Quantmetry
Building AI with pionneers

Stage : Data Engineer (H/F)
Validation automatique de la donnée

Qui sommes-nous ?

Nous pourrions commencer en nous présentant comme l'un des leaders sur le marché international du conseil en data, avec un chiffre d'affaires défiant ceux des G.A.F.A. Mais non. Nous préférons mettre en avant notre cadre de travail, nos réussites et surtout nos consultants. Allez venez, c'est par là…

Quantmetry c'est une centaine de personnes qui travaille de concert pour accompagner nos clients dans leurs réflexions, P.O.C., industrialisation autour de la Data, l'Intelligence Artificielle, le Machine Learning et la Big Data. Nous sommes ce qu'on peut appeler communément un cabinet de conseil pure player en Data.
N'étant pas sectorisés, nous intervenons sur différents sujets (voici une petite liste non-exhaustive) :

- La maintenance prédictive,
- La segmentation clients, le scoring,
- La détection de fraude, de corruption, de blanchiment d'argent,
- Le développement des architectures de plateformes Data,
- L'analyse de textes et d'images dans le cadre de la lutte contre le cancer du sein.

Dans le cadre de notre incessante croissance, nous recherchons des personnes motivées pour nous accompagner et ainsi travailler sur des projets de recherche.

Préalable

Quantmetry propose ci-dessous le volet R&D d'une offre de stage. Tout stagiaire entrant chez Quantmetry, outre le travail de R&D qui lui est proposé et pour lequel il sera encadré, aura aussi pour objectif de participer à certaines missions de conseil chez des clients variés, lui permettant d'aborder le monde du consulting.

Quantmetry
Building AI with pionneers

Contexte

La première étape de tout traitement des données, en particulier du cycle de vie des données, et par extension du cycle de vie des modèles, est la validation de la donnée. En particulier, au même titre qu'un code, la donnée doit pouvoir faire l'objet de tests unitaires automatiques afin de rejeter la donnée de mauvaise qualité et de prévenir les risques y afférant. Une méthode émergente de validation automatique repose sur un schéma de contraintes, capable de sauvegarder toutes les hypothèses faites sur un jeu de données, et de logger toutes les non-conformités observées. Plusieurs librairies en accès libre proposent ce genre de service :

- TDDA (Test-Driven Data Analysis), développé par Stochastic Solutions et le département de mathématique de l'université d'Edimbourg,
- TensorFlow Data Validation, développé par Google Research,
- Delta Lake, développé par l'éditeur de solutions Databricks,
- Deequ, developpé par Amazon Research et l'université d'Augsbourg

Les schémas de contraintes sont un élément clef du cycle de vie des données, et la promesse de ces outils est d'automatiser la découverte et la maintenance des contraintes auxquelles doivent obéir les données.

Les objectifs de ce stage sont :

- Prendre en main les 4 outils et les appliquer à des jeux de données simples en accès libre
- Comparer leurs fonctionnalités et lister leurs avantages/inconvénients
- Construire un guide méthodologique et un tutoriel de validation de donnée avec les outils jugés les plus pertinents
- Animer une formation en interne sur la validation automatique de la donnée