation (), Soundex (Attribute), Last4Chars (Attribute) et NYSIIS (Attribute).

Le tableau 2 présente certaines des différentes fonctions utilisées pour générer la liste des clés de blocage possibles. D'autres fonctions spécifiques ont été utilisées pour chaque ensemble de données ne sont pas mentionnées dans le tableau. Par exemple, «Extract-Number ()» est une fonction utilisée pour extraire le numéro du restaurant du champ d'adresse dans le cas du jeu de données du restaurant.

Function	**Description**
Soundex(Attribute), NYSIIS (Attribute)	Soundex and NYSIIS are both phonetic encoding algorithms (Holmes and McCabe2002) that transform a string to an alphanumeric presentation of how it's pronounced.
First_N_Chars (Attribute)	Extract the first N characters from an attribute field.
Last_N_Chars (Attribute)	Extract the last N characters from an attribute field.
Numerical (Attribute)	Extract the numerical value from a string.
Remove-SP (Attribute)	Remove special characters from a string.
Exact-Value (Attribute)	Use the attribute value without modification

Tableau 2. Les fonctions utilisées pour la génération des clés de blocage

BES Pseudo Code

Pour l'implementation de BES pour le problème de sélection de caractéristiques, la population est un ensemble de tableaux, chacun contenant un certain nombre d'entiers, et chaque entier représente l'indice d'une clé de blocage générée.

```
Algorithm 1. BES for feature selection.
1: Inputs: Dataset D, Number of iterations T, Number of
2:          solutions M, BES parameters: c1, c2,α, R.
3: Outputs: The best subset of blocking keys.
4: Function BES
5:     Initialize a random population of M member.
6:     Initialize the best solution: Pbest = All features.
7:     While (i <= T) do
8:     For each (member P in pop)) do
9:               Pnew = Calculate (Pnew) using equation 1.
10:          IF (f (Pnew) > F (P)) do
11:               P = Pnew
12:          IF (f (Pnew) > F (Pbest)) do
13:               P = Pnew
14:          End IF
15:          End IF
16:          End For
17:   For each (member P in pop)) do
18:               Pnew = Calculate (Pnew) using equation 2.
19:          IF (f (Pnew) > F (P)) do
20:               P = Pnew
21:          IF (f (Pnew) > F (Pbest)) do
22:               P = Pnew
23:          End IF
24:          End IF
25:          End For
26:        For each (member P in pop)) do
27:               Pnew =Calculate (Pnew) using equation 3.
28:          IF (f (Pnew) > F (P)) do
29:               P = Pnew
30:          IF (f (Pnew) > F (Pbest)) do
31:               P = Pnew
32:          End IF
33:          End IF
34:          End For
35:     End While.
36:     Return Pbest.
37: End Function.
```

Algorithm 1. Représente le pseudo-code de l'algorithme BES pour la sélection de caractéristiques.

La première étape consiste à générer la population initiale qui est un sous-ensemble de clés de blocage sélectionnées au hasard dans la liste générée précédemment. Ensuite, nous commençons la phase de sélection de l'espace (lignes 8-16), où pour chaque membre de la population, nous

calculons un nouvel objet (Pnew) en utilisant l'équation 2. α est une variable qui prend une valeur entre 1,5 et 2. R est un nombre aléatoire généré avec R ∈ [0,1] et Pi désigne la position actuelle (Alsattar et al.2019).

Une fois Pnew calculé, ses nouvelles caractéristiques obtenues sont nettoyées, toutes les valeurs continues sont arrondies en nombres entiers et tous les nombres entiers qui se situent en dehors de la plage [0,N] (N est le nombre de clés de blocage générées) sont remplacés par une caractéristique aléatoire (indice de clé de blocage).

L'étape suivante consiste à mesurer l'aptitude à l'aide des caractéristiques de Pnew et à la comparer aux performances de P et de Pbest si elle est meilleure qu'elles. Ensuite, ils sont tous deux remplacés par Pnew.

Dans la phase de recherche dans l'espace (lignes 17-25), pour chaque membre de la population, un nouvel objet Pnew est calculé à l'aide de l'équation 3. X et Y sont deux nombres aléatoires représentant le mouvement en spirale de l'aigle vers la zone de piquage. Enfin, la descente en piqué (lignes 26-34), c'est l'endroit où l'aigle cherche sa cible finale en spirale. Pour chaque objet de la population, Pnew est calculé à l'aide de l'équation 4.

$$P_{new} = P_{best} + \alpha * r * (P_{mean} - P_i) \quad (2)$$

$$P_{new} = P_i + y(i) * \left(Pi - P_{i+1}\right) + x(i) * (P_i - P_{mean}) \quad (3)$$

$$P_{new} = rand * P_{best} + x(i) * \left(P_i - C_1 * P_{mean}\right) + y(i) * \left(P_i - C_2 * P_{best}\right)$$

Dans l'ensemble, l'algorithme général d'optimisation de la recherche du pygargue à tête blanche peut être résumé dans l'organigramme suivant (figure 2).

