Gestion des données

Devoir maison

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2022-2023

Consignes

- À rendre pour le **16 avril**
- Au format pdf (il n'y aucune implémentation à faire, donc pas de notebook ou autre)
- En utilisant le lien suivant https://nuage.isir.upmc.fr/index.php/s/pMQgLeq9MPyLNeY, pas d'email

Remarque Ce devoir est intitulé *devoir* car il contient plusieurs questions pour détailler les étapes, mais tout est volontairement laissé assez libre, il y a plusieurs réponses possibles et des choix à faire (ça n'est pas un projet non plus, et il n'y a aucune implémentation demandée).

Contexte On s'intéresse à la modélisation d'une base de données relationnelle pour stocker les données d'entraînement d'un chatbot. Il ne s'agit pas d'un dataset statique construit une fois par toute mais d'une base de données qui va être enrichie au fur et à mesure des interactions avec le chatbot et qui servira à stocker l'historique des conversations avec les utilisateurs.

Pour assurer la qualité du futur dataset d'entraînement, les messages seront modérés: plusieurs personnes seront chargées de relire les conversations et d'étiqueter les messages.

Description de la base La base de données stockera donc les informations suivantes:

- *Comptes utilisateurs* On stocke seulement un identifiant et un nom.
- *Comptes modérateurs* On stocke seulement un identifian et un nom.
- *Prompts des utilisateurs* On stocke le texte entrée par l'utilisateur et l'auteur.
- *Réponses du chatbot* On stocke le texte de la réponse et le prompt qui a été donné en entrée.
- *Modération* Les modérateurs associent une note entre 0 et 10 aux réponses. Chaque message peut être évalué par un ou plusieurs modérateurs.

Cette description ne constitue pas un schéma complet, il faudra l'utiliser pour définir le schéma relationnel.

Question 1

Dessiner un schéma conceptuel des données.

Question 2

Dessiner un schéma logique des données.

Question 3

Écrire une requête pour extraire les réponses invalides (moyenne des notes <5).

Question 4

Écrire une requête pour extraire les prompts qui ont donné lieu à des réponses invalides (plus de 10 par exemple, peu importe).

Question 5

Écrire une requête pour extraire la liste des utilisateurs qui ont obtenu beaucoup de réponses invalides.

Question 6

Écrire une requête pour extraire la liste des modérateurs classant trop souvent des réponses comme invalides (plus de 80% du temps par exemple, peu importe).

Question 7

Proposer une simplification du schéma de la base de données.

Question 8

La structure précédente ne permet pas de reconstruire l'historique complet d'une conversation, seulement des couples (prompt, réponse). Modifier le schéma pour obtenir un historique complet (c'est à dire l'enchainement de toutes les (promp, réponse) d'une conversation).

Question 9

Écrire une requête pour récupérer une conversation complète.

Question 10

Discuter la pertinence d'une base de données relationnelles pour ce cas d'usage.