

# Spinoza Project

Support de réplication

---

**RSF** REPORTERS  
SANS FRONTIÈRES

**Alliance**  
de la presse  
d'information  
générale

**Ekimetrics**

# Objectifs du projet

# Beaucoup de cas d'usages pour la presse existent

*Spinoza accélère les recherches sur un sujet mais ne travaille pas à la place du journaliste*

**Identifier des enjeux &  
stimuler à la réflexion**

**Spinoza**

**L'IA pour interpréter les  
enjeux à partir de sources  
de confiance**

**L'IA pour imaginer des  
enjeux**

**Accélérer la  
compréhension**

**L'IA pour analyser un  
sujet à partir de sources  
de confiance**

**L'IA pour questionner des  
documents spécifiques  
additionnels**

**Accélérer la rédaction & la  
synthèse**

**L'IA pour aider à la Post  
production**

Style

Relecture

SEO, Texte  
alternatifs

Rédaction

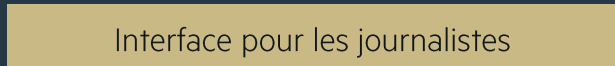
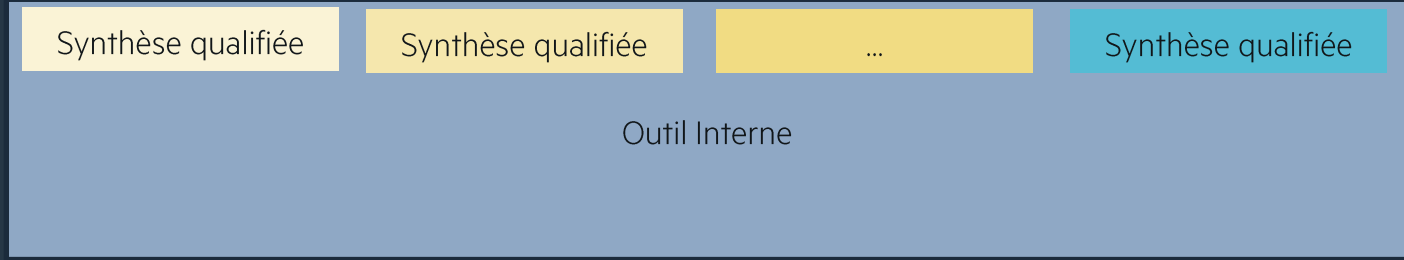
# **La solution open source VS votre solution**

# Une solution MultiRAG dans laquelle vous pourrez intégrer vos données en répliquant l'outil chez vous

Une première étape pour intégrer les Informations sur la structure spécifique de la donnée



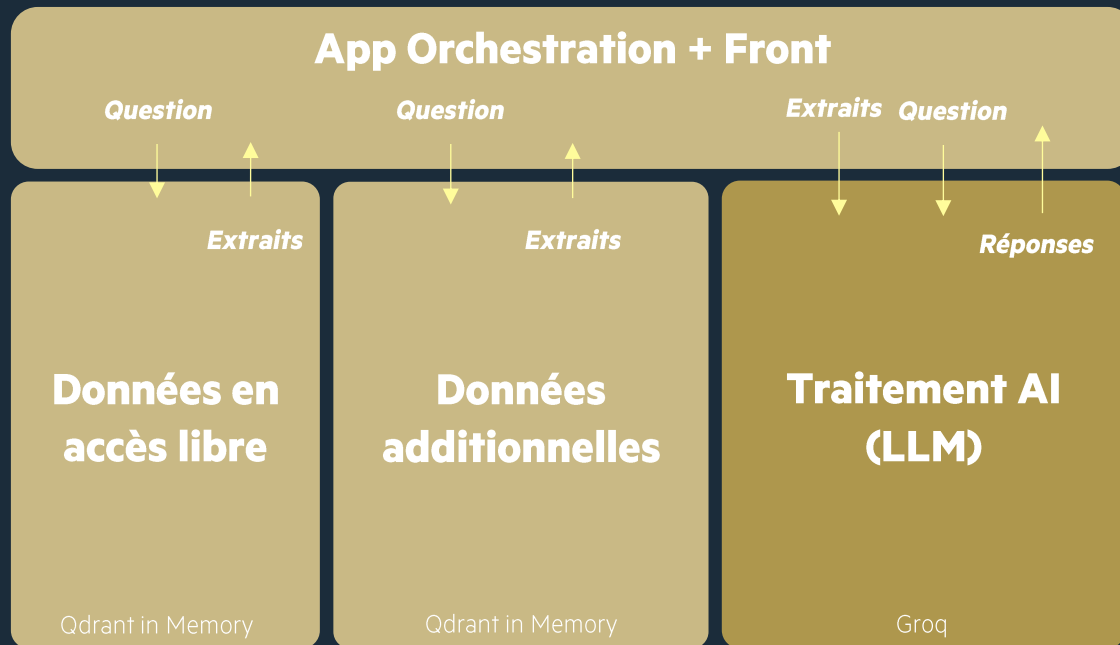
Une seconde pour valoriser la donnée pour un journaliste



