Période de travail : 10/10/2022 => 23/10/2022

Manager de la période : Quentin

Collaborateurs : Lucas, Quentin et Claude.

Samuel :

Les tâches :

- Faire les recherches sur les conditions d'utilisation d'IMDB.
- Écrire l'analyse juridique des conditions d'utilisation de la base de données de IMDB.

Ressenti **:**

Durant ces deux dernières semaines, j'ai travaillé sur l'analyse juridique des conditions d'utilisation de la base de données de IMDB avec mon manager. J'ai effectué des recherches sur les conditions générales d'utilisation de IMDb et sur les droits relatifs aux droits d'auteur et à l'exploration de données. Cela m'a permis de recueillir toutes les informations nécessaires pour écrire un rapport juridique expliquant que nous avons le droit d'utiliser la base de données de IMDb si on respecte leurs conditions. Ensuite j'ai rédigé ce rapport avec l'aide de mon manager Quentin. Cette tâche était très intéressante à faire car cela m'a permis d'utiliser et d'apprendre des notions juridiques dans le cadre de cette SAE.

Lucas :

Les tâches :

- Faire les personas
- Aider Claude pour son problème de syntaxe du fichier title_crew

Ressenti :

J'ai fait les personas plus rapidement, car ce n'étais pas très compliqué. J'ai essayé de faire plusieurs profils bien différents d'entre eux, c'était assez marrant à faire. J'ai donc aidé Claude, car il y avait un problème avec les fichiers tsv. On a eu beaucoup de mal à résoudre ce problème, mais nous l'avons résolu. ça nous a forcé à chercher de nouvelles choses, c'était très intéressant.

Claude :

Les tâches :

- Faire les différents Personas.
- Faire marcher la commande Copy de PostgreSQL avec les différents fichiers csv
    - Syntaxe de la commande à regarder notamment pour les types tableaux
- Ecriture des fichiers csv en Python :
    - Faire correspondre les fichiers csv que nous avons à la syntaxe que prends en compte PostgreSQL

Ressenti :

Les Personas n'ont pas posé de réel problème cependant les deux autres tâches ont été plus complexes que prévues surtout pour la partie Python. Je le savais déjà depuis le début de la SAE que les fichiers qu'on avait en csv ne pouvaient pas être mis directement sur PostgreSQL à cause des types tableaux. Par conséquent, je dois trouver une syntaxe me permettant d'inclure les types tableaux dans la base de données notamment avec la commande Copy.

Voici un exemple de la syntaxe que doit prendre une ligne du fichier title_crew :

- t0000007,`{`nm0374658`,`nm0005690`}`,{\N}
- tt0000008,{nm0005690},{\N}

Comme vous le voyez, ce ne sont pas des guillemets n'y des singles quote (‘ ‘) puisque cela ne marchera pas dans notre cas, la raison est simple à cause du type tableau, il peut y avoir plusieurs valeurs suivies par des virgules. Les virgules étant des délimiteurs, PostgreSQL pense qu'il y a plus de valeur que de donnée (nombre de colonnes du fichier (3)). Par exemple pour la première ligne, il y aurait 4 valeurs pour 3 colonnes ce qui n'est pas possible. Afin de régler ce problème, l'équipe et moi-même avons fait des recherches afin de corriger cela, nous avons donc fini par trouver un moyen qui est d'utiliser les options disponibles de la syntaxe de la commande Copy. Voici donc notre version de la commande Copy :

- COPY nomTable FROM 'C:\Program Files\PostgreSQL\13\scripts\t4.csv' csv DELIMITER ',' header QUOTE '`';

Problème étant résolu, il ne reste plus qu'à écrire sur le fichier csv en impliquant la syntaxe montrée dans l'exemple plus haut. Pour cela, j'utilise Python et notamment la librairie re qui me permet d'utiliser les regex. J'avais déjà commencé à coder lors de la dernière période de travail (14/09/2022 => 25/09/2022) certains changements ont été faits notamment par rapport aux librairies, mais l'idée reste la même. Il y a 4 fichiers csv qui utilise les attributs tableaux les données et le nombre de colonnes n'étant pas les mêmes, le code réalisé pour le fichier title_crew ne pourra pas être réutilisé sur certains autres fichiers csv par conséquent il se pourrait que nous ayons 3 ou 4 fichiers Python pour les différents fichier csv.