Joseph Assu Ondo

Disponibilité & Localisation	Contact
Disponibilité : **immédiate** Basé à : **Paris** Mobilité : **France, Europe**	Tel : **+33 6 58 66 80 03** e-mail : **assujoseph@yahoo.fr**

Data Scientist Senior

Machine Learning / NLP / Python / Pilotage projet

6 ans d'expérience

Compétences clés (étoiles de 1 à 5)

Data Science		**Technologies Big Data**	
Machine Learning	★★★★	Spark (PySpark)	★★★
NLP	★★★★	Hadoop	★★
Exploration statistique	★★★★	Hive	★★
Séries temporelles	★★★★		
Deep Learning	★★★		
Recherche Opérationnelle	★★★		
Langages de programmation		**Outils et environnements**	
Python	★★★★★	GCP	★★★
R	★★★	ElasticSearch	★★★
SQL	★★★★	Git	★★★★
Bash	★★★	PostgreSql – MySql	★★★★
C - C++	★★	Databricks	★★★
Java	★★	Linux / Unix	★★★
Gestion de projet		**Langues**	
Autonomie	★★★★	Anglais	★★★★
Agile / Scrum	★★★	Espagnol	★★★
Objectifs & suivi	★★★★		
Documentation & présentation	★★★★		
Management d'équipe	★★★		
Data Visualisation		**Certifications**	
Python Plotly Dash	★★★★	▪ Data Engineering with Google Cloud - *Coursera licence : 5VSA5WUC85YB*	
R Shiny	★★	▪ Deep Learning Specialization – *Coursera licence : U3SL4656WXQH*	
Tableau	★★★	▪ Spark and Python for Big Data with PySpark – *Udemy licence : UC-c96d6403-da78-4d9a-a78c-6218d9321443*	
Qlik Sense	★★	▪ Interactive Python Dashboards with Plotly and Dash – *Udemy licence : UC-2d9dd7c9-360e-4080-bdc0-d37210079303*	

Depuis 05/2019 : Consultant Data Scientist chez Starclay

Projet interne Starclay : Pilote projet - Data Scientist NLP

03/2021 – 05/2021

Outil de recommandation de CV sur des appels d'offre : développement d'un outil visant à :

- extraire les informations de CV et appels d'offres et à les structurer,
- harmoniser la présentation de ces informations,
- recommander les CVs pertinents pour un appel d'offre donné.

Interlocuteurs : métiers / Data scientist / Développeur full-stack

Equipe : 3 Personnes
Equipe Projet : 2 Data Scientists / 1 Développeur full-stack

Tâches effectuées :

- Pilotage de projet : animation d'ateliers de cadrage des besoins fonctionnels et techniques, réalisation des planning, présentation des résultats, réalisation du suivi agile via Gitlab Issues
- Tech lead : définition du workflow de développement, validation des merge requests (ou pull requests), documentation technique
- Développement de fonctionnalités
- Réalisations techniques : module de structuration de CV via la recherche d'entités et de mots clés, classification supervisée de sections de CV, indexation ElasticSearch, interface graphique, moteur de recherche

Environnement technique	Python, Pandas, scikit-learn, nltk, Spacy, SQL, PostgreSQL, Gitlab
Modèles Machine Learning et NLP	NER (Named Entity Recognition), classification supervisée, vectorisation & embedding contextuels, Regex, ElasticSearch

Direction Interministérielle du Numérique : Pilote projet - Data Scientist NLP

03/2020 – 03/2021 (12 mois)

Pilotage en parallèle de 2 projets visant à implémenter des outils d'extraction d'informations à partir de données textuelles en utilisant les approches Deep Learning et NLP. pour le compte de 2 administrations de l'Etat : Conseil d'Etat et DGCL (Direction Générale des Collectivités Locales)

Juradinfo-IA – Conseil d'Etat : rattachement de requêtes (plaintes contre les administrations publiques) aux séries nationales en cours en rapprochant leurs faits, leurs moyens et leurs conclusions.

ACLIA – DGCL : correction des métadonnées (nature, matière, objet) des actes rédigés par les collectivités locales et à destination des préfectures ; puis détection au sein des actes des éléments pertinents pour le contrôle de légalité.

Interlocuteurs : DSI, magistrats, greffiers, préfectures

Equipe : 4 Personnes
Equipe Projet : Pour chaque projet 1 data scientist sénior (moi) / 1 data scientist confirmé / 1 développeur full-stack

Tâches effectuées :

- Pilotage de projet : animation d'ateliers de cadrage des besoins fonctionnels et techniques, réalisation des planning, présentation des résultats, responsable de la communication avec les interlocuteurs projet, réalisation du suivi agile via Gitlab Issues (projet 1) et Trello (projet 2), animation des daily & weekly meeting
- Tech lead : définition du workflow de développement, validation des merge requests (ou pull requests), documentation technique.
- Développement de fonctionnalités
- Réalisations techniques : module d'océrisation, pipeline d'ingestion de données en BDD, modélisation Machine Learning et Deep Learning, indexation ElasticSearch, interface graphique, moteur de recherche.