3.4. Présentation de l’Application

3.4.1. Fonctionnalités de l'Application

Notre application de couplage d'enregistrements offre plusieurs fonctionnalités clés, visant à simplifier et à automatiser le processus de détection des doublons et de gestion des données. Voici un aperçu de ces fonctionnalités :

Chargement et Prétraitement des Données :

- **Importation de Données :** Les utilisateurs peuvent importer des ensembles de données à partir de divers formats, notamment CSV, Excel, et bases de données SQL.
- **Nettoyage des Données :** L'application propose des outils pour nettoyer les données en supprimant les valeurs manquantes, normalisant les formats de texte, et standardisant les valeurs.

Gestion des Doublons :

- **Fusion des Enregistrements :** L'application propose des mécanismes pour fusionner automatiquement ou manuellement les enregistrements doublons, en choisissant les valeurs les plus pertinentes ou en combinant les informations.
- **Visualisation des Doublons :** Les utilisateurs peuvent visualiser les enregistrements doublons et examiner les correspondances suggérées avant de valider les fusions.

Rapports et Exportation :

- **Génération de Rapports :** Création de rapports détaillés sur le processus de couplage, incluant les enregistrements fusionnés, les doublons détectés et les statistiques de performance.
- **Exportation des Données :** Exportation des données traitées vers divers formats pour une intégration facile avec d'autres systèmes et applications.

3.5. Algorithmes Utilisés

3.5.1. Algorithme d'Optimisation BES (Bald Eagle Search)

L'algorithme BES est inspiré du comportement de chasse des aigles à tête blanche. Il combine exploration et exploitation pour trouver des solutions optimales dans des espaces de recherche complexes. Voici les étapes clés de l'algorithme BES :

- **Exploration :** Les aigles explorent l'espace de recherche de manière aléatoire pour identifier les zones prometteuses.
- **Migration :** Les aigles peuvent se déplacer vers de nouvelles zones si les solutions trouvées ne sont pas satisfaisantes, permettant une meilleure couverture de l'espace de recherche.

Dans notre application, l'algorithme BES est utilisé pour optimiser les paramètres des algorithmes de couplage d'enregistrements, améliorant ainsi la précision et l'efficacité du processus.

3.6. Captures d'Écran

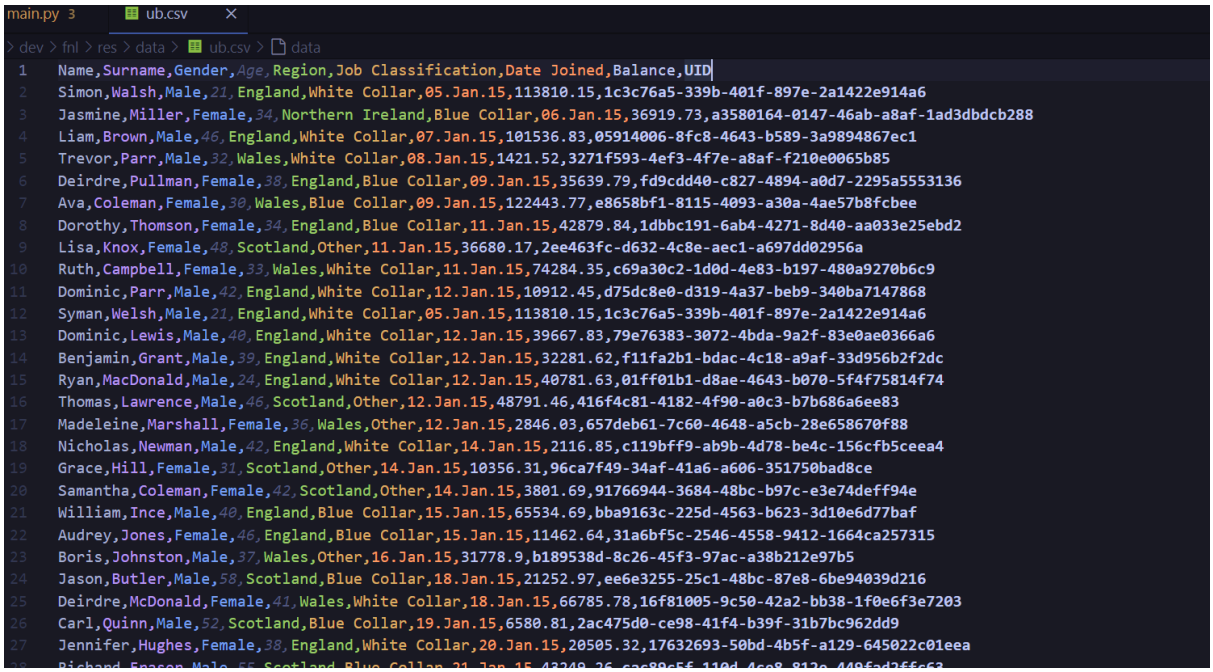

Figure 3 : Afficher le fichier data set

Tout d'abord, nous allons lancer la ligne de commande et naviguer vers le chemin de notre application. Ensuite, nous voulons activer notre environnement virtuel où nous stockons tous les outils nécessaires à l'exécution de l'application (en exécutant activate.exe situé dans ‘record_linkage\scripts\ ‘).

