

Julien PLU

## Ingénieur recherche et développement

10 ans d'expérience dans divers projets liés à l'intelligence artificielle sur le Traitement Automatique du Langage Naturelle, en particulier sur le développement de modèles et leur déploiement en production. Autonome sur mon poste tout en appréciant le travail en équipe avec des collaborateurs sur site ou distants.

### Expériences professionnelles

#### Ingénieur de Recherche et Développement en IA

**Buster.Ai** Paris / Depuis octobre 2022

Mon rôle chez Buster est de faire passer l'IA de la recherche à la production de bout en bout. Cela implique autant de mener des recherches que de rendre ces recherches disponibles en production pour améliorer notre produit. Je travaille principalement avec des LLM pour effectuer des recherches et des vérifications de faits, ainsi que d'optimiser ces LLM afin d'avoir le meilleur rapport taille faible/faible latence dans un environnement de production contraint. La pile technique complète est en Python et le fournisseur de cloud est GCP.

#### Ingénieur Machine Learning

**Hugging Face** Paris / De septembre 2020 à mars 2021

Mon rôle était d'intégrer TensorFlow dans les frameworks Transformers et Datasets d'Hugging Face en implémentant de plus en plus de modèles/outils. J'ai également aidé la communauté NLP à mieux prendre en charge ces modèles et fonctionnalités Tensorflow.

#### Ingénieur Machine Learning Senior

**Beam App** Paris / De mai 2021 à octobre 2022

Mon rôle chez Beam était de travailler sur deux projets différents :

1) Le premier projet consistait à générer des paragraphes lorsqu'un utilisateur prenait des notes dans notre navigateur Web, cette génération prenait en compte tout le contexte des notes, y compris les vidéos qu'elles contenaient, aucun LLM n'était utilisé pour faire cela uniquement du texte simple modèles de génération.

2) Le deuxième projet consistait à regrouper les onglets ouverts dans le navigateur Web par thèmes et à leur donner un nom.

Les modèles des deux projets devaient être exécutés sur site directement sur la machine de l'utilisateur et aussi rapidement que possible. Pour ce faire, de nombreuses optimisations de modèles ont été utilisées et l'intégration a été réalisée en créant des bibliothèques C++ utilisées directement dans Swift, car le navigateur Web était uniquement destiné à MacOS/iOS.

#### Ingénieur Recherche et Développement en IA

**Leboncoin** Paris / D'avril 2018 à août 2020

Mon rôle chez Leboncoin était d'étudier de nouvelles approches d'IA afin d'améliorer l'expérience de nos utilisateurs.

J'ai travaillé sur plusieurs projets mais le plus gros projet était de former et de déployer un modèle linguistique pour le français afin de classer les titres d'annonces. Cela a apporté de nombreux défis tels que :

- pour conserver un énorme ensemble de données (~ 100 millions d'exemples), beaucoup de nettoyage a été nécessaire.
- peu de GPU étaient disponibles, j'ai donc dû faire face à une très petite possibilité de calcul pour entraîner un modèle de langage.
- J'ai dû implémenter et intégrer ce modèle dans la stack Leboncoin en Go.
- l'infra n'était pas faite pour accueillir un tel nouveau service et ensuite, avec l'équipe SRE nous avons dû définir un pipeline complet uniquement pour les services ML.



✉ plu.julien@gmail.com

☎ 06 70 14 19 68

📅 Né le 19/06/1987

🏠 Bagnolet (93170)

🔗 <https://jplu.github.io>

🇫🇷 Français

💻 Télétravail ou présentiel

📄 Permis B

🚲 Mobile en vélo

📍 Île-de-France

👥 Fiancé

### Langues

Anglais

Japonais

### Compétences

Apprentissage Automatique

Traitement Automatique du  
Langage Naturel

Web sémantique

### Informatique

Python

Java

Go

### Divers

Kubernetes

Google Cloud Platform

### Centres d'intérêt

Jeux Vidéo

Plongée

Guitare

Football

## Diplômes et Formations

- **Doctorat en Informatique**

Eurecom Sophia Antipolis / De juillet 2014 à décembre 2017

Mon sujet de thèse consistait à extraire des unités lexicales (entités) de documents textuels afin de les dissocier des ressources Web contenues dans une base de connaissances et de manière adaptative. Cette adaptivité la langue des textes traités, la base de connaissances utilisée et le type de contenu textuel.

- **Master Informatique, spécialité Intelligence Artificielle**


Université Montpellier 2 Montpellier / De septembre 2010 à septembre 2012

- **Licence Informatique**

Université Montpellier 2 Montpellier / De septembre 2007 à juin 2010

## Réseaux sociaux

 @julienplu

 @jplu

 @julienplu