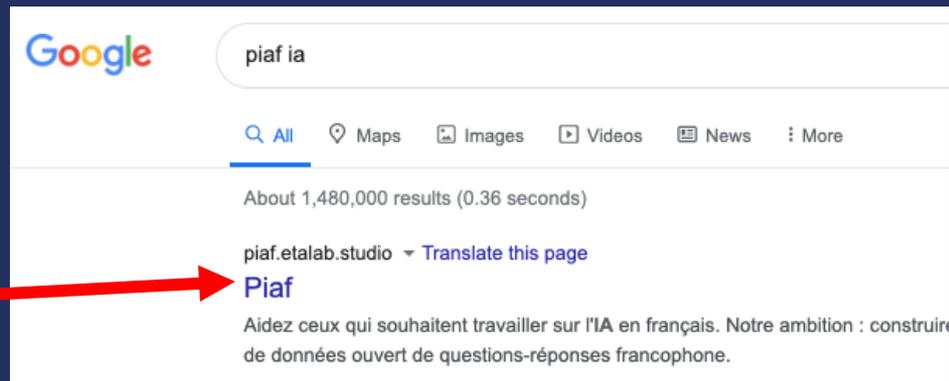


Préliminaire : créez-vous un compte !



Ou directement : <https://app.piaf.etalab.studio/signup/>

Pour des IA francophones

Présentation Le Wagon – 19 août 2020

piaf@data.gouv.fr



Piaf

etalab gouv.fr



le wagon

Programme

Présentation générale du projet et de ses enjeux.

Découverte de la plateforme.

Présentation de la démarche scientifique.

**Pour des IA francophones :
apprendre aux robots à
parler français**

2018, le rapport Villani

D'abord, une politique offensive visant à favoriser l'accès aux données, la circulation de celles-ci et leur partage. Les données sont la matière première de l'IA contemporaine et d'elles dépend l'émergence de nombreux usages et applications. Il est tout d'abord urgent d'accélérer et d'étoffer la politique d'ouverture des données publiques (*open data*), en particulier s'agissant des données critiques pour les applications en IA. La démarche d'*open data* fait l'objet d'une politique volontariste depuis plusieurs années, notamment sous l'impulsion de la loi pour une République numérique³ : cet effort, important, doit être soutenu. La puissance publique doit par ailleurs amorcer de nouveaux modes de production, de collaboration et de gouvernance sur les données, par la constitution de « *communs de la donnée* »⁴. Il lui revient ainsi d'inciter les acteurs économiques au partage et à la mutualisation de données voire, dans certains cas, d'en imposer l'ouverture. La politique de la donnée doit enfin s'articuler avec un objectif de souveraineté et capitaliser sur les standards de protection européens pour faire de la France et l'Europe les championnes d'une IA éthique et soutenable. L'Union européenne s'est engagée depuis

CÉDRIC VILLANI

Mathématicien et député de l'Essonne

DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

POUR UNE STRATÉGIE
NATIONALE ET EUROPÉENNE

IA et politique publique

Un rôle d'**impulsion générale** : financement de la recherche, *open data*.

Un rôle d'**animateur** : consolider et rendre visible l'écosystème français.

Un rôle de **contrôle** : protection des données, explicabilité des algorithmes.

Un rôle d'**exemple** : intégrer les bénéfices de l'IA dans les services publics.

PIAF : le produit et ses enjeux

Un jeu de données ouvert de questions-réponses francophone : mettre des données pour des IA francophones à disposition des administrations, des laboratoires de recherche, des entreprises, des citoyens.

Open data



Place du français
dans l'IA



Micro-travail ?
Contribution engagée ?



Question scientifique :
Natif VS Multilingue



Un projet ouvert, documenté, qui fait le choix de la « contribution engagée »

Une plateforme d'annotation *open source*.

Un premier jeu de données *open data* de questions-réponses construit selon une méthodologie scientifique.

Une méthode ouverte : contributions volontaires et communauté.

Un enjeu de pédagogie : démythifier l'IA et montrer qu'elle est toujours le résultat de choix humains.