Figure 1 : Lancez la ligne de commande et naviguez jusqu'au chemin d'accès à notre application

Avant d'exécuter l'application, nous avons seulement quatre dossiers dans le chemin principal : le dossier généré par VS Code, le dossier record_linkage qui représente l'environnement virtuel, et le dossier src qui contient tout le code source pour construire cette application.

Nous avons également un autre dossier important pour l'exécution des scripts, qui est le dossier keys, Il contient les clés initiales que nous allons utiliser pour indexer les données.

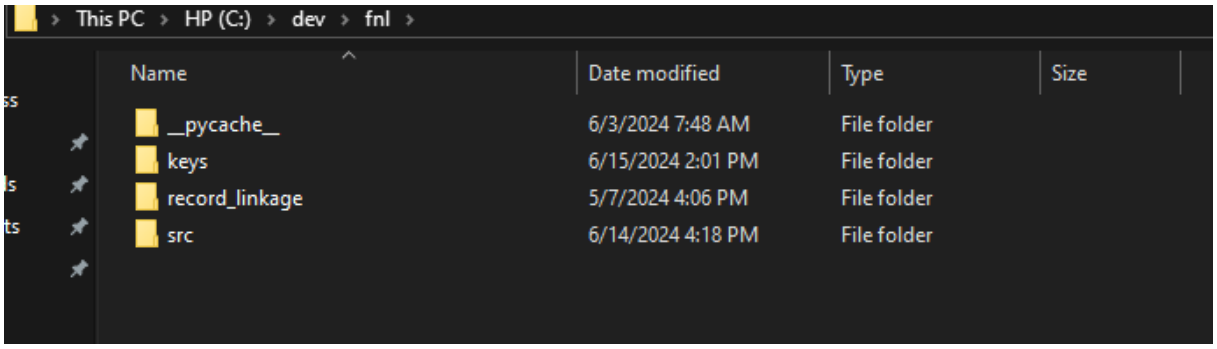

Figure 4 : dossiers principaux

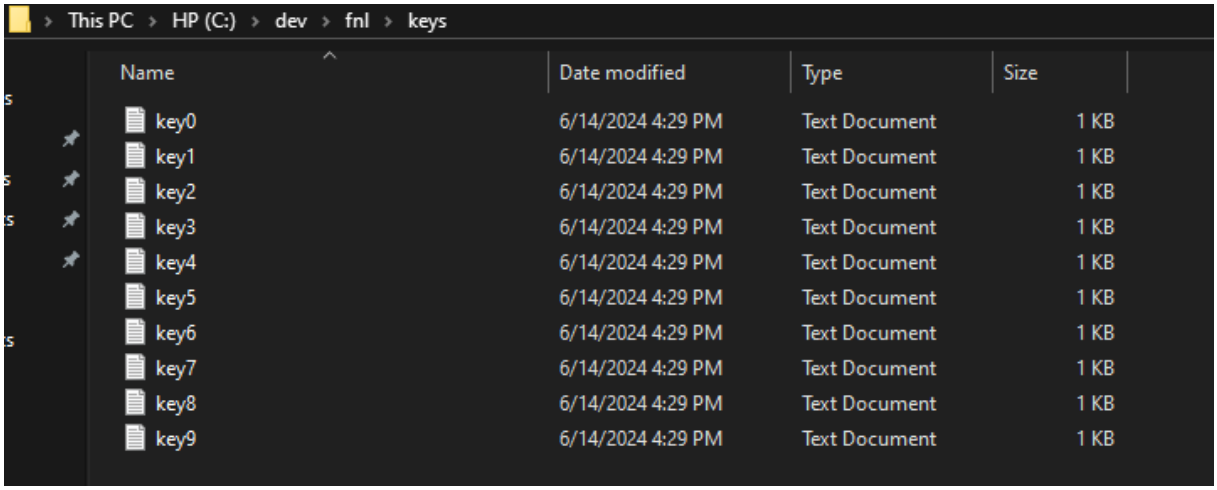

Figure 5 : Structure d'une clé

Exécuter le script Python en tapant la commande "python main.py <dataset>". La première tâche effectuée est la phase d'indexation en utilisant les clés initiales.

