Anass Mezroui

Data Scientist

mezroui.anass7@gmail.com

Data Scientist passionné par la Data et IA

Expérience professionnelle

08/2019 - Aujourd'hui **Data Scientist**, *Quinten Maroc*, (5 ans)

Collaborer avec l'équipe DEV pour exploiter la valeur des données et créer des solutions data-driven, nettoyer et structurer les données brutes, développer des modèles prédictifs à l'aide d'algorithmes de machine learning pour anticiper des comportements futurs ou automatiser certaines tâches et faciliter l'intégration de ces modèles dans les pipelines de production.

- Création de script **python** pour **Préparer** et **Transformer** des données textuelles ou tabulaires
- **Optimisation** des temps de calculs et de consommation des ressources des pipelines de **dataprep** existantes
- **Scraping** de différents fichiers **PDF** et de ressources **Web** pour collecter des **dataset** et créer des bases d'apprentissage pour des modèles de **deep learning**
- Créer des **Parsers** (analyseurs syntaxiques textuels) pour exécuter/évaluer des **grammaires formelles LL(1)** prédéfinies par le besoin.
- Explorer et concevoir architectures de **Modélisation** des données pour créer des modèles **Prédictifs**
- Explorer différentes **Métriques** pour **Evaluer** la performance des modèles prédictifs
- Collaborer avec l'équipe DEV afin d'**Intégrer** et **Déployer** les modèles dans les projets existants
- Participer aux différentes réunions d'analyse et de conception

02/2024 - 07/2024 **Stagiaire Data Scientist**, *Quinten Maroc*, (6 mois)

Développement d'un système de scoring basé sur des techniques de Text Mining pour créer un jeu de donnés à partir d'un dictionnaire de médicaments, élaborer une fonction pour annoter les données puis entrainer un modèle de machine learning SVM pour la régression sur la criticité des prescriptions médicales.

Éducation

2016 - 2019 **Formation d'ingénieur**, *Ecole nationale d'informatique et d'analyse des systèmes (ENSIAS)*, Rabat

Ingénierie Informatique et Systèmes Embarqués et mobiles

2014 - 2016 **Classe préparatoire aux grandes écoles (CPGE)**, *Lycée Omar Ibn Abdelaziz*, Oujda

MPSI/MP

2013 - 2014 **Baccalauréat**, *Oued Eddahab*, Oujda

Science mathématiques option A

Compétences techniques

Systèmes d'exploitation: **Linux (Debian, Ubuntu, Raspbian), Windows**

Langages de programmation: **Python , Java, C**

Bases de données: **SQL (PostgresSQL, MySQL, Oracle DB), NoSQL (Mongo DB)**

Data Transformation & analysis: **Pandas, Numpy, PySpark, Dask**

Machine Learning: **PyTorch, Tensorflow, Scikit-learn, XGBoost, Transformers**

Model Serving: **Flask, TensorFlow Serving**

Model Monitoring: **TrochDrift, AlibiDetect**

MLOps: **MLflow**

Data visualisation: **Grafana, Seaborn, Matplotlib**

Indexing: **FAISS, Solr**

Image Processing: **OpenCV, Scikit-image, Pillow/PIL**

Autre compétences: **Docker, Git, Jupyter Notebooks, Selenium, PyParsing, PdfQuery, RegEx**

Compétences générales

Data Preparation, Data Visuallisation, Machine Learning, Statistical Analysis, Time-Series Forecasting, Classification, Regression, Clustering, Dimensionality Reduction, Natural Language Processing, Text Embeddings, Semantic Search, Text Parsing, Image Processing, Unit Testing, Web Scrapping, Agile Methodologies

Algorithmes

Decision Trees, Random Forest, Support Vector Machines (**SVM**), Linear Regression, Logistic Regression, Gradient Boosting, K-Means, Synthetic Minority Over-sampling Technique(**SMOTE**), Principal Component Analysis (**PCA**), Neural Networks (**CNN, RNN, LSTM, Autoencoders**), Bidirectional Encoder Representations from Transformers(**BERT**)

Projets

Système de Recommandation

Implémentation et déploiement d'un système de recommandation pour des produits cosmétiques sur un site e-commerce en se basant sur une approche de **Collaborative Filtering**. À partir d'une matrice d'interactions (**Utility Matrix**) entre les utilisateurs et les produits (les évaluations lors des achats), on a appliqué une **réduction de dimensionnalité** en utilisant des techniques comme la décomposition en valeurs singulières (**SVD**) pour améliorer l'efficacité et **réduire la complexité** des calculs, Ensuite, on a construit une **matrice de corrélation** (*Pearson product-moment correlation coefficients*) entre les différents utilisateurs pour identifier les **utilisateurs similaires**. Cela a servi à recommander des **nouveaux produits** cosmétiques que d'autres **utilisateurs similaires** avaient déjà consultés ou achetés, optimisant ainsi l'expérience utilisateur. Le système a été évalué avec différentes métriques (**MSE**, R^2 **score, Precision@k, Recall@k**)

Prédiction des ventes trimestrielles

Pour ce projet de prédiction des ventes trimestrielles, j'ai développé un modèle de **deep learning** en utilisant **TensorFlow**. Le modèle s'appuie sur une architecture hybrid **CNN-LSTM** combinant une couche **CNN** pour l'extraction des caractéristiques à partir des sous séquences (15 jours) suivie d'un **LSTM** pour capturer les **dépendances temporelles** des 30 derniers jours (2 sous séquences) de données historiques. J'ai mis en place un suivi détaillé des **performances** du modèle via **MLflow**, en loguant les **hyperparamètres**, les **métriques** et les **résultats** de chaque entraînement. Après le **déploiement** du modèle sur **MLflow**, j'ai intégré un mécanisme de détection de drift à l'aide de **TorchDrift**, et j'ai également automatisé la remontée des résultats de drift (**drift score, p_value**) dans **MLflow** pour un **monitoring continu en production**. Le système a été évalué avec différentes métriques (**MSE**, R^2 score)

Moteur de recherche sémantique pour les textes de législation européenne

Dans le cadre de ce projet, j'ai conçu et mis en œuvre un système capable de répondre à des **requêtes en langage naturel** en s'appuyant sur les textes du **EUR-Lex** (*le point d'accès officiel et le plus complet aux documents législatifs de l'UE*). J'ai d'abord créé un dataset d'entraînement à l'aide de **BEIR** (*Benchmarking Information Retrieval est un cadre de référence pour évaluer les modèles d'extraction d'informations qui propose différents outils pour générer des requêtes à partir d'un corpus*). Ensuite, j'ai effectué un **fine-tuning** d'un modèle pré-entraîné **Sentence-BERT** (**Bi-Encoder**) sur ces données afin d'améliorer la qualité des **vecteurs d'embeddings**. Enfin, j'ai exploré plusieurs pistes de **déploiement** du système (**FAISS, Solr**) pour indexer la base de données des vecteurs et récupérer les **top 10 résultats** les plus **pertinents** en utilisant **cosine similarity** comme métrique.

Langues

Arabe	**Maternelle**
Anglais	**Couramment** TOEFL iBT Novembre 2018 score: 86
Français	**Couramment** TCF Octobre 2018 niveau C2

Centre d'intérêt

Natation, Raod trip à Moto.