Découvrez comment ça fonctionne en 3 min



PIAF : la démarche scientifique

Les modèles de questions-réponses

- Sont entraînés pour trouver la « bonne réponse » à une question dans un texte qui contient la réponse.

Ex. Dans une biographie de Louis XIV, trouver la réponse « 1638 » à la question « Quelle est la date de naissance de Louis XIV? ».

- Une technologie transformante pour les tâches de recherche ou d'extraction
=> plus besoin de structurer les données texte avant des les interroger =>
une nouvelle génération d'IA.

Les bases de données sont déterminantes dans la constitution de l'IA: elle apprend en voyant de nombreux exemples.

Garbage in => Garbage out

Le problème

- Les *datasets* d'entraînement / évaluation existent exclusivement en anglais / chinois (SQuAD, QuAC, HotpotQA, NewsQA, etc.).
- Peu ou pas de données dans les autres langues. Pas de *dataset* significatif en français.
- La traduction automatique des *datasets* ne suffit pas (env. -10 points de performance selon nos évaluations => 4 ans de retard). Cf Annexe 1



Un protocole inspiré du *dataset* de référence proposé par Stanford

The screenshot shows the SQuAD 2.0 website interface. At the top, there is a navigation bar with 'SQuAD' on the left and 'Home', 'Explore 2.0', and 'Explore 1.1' on the right. The main header features the text 'SQuAD2.0' in a large, white, serif font, with 'The Stanford Question Answering Dataset' in a smaller, white, sans-serif font below it. The page is divided into two main columns. The left column is titled 'What is SQuAD?' and contains a paragraph describing the dataset and a 'New' section about SQuAD2.0. Below the text are two blue buttons: 'Explore SQuAD2.0 and model predictions' and 'SQuAD2.0 paper (Rajpurkar & Jia et al. '18)'. The right column is titled 'Leaderboard' and contains a paragraph explaining the task and a table of model performance. The table has four columns: Rank, Model, EM, and F1. It lists several models, including Human Performance, ALBERT (ensemble model), XLNet + DAAF + Verifier (ensemble), ALBERT (single model), UPM (ensemble), XLNet + SG-Net Verifier (ensemble), and XLNet + SG-Net Verifier++ (single model). Each model entry includes a date in a small box and the team name.

What is SQuAD?

Stanford **Question Answering Dataset (SQuAD)** is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

New **SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) Google Language ALBERT Team	89.731	92.215
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) Google Language ALBERT Team	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) Anonymous	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	88.174	90.702
4	XLNet + SG-Net Verifier++ (single model)	87.238	90.071

<https://rajpurkar.github.io/SQuAD-explorer/>

Le protocole : annoter des articles Wikipédia en français et disposer de données de qualité et comparables à SQuAD.

Nous avons construit un protocole pour rendre les évaluations comparables avec SQuAD :

- Sélection d'articles similaires en « complexité ».
- Protocole d'annotation différent (pas de *Mechanical Turk*) mais comparable en *output*.

Participer en 3 min



Les données PIAF permettront

Pour la recherche :

Mesurer les écarts de performance entre modèles multilingues et natifs.

Pour les *data scientists* :

Entraîner des modèles natifs francophones.