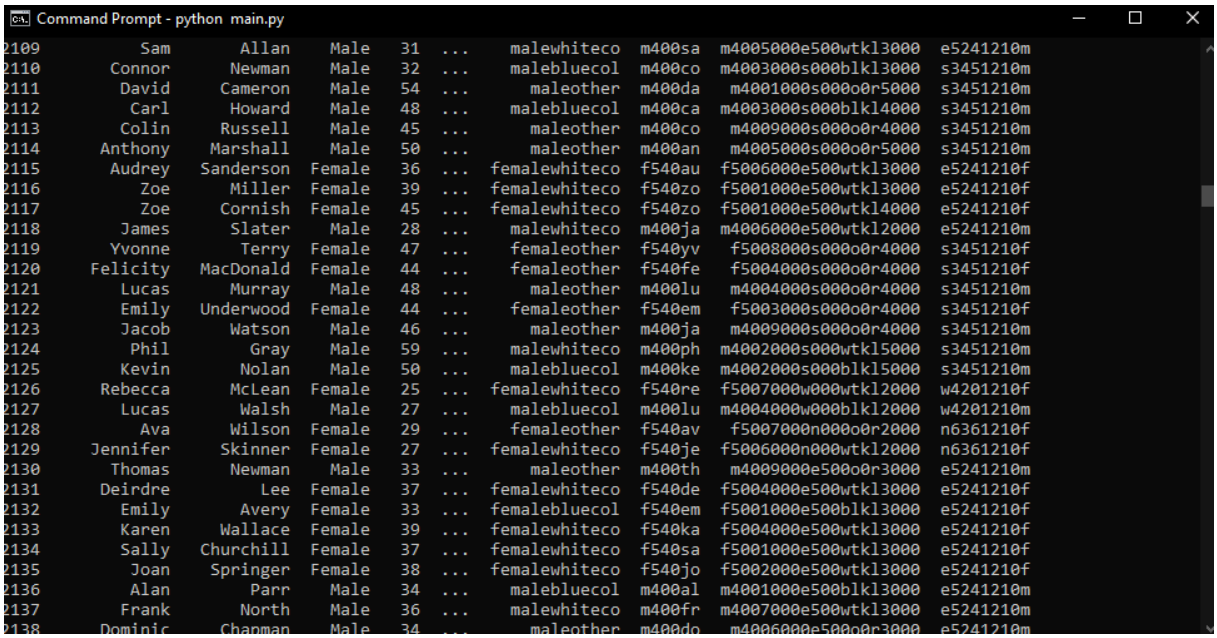

Figure 6 : Exécuter le script Python

Après la phase d'indexation, chaque clé sera évaluée et recevra un score (entre 0 et 1). Si les blocs générés par une clé n'ont aucune correspondance ou si les blocs sont très grands ou très petits, le score sera alors 0 et la clé sera remplacée.

```
[4040 rows x 19 columns]
Key0 ---------  0.5925925925925926
Key1 ---------  0.05128205128205127
Key2 ---------  0.4489795918367347
Key3 ---------  0
Key4 ---------  0.05128205128205127
Key5 ---------  0
Key6 ---------  0.8125000000000001
Key7 ---------  0
Key8 ---------  0
Key9 ---------  0.05128205128205127
```

Figure 7 : blocs générés par une clé

Les blocs seront stockés dans un dossier portant le nom de la clé.