Environnement technique	Python, BDD SQL (Postgres, MySql), Gitlab, Tesseract (moteur d'océrisation de documents), Apache Tika, Linux, Windows, Pandas, Flair, Fastai, Natural Language Toolkit, Scikit-Learn, Spacy
Modèles Machine Learning et NLP	NER (Named Entity Recognition), LSTM, classification supervisée (Xgboost, SVM), vectorisation & embedding contextuels (tf-idf, Elmo, Bert, Glove, Vord2vec), regex, ElasticSearch

Direction Interministérielle du Numérique : Data Scientist NLP

12/2019 – 02/2020 (3 mois)

Marchés subséquents 9, 10, 11 & 12 : Rédaction de réponses à 4 appels d'offres (12 projets) pour plusieurs administrations de l'Etat. Ces projets permettent à l'Etat d'expérimenter l'intelligence artificielle (IA) afin de capitaliser sur les connaissances des agents et de fluidifier les processus internes.
Equipe : 6 Personnes
Equipe projet : 4 data scientists / 1 concepteur-développeur / 1 commercial et le directeur du cabinet

Tâches effectuées :

- Etat de l'art des méthodes IA de résolution des problématiques posées : embedding de textes, algorithmes Machine Learning, algorithmes Deep Learning, Named Entity Recognition (NER), Entity Linking, Learning to rank
- Etat de l'art du socle technique de résolutions des problématiques posées : océrisation, extraction données PDF, moteurs de recherche, conteneurisation, bases de données relationnelles, bases de données orientées graphes,
- Benchmarking des différentes solutions existantes : librairies, outils, langages de programmation, bases de données
- Présentation des approches de réalisation technique

Domaines concernés	Economie - Finance, Santé, Sécurité intérieure, Collectivités territoriales
Résultats obtenus	Appels d'offres 9 et 10 remportés, représentant 8 projets NLP

Autorité de Sûreté Nucléaire : Data Scientist NLP

09/2019 – 11/2019 (3 mois)

Implémentation d'un moteur de recherche : Elaboration d'un moteur de recherche thématique sur les lettres résultant des contrôles d'inspecteur

Interlocuteurs : DSI, coordinateur projet

Equipe Projet : 1
Equipe Projet : 1 pilote de projet / 1 data scientist confirmé (moi) / développeur full-stack

Tâches effectuées :

- Configuration des cartes graphiques Nvidia sur un serveur Unix Centos afin d'accélérer l'apprentissage des modèles Deep Learning NLP
- Implémentation d'un modèle de reconnaissance d'entités nommées (NER) pour prédire au niveau paragraphe la catégorie et la sous-catégorie d'appartenance. Multi-labelling autorisé.
- Prédictions réalisées sur de nouvelles lettres stockées en base de données
- Mise en production via l'ordonnancement crontab des scripts de manière quotidienne

Environnement technique	Python Spacy, PostgreSql, SQL, Nvidia Cuda, Gitlab
Modèles mathématiques	Named Entity Recognition

Groupe Hospitalier Paris Saint-Joseph : Data Scientist NLP

06/2019 – 08/2019 (3 mois)

Outil de suivi de patients aux urgences : implémentation de modèles de détection de termes médicaux et de médicaments dans des rapports médicaux et implémentation de modèles de prédiction de durée de séjour

Interlocuteurs : métiers / Data scientist / Développeur full-stack

Equipe : 3 Personnes
Equipe Projet : 2 Data Scientists / 1 Développeur full-stack

Tâches effectuées :

- Implémentation de requêtes SQL d'import de données d'une base PostrgreSQL
- Prétraitement des données : regex, tokenisation, lemmatisation/stemming
- Parallélisation de traitements avec la librairie Python multiprocessing
- Modèles d'extraction d'entités (NER) et machine learning

Environnement technique	Python, Pandas, scikit-learn, nltk, Flair, Spacy, SQL, PostgreSQL, Gitlab
Modèles mathématiques	Xgboost, random forest, SVM, bi-LSTM

01/2015 – 04/2019 : Consultant Data Scientist chez Ernst & Young

Agence Régionale de Santé (ARS) IDF : Data Scientist

04/2019 – 04/2019 (1 mois)

Segmentation d'établissements de santé : analyse des indicateurs financiers des ARS de France en vue d'un plan de restructuration stratégique

Interlocuteurs : Stratégie et management

Equipe: 3 Personnes
Equipe Projet : 1 Data scientist / 1 Manager Data Scientist / 1 Consultant Stratégie

Tâches effectuées :

- Ateliers de collecte et compréhension des données
- Elaboration de modèles de segmentation
    - préprocessing des données : valeurs manquantes, création d'indicateurs, normalisation des données
    - Réduction de la dimension des données : ACP
    - Modélisation CAH (classification ascendante hiérarchique), K-means et arbre de décision explicatif
- Ateliers de boucles retours Métier pour améliorer le modèle

