**Sylvain PUYRAVAUD**

**Machine Learning Engineer**

**3 ans d’expérience**

**EXPERIENCES SIGNIFICATIVES**

* **Samsung** : NLP engineer (7 mois)
* **Softathome**: Data Scientist(7 mois)
* **Dassault Systèmes** : Research Engineer (2 ans)

**COMPÉTENCES**

**Compétences fonctionnelles**

* Data Science, Data Engineering
* Machine Learning & Deep Learning
* Computer Vision, NLP
* Etat de l’art
* Formation

**Compétences techniques**

* Langages : python(expert), R, Matlab, C++, Java
* Frameworks : tensorflow, cmusphinx,caffe, theano,keras, spark, kafka,cuda, nutch,scikit-learn
* Base de données : mongodb, cassandra, hadoop
* Méthodes : SCRUM, confluence,slack, github, jira
* Systèmes : Windows, Unix, Linux

**Langues**

* Français : langue maternelle
* Anglais : langue maternelle

**FORMATION**

* ENSAE : Ingénieur en statistique - 2016
* Ecole Normal Supérieur (ENS) : Master MVA (Mathématiques, Vision, Apprentissage.
* Prépa MP/MPSI – Lycée Henry IV
* Formation continue sur des moocs, des tutoriels disponibles sur internet et des projets personnels.

Exemple : développement d’algorithmes de reconnaissance d’images sur des données Kaggle, tests de l’algorithmes de séries temporelles sur des données d’abonnements.

**REFERENCES PROFESSIONNELLES**

**Samsung – Issy-les-Moulineaux**

**10/2018 à 6/2019 – 9 mois**

*Machine Learning Scientist et Engineer*

**Projet :** ASR (automatic speech recognition)

*L’équipe ASR s’occupe de la transcription de la voix en texte. Les performances de l’équipe ASR ont une influence directe sur la reconnaissance vocale de Bixby qui est installé sur le téléphone des particuliers. L’équipe est divisée en plusieurs sous modules : language modeling, ITN, wake up, G2P.*

**Contexte et objectif :** L’objectif était d’améliorer la reconnaissance vocale de Bixby sur les Samsung des utilisateurs.

**Environnement travail/fonctionnel :** Equipe de 10 personnes en SCRUM dont un lead.

**Contraintes :** Encadrement du projet par Samsung Pologne qui a mis en place l’architecture permettant de sous-traiter les langues Européennes. Contraintes de temps avec la sortie du produit en novembre 2018. Contrainte de performances pour l’acquisition de nouveaux clients. Contraintes légales pour le crawling de sites webs.

**Domaine d’intervention :**

* Crawling de sites web pour entrainer le modèle de langage. Utilisation de rotating proxies, beautiful soup, headless chrome, selenium(utiliation par exemple pour des pages à défilement infini sur le forum quora), VPN, apache nutch pour par exemple crawler le monde, wikipédia, français facile, Opus.... Utilisation d’expression régulières pour faire du cleaning. Utilisation de jsgf avec des grammaires pour la génération automatique de texte. Test de différents agencements de textes de language naturels pour améliorer la perplexité.
* Utilisation de lucenes pour agrandir les domaines corporas
* Mise en place de slot dictionaries
* Automatisation de tâches de correction automatique de corpus
* Utilisation de LSTM pour la phonétisation automatique
* Collaboré avec l’équipe NLU pour la détection d’intentions des utilisateurs.

Environnement technique :   
regular expression, nutch, python, cmusphinx

**SoftatHome – Colombes**

**3/2018 à 10/2018 – 7 mois**

*Data scientist*

**Projet :** Eyes’on

*Surveiller et optimiser l’utilisation du wifi des utilisateurs*

**Contexte et objectif :** Softathome est spécialisé dans les logiciels des box wifi. Le projet Eyes’on se focalise sur la récupération des données d’utilisation des box wifi(rsse, force du signal,…) pour les stocker dans une base de données. Ces données sont ensuite analysées pour apporter de la plus value.

**Environnement travail/fonctionnel :** Equipe de 10 personnes de data engineer, data scientists et data analysts en SCRUM.

**Contraintes :** Utilisation du cloud orange, 2 Alienwares

**Domaine d’intervention :**

* Analyse d’une très grande quantité de données (plusieurs tera) collectés des clients de Softathome(O2,Orange) afin d’aider les opérateurs à prendre des décisions plus rapidement.
* Mise en place d’un environnement spark pour faire du calcul parallèle. Un cluster S3 local a aussi été mis en place. Transfert des données de la base de données Cassandra vers S3.
* Combinaison avec des données externes (par exemple mozilla location services) pour localiser les box.
* Analyse statistique des données de séries temporelles pour ensuite les visualiser

Environnement technique :   
Cassandra, Spark, python, unix, git, jira, amazon S3, orange cloud, unix

**Dassault Systèmes– Vélizy-Villacoublay**

**11/2016 à 03/2018 – 1 ans 6 mois**

*Data Scientist*

**Projet : Systèmes de recommandations**

*Etat de l’art des systèmes de recommandations*

**Contexte et objectif :** La 3D Experience plateforme de Dassault Systèmes est une market place de produits 3D. L’objectif du projet était de faire un POC et un état de l’art dans le but futur d’intégrer un système de recommandation liant consommateur de produits 3D et les vendeurs de produits 3D.