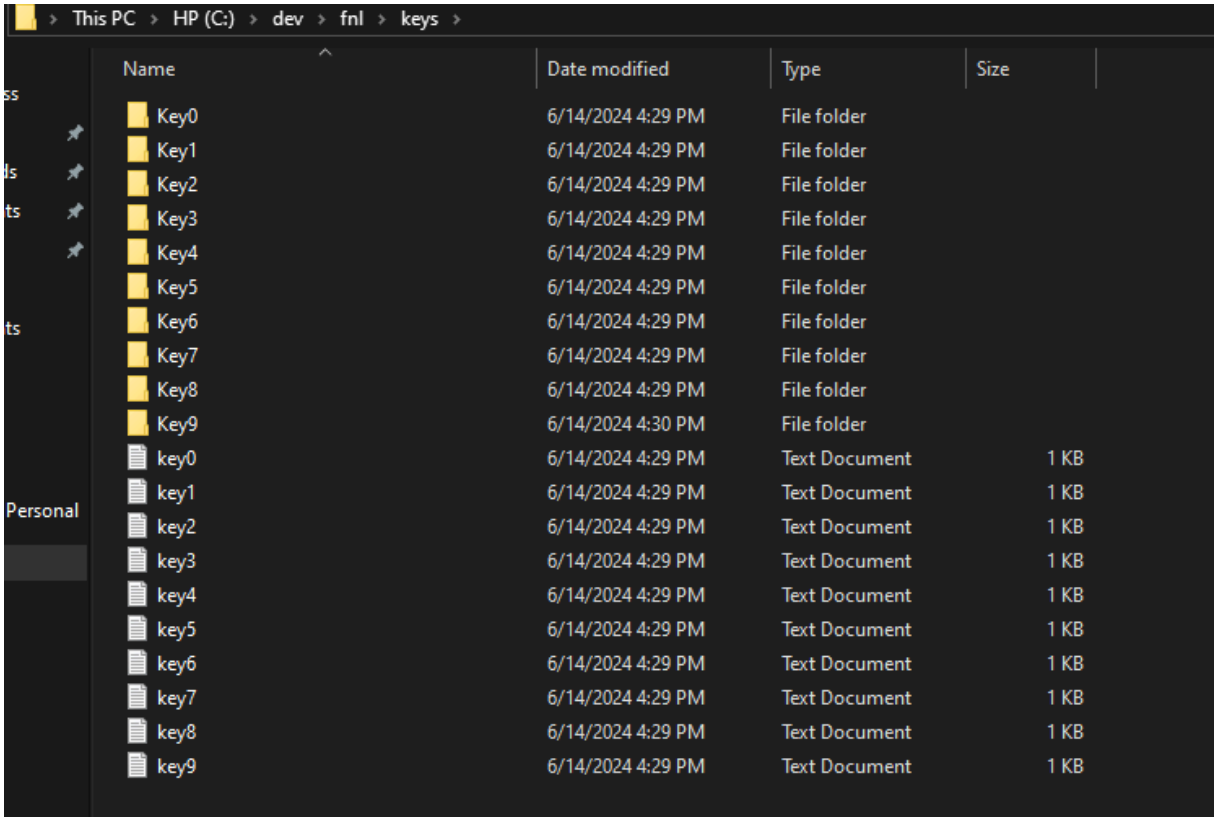

Figure 8 : Les blocs

Voici un exemple d'un bloc qui regroupe tous les enregistrements similaires en fonction de la clé sélectionnée.

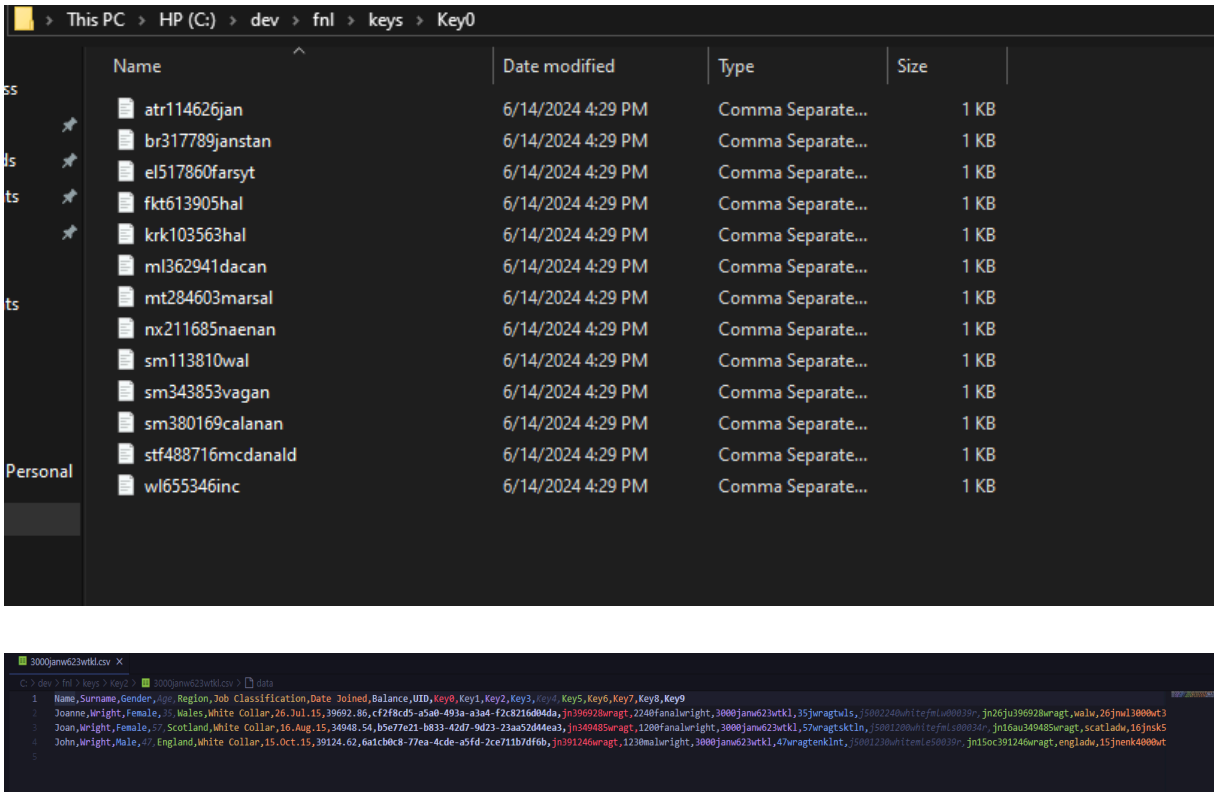

Figure 9 : exemple d'un bloc

Juste après avoir terminé l'évaluation des clés initiales, les mêmes clés sont utilisées pour lancer l'exécution de l'algorithme BES pour un nombre donné d'itérations. À chaque itération, il effectue trois opérations principales : select, search, swoop.

En se dirigeant vers la meilleure combinaison de clés à chaque itération jusqu'au résultat final.

