Université de Lille 3

Rapport de TER

Nettoyage d'une Base de données

DEREGNAUCOURT Paul
POULAIN Romain

Tuteur : M. STAWORKO

Octobre 2019 — Mai 2020

1 Introduction

Les problèmes liés au nettoyage de données sont apparus au début des années 2000 avec l'explosion d'Internet et des entrepôts de données. A cause du volume croissant de données, il y a eu un problème dans les qualités des données des différentes bases.

La qualité des données correspond à la conformité des données aux usages prévus. En effet la qualité des données est très utile pour une entreprise. Elle lui permet évidemment de gagner de l'argent. Par exemple une entreprise comme Free collecterait les données de numéro de téléphone de ses clients. Ces données seraient utilisées dans les forfaits téléphoniques de leurs clients. Mais ces données seraient de mauvaise qualité suite à un problème de détection des antennes. L'entreprise récupèrerait donc de mauvais numéros de téléphone. Ce serait un très gros problème car ils pourraient perdre des millions d'euros car les numéros de téléphone dans la base de données seraient de mauvaise qualité or l'entreprise de téléphonie mobile aurait besoin des bons numéros de téléphone associés aux clients (Par exemple un appel d'une 1 heure est décompté sur une personne avec un forfait bloqué donc il y a des frais téléphoniques que Free comptabilise pour la mauvaise personne). Pour réparer ces erreurs il faut donc débourser beaucoup d'argent et il faut également nettoyer la base de données qui est de mauvaise qualité. C'est un gros coup dur pour les entreprises. La qualité de données permet donc également de gagner du temps parce que si les données sont de mauvaise qualité, il faudra améliorer la qualité pour que le travail de l'entreprise se fasse dans de bonnes conditions et qu'il n'y a pas à réparer les erreurs que peuvent engendrer beaucoup de perte de temps et d'argent.

Il peut y avoir plusieurs problèmes de données qui affectent la qualité de données. Premièrement il y a des problèmes de syntaxe dans l'écriture de données. Par exemple mettre l'âge d'une personne dans la colonne du sexe de la personne, des erreurs de formatage avec des données qui ne sont pas du même format. Il y a aussi des problèmes d'irrégularité par exemple des distances exprimées en kilomètres mais des données sont ecrie en miles. Deuxièmement il y a des problèmes de sémantiques. Elles concernent notamment une violation des contraintes d'intégrité de la base, par exemple une personne avec un âge négatif. Il y a des problèmes de contracdiction, par exemple l'âge ne correspond pas avec la date de naissance. Il y a des problèmes de qualité de données de duplication, c'est à dire plusieurs données relatives à une même données. Et enfin des problèmes de données invalides. Dernièrement il y a des problèmes de qualité de données de couverure. comme valeurs manquantes, par exemple il manque la date de naissance d'un personne ou données manquantes, par exemple aucun âge est spécifé dans la base de données.

Dans notre cas nous allons nous concentrer sur l'erreur de duplication des données. Principalement, on parle des doublons, ce sont des erreurs qui représentent deux fois le même élément. On estime environ à 25% des données critique. De plus, 2% des enregistrements dans un fichier client peuvent devenir obsolètes en un mois (changement d'adresse, de numéro), soit environ 50% de données erronées dans un fichier en 2 ans. Ceci est très dérangeant car des données sales coutent excrément cher. Imaginons des clients dans une banque qui aurait des identités en double ou encore des envoies de catalogues en double ou à une mauvaise adresse, il faudrait alors renvoyer le catalogue à la bonne adresse, ce qui à pour conséquence une perte de temps et d'argent. Il existe un site qui recueille toutes sortes de livres, page web, télévisions. Il s'agit Open Library Data Dumps basée à San Franciso, en Californie avec des satellites dans le monde entier. Ceci à pour but de rendre toutes les œuvres publiées de l'humanité accessibles à tous dans le monde afin de construire un accès universel à toutes les connaissances. Dans ce projet, nous travaillerons sur une base de données qui constitue des livres, des auteurs.

Cependant la base de données Open Library Data Dumps est impure, ce qui veut dire qu'elle n'est pas nettoyée et qu'elle possède de nombreuses données incohérentes, de doublons et de fausses informations. La détection automatique des doublons est difficile : premièrement, les représentations en double ne sont généralement pas identiques mais légèrement différentes dans leurs valeurs. Deuxièmement, en principe, toutes les paires d'enregistrements devraient être comparé, ce qui est impossible pour de gros volumes de données.

Pour parvenir à nettoyer notre base de données Open Library Data Dups, nous allons procéder en 3 étapes pour effectuer la déduplication de nos données :

1. Identification des doublons

   Pour réaliser ce projet, nous allons commencer par étudier la base de données, pour cela nous aurons besoins de rendre la base exploitable, car celle-ci est sous format json, ce qui n'est pas pratique pour l'exploiter afin de la nettoyer. Une fois que nous aurons la base souhaités, sous formes de tables exploitables alors nous pourrons commencer à l'étudier. Tout d'abord, nous allons étudier les doublons de notre base de données. En effet, cette base en contient, notamment quand il s'agit des titres de livres ou des noms et prénoms d' auteurs. Donc le but de notre projet est de faire la déduplication de nos doublons. C'est un moyen de stockage de données qui consiste à factoriser des séquences de données identiques pour libérer de l'espace dans la base de données. Donc ici nos doublons seront analysés élément par élément pour savoir s'ils sont suffisamment proches pour être considéré comme le même objet. Pour cela, nous allons utiliser des fonctions de similarités qui auront pour but de comparer les objets entre eux avec un certain seuil de similitude pour déterminer si 2 objets sont considérés comme similaires. Par exemple pour allons prendre la distance entre deux mots.(John et Johny par exemple).

2. Clustering

   Si nous obtenons des données similaires nous allons les placer dans ce qu'on appelle un graphe de similitude. Tous les objets qui auront un seuil suffisamment proche seront clusturisé, c'est-à-dire qu'on va les regroupés ensemble et ils seront proches dans le graphe de similitude. Ensuite de par ce graphe nous pourrons déterminer si les groupes d'objets sont suffisamment similaires pour être considéré comme des doublons.

3. Fusion des clusters

   Si c'est le cas nous procéderons à la fusion de nos groupes (clusters). En effet nous allons fusionner les doublons pour que ces groupes d'objets similaires ne soient plus qu'un seul et unique objet. C'est après cette étape que nous allons commencé à nettoyer notre base pour qu'elle soit propre.

Dans un premier temps nous allons étudier l'implémentation de Library Open Data Dups, en format json, dans une base de données puis prendre connaissance de cette dernière et nous allons identifier les doublons notamment avec des fonctions de similitudes. Dans un deuxième temps nous allons etiqueter ces doublons, puis procéder à leur regroupement (clustering) dans un graphe de similitude. Dans un dernier temps nous allons fusionner nos groupes (clusters) pour otenir une base de donnée nettoyée.