Fiche descriptive : TPT
Plateforme Collaborative d'Aide à l'Industrialisation des Modèles Prédictifs

Technologies utilisées: Python, JavaScript, Java, CSS, HTML

Membres du groupe (5 étudiants)

- **à définir, choix par les étudiants / répartition par les encadrants, pas encore décidé**

Volume horaire

~100h par étudiant (+ 3 jours par semaine pour les non alternants)

Encadrement universitaire

- **Gabriel Mopolo-Moke :** gabriel.mopolo@gmail.com
- **Grégory Galli :** greg.galli@gmail.com

Encadrement du cahier des charges

- **Alaaeddine Hammoudi :** a.hammoudi@probtp.com

Contexte technique

Python, Java, Stack programmation web python

Traitement des données (y compris les données non structurées : texte, image, …), création de modèle

Genèse du projet

Dans le cadre de l'intégration du langage python et ses dérivés dans la chaine de production de PRO BTP, nous souhaitons mettre en place un outil permettant l'accélération de la mise en production et la mise à disposition des travaux Data Science et Intelligence Artificielle en particulier et les travaux data en général.

En effet, pour réussir un projet data, il est primordial de faire particulièrement attention au pré traitement de la donnée et à la mise en production des modèles.

Collecter, agréger et auditer les données sont des étapes chronophages mais nécessaires si nous souhaitions générer un produit data de qualité.

Dans le cas des données entreprises, souvent nous appliquons les mêmes opérations de

prétraitement, selon le type de la donnés et/ou sa provenance. C'est à ce niveau qu'une première optimisation/standardisation est possible.

Une fois cette étape terminée, le data scientist s'intéresse au choix et à l'étude du modèle. C'est souvent l'étape la plus rapide du projet. L'automatisation et l'uniformisation est également possible ici aussi pour la phase de sélection du modèle.

Si le prototype est satisfaisant pour les différentes parties prenantes, le passage en production s'impose et selon l'agilité du SI de l'entreprise, ce passage sera plus ou moins compliqué, plus ou moins lent.

Un dernier objectif serait de faciliter cette mise en production, mise à jour et consommation du produit data.

Le but de la plateforme est de s'affranchir au maximum de certains processus techniques et de simplifier la création et l'adoption des modèles en production dans un SI donné.

Description de l'existant

- Bibliothèque de prétraitements pour les types de donnés suivants :
  - Tables de données (sous forme classique)
  - Données textuelles
  - Documents images

Objectifs

Création d'une plateforme permettant en s'appuyant sur Apache AirFlow ou Luigi:

- ✓ Benchmarker deux solutions de création de pipeline via Python : Luigi et AirFlow (autre idées seront les bienvenues)
- ✓ L'ingestion de la donnée de différentes sources
- ✓ Offrir des options de prétraitements récurrents
- ✓ Offrir la possibilité d'annoter les données (augmentation de la donnée)
- ✓ Création d'un pipeline de production
- ✓ Mise en production des modèles à consommer via Web Service et Docker
- ✓ S'intégrer facilement au SI d'une entreprise

Exigences proposées

1. Plateforme de préférence distribuée et résiliente
2. L'ingestion de la donnée de différentes sources : vérifier que les connecteurs les plus demandés sont disponibles, sinon les créer
   a. Connecteur pour HIVE

b. Connecteur pour MySql

3. Création de 4 pipelines typiques pour accélérer la mise en production des modèles et des restitutions dont un pipeline Spark
4. Création d'une interface d'annotation collaborative permettant aux membres d'une équipe de procéder à plusieurs type d'annotation : détection d'objet, classification, extraction d'entité

Architecture

Libre, mais prendre en compte un SI existant :

Serveur LDAP, Dataware, Cluster Spark

Contraintes de développement

- Linux, Java, Python, HTML, CSS, jQuery

Résultats attendus

- Documentation et code permettant la mise en place d'une architecture distribuée pour la mise en production de modèle DATA
- Code python sous format de package documenté / installable via le gestionnaire de package.
- Quelques démos sur des exemples concrets (data à communiquer ultérieurement)