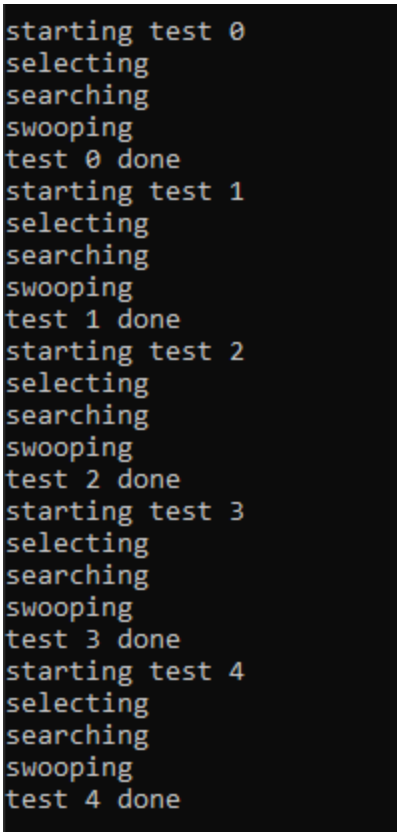

Figure 9 : Le processus de l'algorithme BES à chaque itération.

Le résultat final est un ensemble des meilleures clés obtenues par l'algorithme BES, garantissant d'obtenir au moins une clé avec un meilleur score que la meilleure précédente.

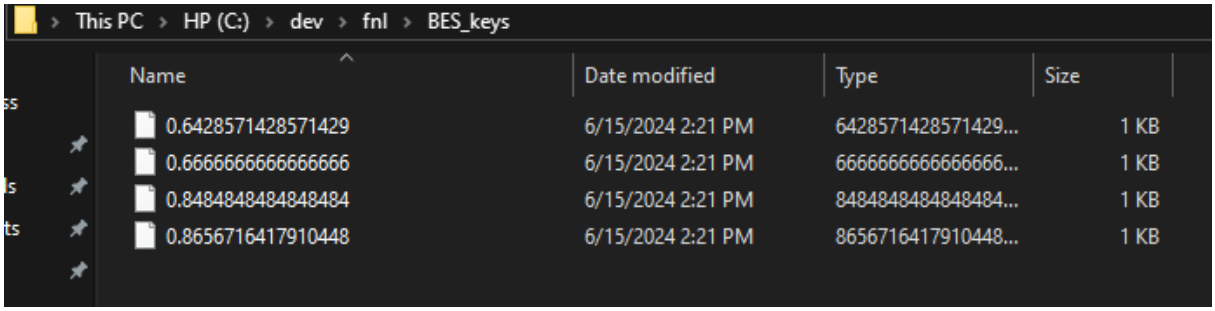

Figure 10 : Le résultat final est un ensemble des meilleures clés obtenues

Le résultat final est un ensemble des meilleures clés obtenues par l'algorithme BES, garantissant d'obtenir au moins une clé avec un meilleur score que la meilleure précédente.

3.7. Analyse des résultats

Dans cet exemple d'exécution (algorithme BES), nous pouvons voir clairement les améliorations de la qualité des clés après chaque essai à partir de la clé initiale, les opérations effectuées qui est censé aller vers une meilleure solution chaque itération peut conduire à trouver des clés similaires avec le même score d'évaluation.

Nous voyons également que certaines clés sont conservées telles qu'elles sont puisqu'il n'y a pas de meilleures clés trouvées pendant ce test.

	Initial keys	Test 1	Test 2	Test 3	Test 4	Test 5
Key 0	0.61	0.61	0.61	0.61	0.72	0.72
Key 1	0.37	0.78	0.78	0.88	0.88	0.88
Key 2	0	0	0.88	0.88	0.88	0.88
Key 3	0.13	0.80	0.80	0.80	0.80	0.80
Key 4	0.62	0.62	0.62	0.64	0.67	0.67
Key 5	0.34	0.67	0.72	0.72	0.72	0.72
Key 6	0.53	0.53	0.76	0.78	0.78	0.78
Key 7	0.76	0.76	0.83	0.88	0.88	0.88
Key 8	0.79	0.79	0.79	0.79	0.79	0.79
Key 9	0.08	0.59	0.59	0.67	0.67	0.67