Environnement technique	Python, scikit-learn, Pandas
Modèles mathématiques	CAH, K-means, arbres de décision

EDF R&D : Data Scientist

04/2018 – 01/2019 (8 mois)

R&D : Appui aux projets Analitycs

Interlocuteurs : Marketing B2C

Equipe: 3 Personnes
Equipe Projet : 1 chef de projet / 1 Data Scientist / 1 Data Engineer

Tâches effectuées :

- Ateliers de compréhension de besoins
- Prévision de la consommation électrique des ménages français
    - Clustering des ménages

    - Régressions linéaires (Ridge, Lasso) en utilisant les informations superficie, nombre de pièces, présence de certains types d'appareil, etc.
- Détection de surconsommation ou sous-consommation des ménages en utilisant des modèles de régressions quantiles
- Rédaction de la documentation technique
- Intégration de nouvelles données dans un datawarehouse contenant les signatures d'appareils électriques
    - Collecte et nettoyage des open data et des données issues d'une expérimentation interne
    - Scripting Python de transformation et chargement de données
    - Scripting Bash d'industrialisation de la pipeline data
- Industrialisation d'un pipeline de traitement R dans un cluster Big data
    - Utilisation de l'ordonnanceur Slurm et scripting Bash avancé
    - Gain considérable en temps d'exécution : de 2 mois à 1 semaine

Environnement technique	Python, scikit-learn, Pandas, Bash, SQL, Slurm, Oracle, Gitlab
Modèles mathématiques	Régressions linéaires, régressions quatiles, clustering, arbres de décision

Ernst & Young – Département RH : Ingénieur de recherche

11/2017 – 03/2018 (4 mois)

Nom du projet : Application web de reporting RH

Interlocuteurs : RH / Contrôleurs de gestion / Chef de projet

Equipe : 2 Personnes
Equipe Projet : 1 Data scientist / 1 manager

Tâches effectuées :

- Ateliers de collecte et compréhension de données et recueil de nouveaux besoins
- Intégration des données dans le back-end en Python et création d'indicateurs
- Implémentation des visualisations front-end en utilisant le framework Flask
- Monitoring du serveur applicatif sous Docker

Environnement technique	Python, Pandas, Flask, Docker, Gitlab
Modèles mathématiques	N/A

Air France – Département Recherche Opérationnelle : Data Scientist

01/2017 – 09/2017 (9 mois)

Revenue Management Cargo : Amélioration d'un modèle d'optimisation linéaire sous contraintes

Interlocuteurs : Revenue management

Equipe : 4 Personnes
Equipe Projet : 2 Data scientist / 1 manager / 1 product owner

Tâches effectuées :

- Ateliers de collecte de nouveaux besoins
- Amélioration du modèle d'optimisation du revenu de l'activité cargo
    - Mise-à-jour sous Python d'un modèle de prévision de la capacité des vols en utilisant un arbre de décision se basant sur le type d'avion, les données calendaires et l'historique des capacités
    - Ajout de nouvelles contraintes sur la configuration spatiale dans le modèle d'optimisation sous contraintes écrit en C++
    - Minimisation du nombre d'escales subies par les marchandises
    - Restructuration et refactorisation du code
- Intégration des résultats sur une application web Java
- Mise en production du pipeline data : scripting Bash, ordonnanceur Crontab
- Formation des utilisateurs aux nouvelles évolutions

Environnement technique	C++, Ilog Cplex, Python, Bash, Java, Gitlab
Modèles mathématiques	Optimisation linéaire sous contraintes, arbres de décision

Enedis – DSI : Data Scientist

02/2015 – 07/2016 (1,5 ans)

QDD : Mise en qualité des données sur une chaine de quatre systèmes d'information gérant les clients professionnels

Interlocuteurs : Responsables DSI / Experts métiers

Equipe : 4 Personnes
Equipe Projet : 3 Data scientist / 1 Manager

Tâches effectuées :

- Ateliers de collecte de nouveaux besoins
- Mise en place de flux de données automatiques FTP provenant des 4 systèmes d'information
- Développement d'un programme en SAS de détection d'incohérences au sein des systèmes d'information allant de la relève des compteurs jusqu'à la facturation des clients
    - Gestion des imports et des jointures entre tables avec des 'proc SQL'
    - Implémentation des règles de gestion
- Réalisation d'un reporting Excel
- Mise en production du pipeline via l'outil Windows Task Scheduler
- Mise en place d'un processus pérenne de correction des écarts
    - Animation d'ateliers de présentation des incohérences détectées, de collecte de nouvelles règles de gestion et de collecte des nouveaux besoins
    - Pilotage des actions de correction

Environnement technique	SAS, Excel
Modèles mathématiques	N/A

Formation

2010 – 2014 : Diplôme d'Ingénierie d'Aide à la Décision
Ecole Internationale des Sciences du Traitement de l'Information - Cergy
2008 – 2010 : Classes préparatoires Maths - Physique
Lycée National Léon Mba – Gabon