Mener un projet d'IA contributif : enseignements et méthodologies

Comment le projet PIAF explore les enjeux de la contribution et des communautés

Le projet PIAF "Pour des intelligences artificielles francophones" a pour objectif de **construire le premier jeu de données ouvert de questions-réponses en français**. La cible : permettre à des services qui utilisent l'IA (et plus spécifiquement le traitement automatique du langage) d'améliorer leur performance grâce à des données d'entraînement en français et de qualité.

Lancé en juin 2019, ce projet a été mené par une petite équipe au sein d'Etalab, plus précisément du Lab IA, dont la mission est de diffuser des méthodes et des ressources pour l'IA publique. Pendant 6 mois, l'équipe a défini sa méthodologie et un axe a été rapidement privilégié : faire de PIAF un projet **contributif**, qui permette autant de **diversifier les participants au projet (ie : ne pas uniquement en faire un projet technique)** que de **constituer une communauté autour des suiets d'IA et d'open data**.

Qu'avons-nous appris de cette expérience ? Quelles sont les principales difficultés auxquelles nous avons été confrontés ? Quelles ont été nos surprises ? Nous vous proposons, dans cet article, un retour d'expérience qui pourra être utile à d'autres équipes.

Un projet technique, une approche moins technique

Lorsque nous avons décidé de nous aventurer dans l'univers des jeux de données francophones, le sujet nous paraissait assez complexe techniquement : au sein de l'administration, de plus en plus de projets d'IA sont menés (grâce notamment au programme Entrepreneurs d'intérêt général et aux appels à manifestation d'intérêt IA), mais le vocabulaire de l'IA n'est pas encore pleinement installé. Les termes comme *traitement automatique du langage, données d'entraînement, plateforme d'annotation, "leader-board", " speech to text"*, etc. ne se retrouvent pas tous les 2 jours dans le journal officiel ! Parfois même, on emploie des termes de l'IA avec quelques contre-sens et une maîtrise limitée.

Pour palier le risque d'un projet trop technique et uniquement tourné vers la communauté data science, nous avons pris le parti de faire de PIAF un "prétexte" pour parler de l'IA autrement au sein de l'administration.

communauté scientifique nous a permis de conforter nos hypothèses et d'imaginer des collaborations futures.

Les volets participatif et pédagogique : annotathons, tournées

- **Une liste de diffusion** : lors du lancement du projet, nous avons ouvert une liste de diffusion, qui s'est enrichie au fil des semaines, des événements et des rencontres.

- **Les "annotathons"** : chaque semaine, nous réunissons tout contributeur volontaire et/ou intéressé par le projet au Lieu de la transformation publique (merci la DITP). **En 4 mois (octobre-janvier), nous avons organisé 12 annotathons et réuni plus de 150 personnes**. Lors de ces moments contributifs, animés par Benjamin, notre chargé de déploiement, les contributeurs "s'embarquent" dans le projet, en créant un compte et en réalisant une première annotation. Des échanges ont lieu sur le projet, l'expérience utilisateur et de nouvelles annotations sont réalisées.

Ces événements nous ont permis, au fil de l'eau : - d'apporter des améliorations UX sur la plateforme d'annotation - d'affiner notre discours sur le projet - d'identifier des cas d'usages et des partenariats avec des administrations

- **Les tournées** : grâce aux annotathons et à nos échanges avec les administrations sur l'IA, nous avons été sollicités par plus de 5 ministères pour présenter PIAF à différents publics. Ces événements étaient l'occasion de faire parler du projet tout en proposant des formats pédagogiques sur l'IA.

En complément, nous avions en tête qu'un travail éditorial était nécessaire pour faire vivre le projet et le documenter. Nous avons donc fourni un effort supplémentaire pour disposer d'un site attractif et écrire du contenu au fil de l'eau. Et bien sûr, tout au long du projet, nous nous attachons à tout ouvrir (la méthode comme le code). D'autres actions sont en réflexion : créer une identité de PIAF sur les réseaux sociaux, ouvrir un forum de discussion en ligne, etc.

Les frontières de la contribution volontaire

Vous l'aurez compris, PIAF c'est d'abord générer des questions-réponses en français de manière contributive. Pour cela, nous avons mis en place une plateforme d'annotation qui permette à tout contributeur de proposer des questions et des réponses sur un texte.

Pendant les 4 premiers mois d'annotation (octobre 2019-janvier 2020), notre cible était d'atteindre 20 000 questions-réponses pour constituer un **premier jeu de données d'évaluation**, afin *d'évaluer* la performance de modèles sur un jeu de données nativement en français. Les annotathons ont été mis en place pour atteindre cet objectif et répondre à une exigence de qualité : ces premières données devaient être produites par des "annotateurs certifiés", c'est-à-dire formés par l'équipe.

Vous l'aurez compris aussi, nous avions plusieurs options pour générer ces questions-réponses :