Tableau 4 : de résultat

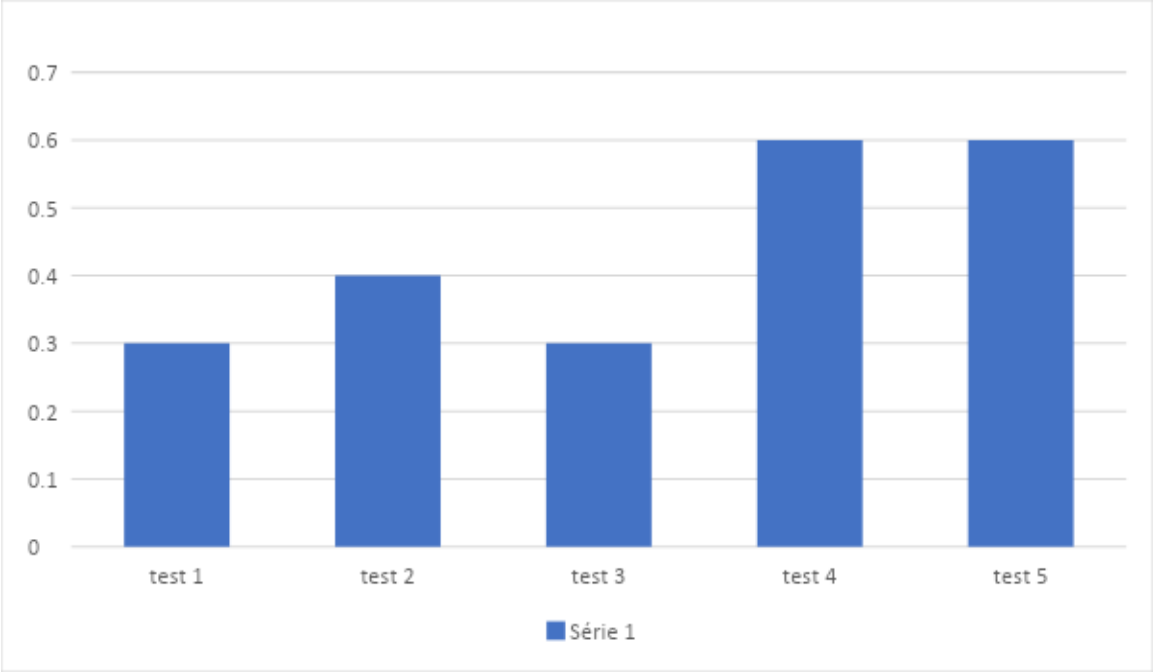

Figure 11 : Affiche les résultats du record (Graphe)

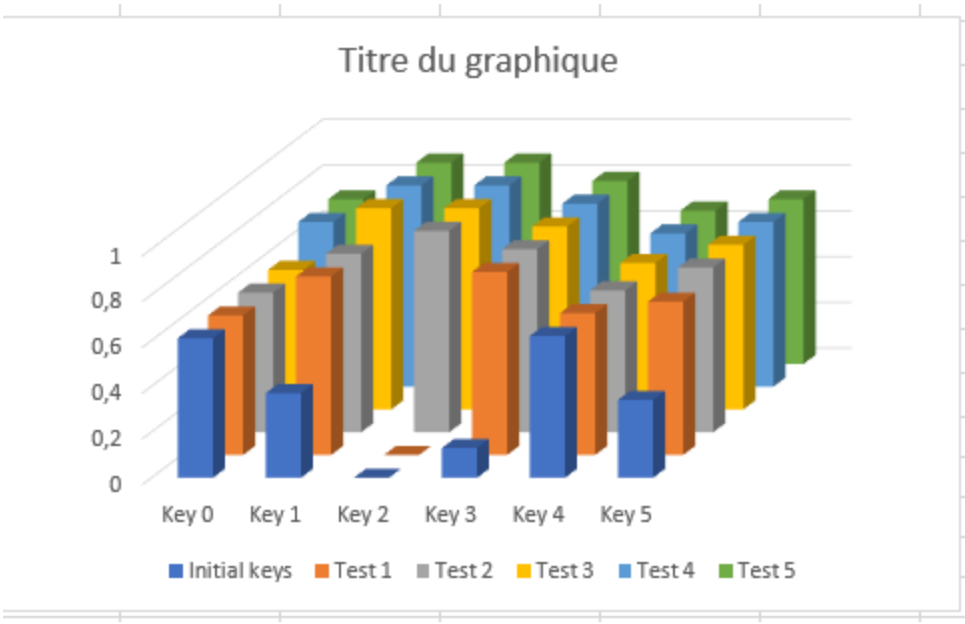

Figure 12 : Affiche les résultats du record (Graphe)

3.8. Conclusion

En conclusion, ce chapitre a présenté l'environnement et les technologies utilisés pour le développement de notre application de couplage d'enregistrements, ainsi que notre contribution au domaine à travers des solutions innovantes pour les défis rencontrés. Les captures d'écran fournies illustrent les résultats de nos efforts et l'aboutissement de notre projet.

4. Conclusion Générale

La qualité des données est un élément crucial pour assurer la fiabilité et l'exactitude des informations utilisées dans divers domaines. À travers ce travail, nous avons exploré les multiples dimensions de la qualité des données et l'importance de maintenir des standards élevés pour éviter les problèmes courants associés à la non-qualité des données. Les critères et les dimensions de la qualité des données ont été analysés, soulignant leur pertinence dans la gestion de l'information.

Nous avons également examiné les approches générales pour détecter et corriger les problèmes de qualité des données, mettant en avant l'importance de l'adoption de méthodologies rigoureuses pour garantir des données fiables et précises. Les objectifs de la qualité des données ont été discutés, soulignant leur rôle essentiel dans la prise de décision éclairée et la réalisation de recherches solides.