**Environnement travail/fonctionnel :** Equipe recherche de 10 personnes spécialisé en machine Learning en scrum.

**Contraintes :** Peu de contraintes étant un projet de recherche.

**Domaine d’intervention :**

* Implémenté un algorithme à base de graphes combiné avec du tf-idf les données étant en faible quantité. Cet algorithme avait donné les meilleurs résultats au challenge 2016 de recommandation de Dassault Systèmes.
* Etat de l’art des systèmes de recommandation
* Implémenté 3 algorithmes de systèmes de recommandation : Ils étaient respectivement basés sur de la SVD, ALS, denoising auto-encoders.
* Modifié l’architecture du denoising autoencoder pour qu’il puisse prendre en compte les features utilisateurs et les features items. Ce nouvel algorithme donnait de meilleurs résultats que le simple denoising autoencoder qui donnait les meilleurs résultats jusqu’à présent.
* Implémenté un système d’interprétation permettant d’expliquer les recommandations du denoising auto encoder(LIME algorithm).
* Aidé à mettre en place un serveur démontrant un cas d’application du denoising autoencoder (Flask, visualisations avec D3.js).
* Transfert de la technologie de denoising auto encoder vers spark pour une possible intégration.
* Etat de l’art sur le reinforcement learning.
* Collaboré avec des membres de l’équipe sur la reconnaissance d’image : image segmentation(pour délimiter les portes et fenêtres sur un plan 3D), image to text(décrire une image avec du texte), reconnaissance de mouvements sur une video(pour détecter par exemple ce qu’un ouvrier est en train de faire dans une usine). Les technologies : GAN, res net, U-net, CNN, RNN

Environnement technique :   
Python, Réseaux de neurones, machine learning, Caffe, Theano, Cuda, UNIX, Flask, D3.js, Modèles bayésiens, spark, multiprocessing, Camtasia, machines virtuelles, mongoDB. Git.

**Dassault Systèmes– Vélizy-Villacoublay**

**5/2016 à 10/2018 - 6 mois**

*Stagiaire Data Scientist*

**Projet : Text classification**

*Classification de queries des utilisateurs sur la knowledge base*

**Contexte et objectif :** L’équipe knowledge qui faisait partie de l’équipe customer success qui faisait elle-même partie de l’équipe R&D support client.

**Environnement travail/fonctionnel :** Equipe knowledge de 5 personnes qui avait la responsabilité de l’amélioration continue de l’expérience d’utilisation de la knowledge base de Dassault Systèmes.

**Contraintes :** Faire un rapport de stage et un POC qui permettrait d’améliorer l’expérience utilisateur.

**Domaine d’intervention :**

* Analysé les logs de recherche des clients de Dassault sur la knowledge base dans le but de faire du clustering(pour détecter les différents sujets qui intéressent les utilisateurs) et ensuite de la détection d’anomalie pour mettre en place un système de feedback automatisé au support client.
* Cleaning de texte avec des outils de nlp(lemming,stemming, language detection, nltk, stopwords,…).
* Mise en place d’algorithmes de clustering (K-means, algorithmes basés sur la théorie de graphes,…) sur des features extraits grâce aux méthodes de nlp(tf-idf, word embeddings, ngrams,…) afin de classifier les requêtes utilisateurs.
* Aide à la mise en place d’une démo pour la présenter à l’EXCOM.
* Collaboré avec un membre de l’équipe recherche pour la classification d’images de la Knowledge base pour améliorer le moteur de recherche (CNN classique).

Environnement technique :   
Python, nltk, R, theano

**Openclassrooms – en remote**

**03/2017 à présent – 7 mois**

*Tutorat d’élèves data scientiste*

**Projet :** Mentorat

*Encadrement des élèves du parcours data scientiste, data analyste et data engineer.*

**Contexte et objectif :** L’objectif est le suivi de l’élève jusqu’à l’obtention du diplôme accordé par OpenClassrooms.

**Environnement travail/fonctionnel :** mentorat par Visio conférence à distance

**Contraintes :** l’élève doit valider tous les projets pour valider le parcours. Réunions d’une heure hebdomadaire.

**Domaine d’intervention :**

* Encadré une dizaine d’élèves sur le parcours data scientiste dont trois diplômés qui ont intégré l’industrie.
* Encadrement de 3 élèves sur le parcours data engineer dont 1 diplômé
* Encadrement de 3 élèves sur le parcours data analyste dont 1 élève presque diplômé.