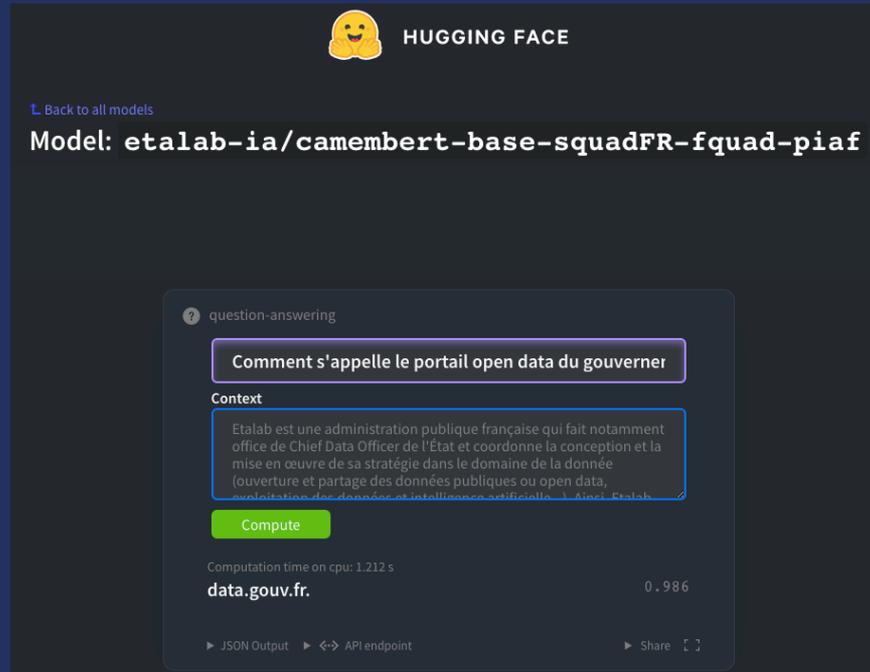
Pour tous :

Le *dataset* est disponible (*open data*) sur la plateforme **data.gouv**.



The screenshot shows the data.gouv.fr website interface. At the top, there is a search bar with the text "plaf" entered. Below the search bar, there are navigation tabs for "Données", "Réutilisations", "Organisations", "Tableau de bord", and "Documentation". To the right of these tabs, there is a language selector (France flag) and a "Connexion / Inscription" link. Below the search bar, there are two buttons: "1 jeu de données" and "1 réutilisation". To the right of these buttons, there is a "Trier par" dropdown menu set to "PERTINENCE" and a download icon. The main content area displays the dataset "PIAF - Le dataset francophone de Questions-Réponses (PIAF - Q&A)" with a description: "PIAF, construire un jeu de données francophones ouvert pour l'IA. Le recours à l'intelligence artificielle au sein de l'action publique est souvent identifié comme une opportunité pour interroger des textes documentaires et réaliser des outils de". There is also a "Mots-clés" dropdown menu with "chatbot" and "intelligence-artificielle" listed.

Modèle entraîné Piaf



The screenshot shows the Hugging Face interface for the model `etalab-ia/camembert-base-squadFR-fquad-piaf`. The question being asked is "Comment s'appelle le portail open data du gouverner". The context provided is: "Etalab est une administration publique française qui fait notamment office de Chief Data Officer de l'État et coordonne la conception et la mise en œuvre de sa stratégie dans le domaine de la donnée (ouverture et partage des données publiques ou open data, exploitation des données et intelligence artificielle...). Ainsi, Etalab". The model's response is "data.gouv.fr.". The interface also shows a "Compute" button, the computation time on CPU (1.212 s), and options to view the JSON output, API endpoint, and share the result.

HUGGING FACE

[Back to all models](#)

Model: `etalab-ia/camembert-base-squadFR-fquad-piaf`

question-answering

Comment s'appelle le portail open data du gouverner

Context

Etalab est une administration publique française qui fait notamment office de Chief Data Officer de l'État et coordonne la conception et la mise en œuvre de sa stratégie dans le domaine de la donnée (ouverture et partage des données publiques ou open data, exploitation des données et intelligence artificielle...). Ainsi, Etalab

Compute

Computation time on cpu: 1.212 s

data.gouv.fr. 0.986

▶ JSON Output ▶ ↔ API endpoint ▶ Share

<https://huggingface.co/etalab-ia/camembert-base-squadFR-fquad-piaf>

Exemple d'utilisation du modèle & limites



Piaf & Service-public.fr

Une autre façon de trouver une information

[Chercher](#)

**Combien de paragraphes
annotés en 5 minutes ?**

<https://app.piaf.etalab.studio>

**Nous répondons à vos
questions !**

Merci pour votre écoute !

Présentation Le Wagon – 11 août 2020

piaf@data.gouv.fr



Piaf

etalab gouv.fr



le wagon

Annexe 1

