Sujet de stage M2 :

Enrichissement de méta-données dans un datalake

Contacts : Ce stage est en co-encadrement laboratoire/entreprise :

Contact LIP6 : Bernd Amann : bernd.amann@lip6.fr
Contact Zeenea : Julien Buret : jburet@zeenea.com

Entreprise

Zeenea propose une solution SaaS de catalogue de données qui aide les entreprises à accélérer leurs initiatives en matière de données. Notre plateforme basée sur le cloud offre une base de données fiable et compréhensible disponible avec un maximum de simplicité et d'automaticité. En quelques clics, vous pouvez trouver, découvrir, gouverner et gérer les informations de votre entreprise.

Contexte

La solution de catalogue de données connectée de Zeenea est basée sur 3 grandes fonctionnalités :

- Un système de scanner extensible permettant de collecter les métadonnées techniques (emplacement, schéma, type de source de données, …) et statistiques sur un grand nombre de sources de données tels que des bases de données relationnelles, des bases NOSQL, des datalakes, des systèmes de fichiers distribués, …
- Un studio de modélisation et d'enrichissement manuel du méta modèle qui permet de documenter de manière manuelle les objets synchronisés par le scanner et d'enrichir le méta modèle avec des concepts sémantiques ou métiers.
- Une interface graphique dédiée à l'exploration ou à la recherche d'informations dans le métamodèle.

Cette architecture permet à notre data catalogue, de documenter automatiquement et rapidement le patrimoine de données d'une entreprise, qui peut représenter des milliers de datasets, des centaines de milliers de champs synchronisés et des millions de propriétés. Une fois ces données « techniques » cataloguées, elles doivent être ensuite documentées, soit en valorisant leur description, leurs propriétés ou en **créant des liens avec des concepts sémantiques ou métiers**. La valorisation manuelle de cette documentation est très coûteuse à produire à cause de l'énorme volume d'objets à documenter et à maintenir à jour.

Un de nos enjeux pour 2022 est de proposer des solutions pour faciliter et/ou automatiser la production et la maintenance de cette documentation. Nous développons actuellement une nouvelle brique logicielle permettant de suggérer l'existence d'un lien entre différents objets du catalogue, par exemple plusieurs champs de base de données identiques ou un lien entre un

champ et un concept métier. Les approches choisies sont, pour le moment, simplement basées sur la similarité du nom ou de la description. En parallèle de cette première implémentation, l'objectif est de concevoir et mettre en œuvre des approches plus élaborées basées sur de l'apprentissage automatique.

Travail à faire

Vous devrez en premier lieu, constituer un état de l'art sur des approches de machine learning adapté à nos contraintes et de proposer un ensemble d'approches pertinentes.
En second lieu vous devrez prototyper les approches les plus prometteuses. Les prototypes devront prendre la forme d'un pipeline de machine learning complet.
Les données du catalogue sont exportées sous forme de fichiers structurés dans AWS S3, vous devrez

- Préparer les données pour l'apprentissage et la validation des modèles
- Proposer une méthode de validation et scoring
- Mettre en œuvre la solution proposée
- Collaborer à l'intégration et l'exploitation du modèle.

Les librairies et technologies utilisées durant ce stage seront définies lors de la constitution de l'état de l'art. Zeenea utilise AWS pour provisionner les infrastructures nécessaires à la constitution et au scoring des modèles.

Divers

Le stage se déroulera dans nos locaux 156 Boulevard Haussmann à Paris (présence à mi-temps) pour une durée de 5 à 6 mois. Indemnité de stage de 1500€ brut mensuelle plus ticket-restaurant.

Bibliographie

Riccardo Cappuzzo, Paolo Papotti, Saravanan Thirumuruganathan: Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. In International Conference on Management of Data (SIGMOD), pages 1335-1349, 2020.

V. Shah, J. Lacanlale, P. Kumar, K. Yang, A. Kumar. Towards benchmarking feature type inference for autoML platforms. In International Conference on Management of Data (SIGMOD), pages 1584–1596, 2021.

M.-A. Baazizi, C. Berti, D. Colazzo, G. Ghelli, C. Sartiani : "Human-in-the-Loop Schema Inference for Massive JSON Datasets", International Conference on Extending Database Technology (EDBT), pp. 635-638, OpenProceedings.org, 2020