Le deuxième chapitre s'est concentré sur le Record Linkage, une méthodologie cruciale pour l'intégration et la liaison de données provenant de différentes sources. Nous avons détaillé les étapes du Record Linkage, incluant le nettoyage et la normalisation des données, l'indexation, et le processus de matching. Ces étapes sont essentielles pour garantir que les enregistrements correspondants soient correctement identifiés et associés, minimisant ainsi les erreurs et améliorant la qualité des ensembles de données fusionnés.

Enfin, dans le troisième chapitre, nous avons présenté une implémentation pratique et une expérimentation de la méthodologie de Record Linkage. Nous avons décrit l'environnement de développement et l'application développée, démontrant l'efficacité des techniques abordées dans les chapitres précédents. Cette application pratique a permis de tester et de valider les concepts théoriques discutés, offrant des perspectives tangibles sur les défis et les solutions liés à la qualité des données et au Record Linkage.

En conclusion, ce travail a mis en lumière l'importance critique de la qualité des données et du Record Linkage dans la gestion et l'analyse des informations. En adoptant des pratiques rigoureuses et des méthodologies éprouvées, il est possible de surmonter les défis associés à la non-qualité des données et d'assurer des résultats fiables et exploitables. Les perspectives futures

incluent l'amélioration continue des techniques de qualité des données et du Record Linkage, ainsi que l'intégration de nouvelles technologies pour répondre aux besoins croissants de précision et de fiabilité des données dans un monde de plus en plus axé sur les données.

5. Bibliographie

[1] C. Toulemonde, JEMM research_Informatica, “Exploiter le capital de votre organisation”, Un livre blanc de JEMM research - Des données de qualité, 2008, pp. 1-26.

[2] Franck Rgnier-Pcastaing, Michel Gabassi, et Jacques Finet. Enjeux et mthodes de la gestion des données. Paperback, 2008.

[3] Jamm. Des Données Qualités : Exploitez le capital de votre organisation. livre blanc, janvier 2008.

[4] Barrau Delphine et al: Gestion de la qualité des données ouvertes liées – État des lieux et perspectives

[5] Gestion de la qualité des données : Quoi, pourquoi, comment et meilleures pratiques - Data Ladder

[6] Laure Berti-Equille. Qualité des données. Techniques de l’ingénieur. Informatique, 2006.

[7] An Automatic Blocking Keys Selection For Efficient Record Linkage , International Journal of Organizational and Collective Intelligence ,Volume 11 • Issue 1 • January-March 2021

[8] Hamid Naceur BENKHALED A novel approach to improve the Record Linkage process

[9] [Rajkovic Jankovic 2007] P Rajkovic and Jankovic. Adaptation and application of Daitch-Mokotoff Soundex algorithm on Serbian names. In XVII Conference on Applied Mathematics,volume 12, 2007.

ملخص

أحد العمليات الأساسية في جودة البيانات هو التوافق بين السجلات (RL)، يكتشف RL التكرارات في مجموعة أو أكثر من مجموعات البيانات. المرحلة الأكثر أهمية هي الحجب، التي تقلل من التعقيد عن طريق تقسيم البيانات إلى كتل، مما يحد من المطابقة بين السجلات داخل نفس الكتلة. اختيار أفضل مفاتيح الحجب أمر صعب وغالبًا ما يتم بواسطة خبير. تقترح هذه المقالة نهجًا جديدًا غير مراقب لاختيار مفاتيح الحجب تلقائيًا، استنادًا إلى خوارزمية التحسين BES للصقور الصلعاء. تظهر التجارب على مجموعات بيانات حقيقية أن BES يتفوق على النهج الحالي ويجد أفضل مفاتيح الحجب.

Resume

L'un des processus clés en qualité des données est le couplage d'enregistrements (RL), ou résolution d'entités. RL détecte les doublons dans un ou plusieurs ensembles de données. L'étape la plus critique est le blocage, qui réduit la complexité en divisant les données en blocs, limitant ainsi la mise en correspondance aux enregistrements du même bloc. Choisir les meilleures clés de blocage est difficile et souvent fait par un expert. Cet article propose une nouvelle approche non supervisée pour la sélection automatique des clés de blocage, basée sur l'algorithme d'optimisation BES pour les pygargues à tête blanche. Les expériences sur des ensembles de données réels montrent que le BES dépasse les approches existantes et trouve les meilleures clés de blocage.

Mots clés : Qualité des données, Algorithme BES, clés de blocage, couplage d'enregistrement.

Abstract

One of the key processes in data quality is record linkage (RL), or entity resolution. RL detects duplicates within one or more datasets. The most critical stage is blocking, which reduces complexity by dividing data into blocks, limiting matching to records within the same block. Choosing the best blocking keys is challenging and often done by an expert. This article proposes a new unsupervised approach for automatic selection of blocking keys, based on the BES optimization algorithm for bald eagles. Experiments on real-world datasets show that BES outperforms existing approaches and identifies the best blocking keys.

Keywords: Data quality, BES algorithm, blocking keys, record linkage.