# La solution disponible en ligne repose sur des briques gratuites dont nous ne garantissons pas la souveraineté

Hugging Face public

Provider gratuit



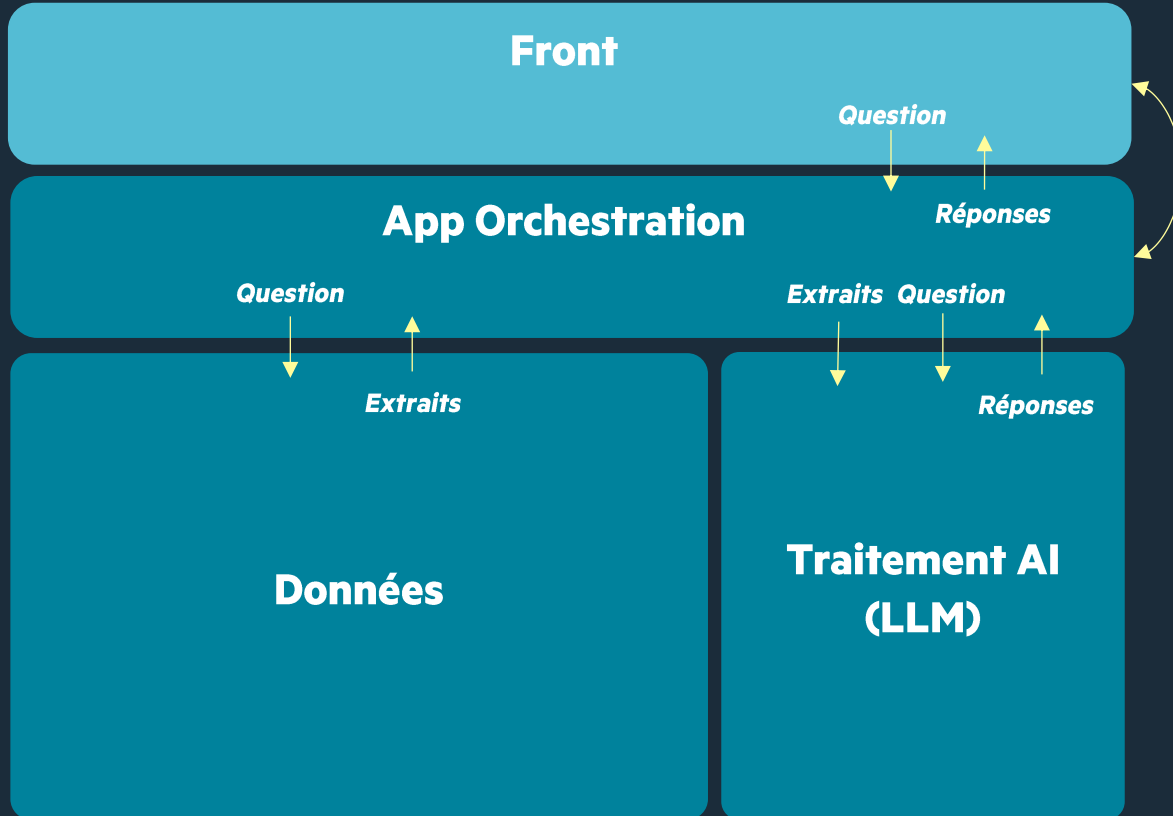
## Points d'attention :

- Les questions, extraits et réponses affichées dans l'app pourront être réutilisés par le provider de service gratuit
- Il faut donc éviter d'utiliser des données sensibles, et de poser des questions à enjeu.
- Il conviendra ensuite de déployer la solution en interne avec des briques sécurisées

# Afin d'internaliser cette solution et travailler sur un environnement sécurisé voici ce que nous vous recommandons

URL publique ou interne

Infrastructure interne



*Chiffrement RSA si la solution n'est pas si un réseau interne*

## Points à mettre en place :

- Hébergement interne de l'ensemble des briques
- Chiffrement des données internes, acces review, MFA
- Chiffrement RSA (si solution sur internet)
- Système de logs pour analyser les flux de potentiel extraction

# Quelles licences avec quels droits ?

# License/ Droit d'auteur

*Il y a deux éléments liés à des licences différentes :*

**L'application** (le code) est sous une [licence](#) GNU General Public License v3.0

- Vous pouvez l'utiliser, la modifier et la commercialiser comme bon vous semble
- Vous êtes obligés de conserver la mention de ses créateurs sur les interfaces qui pourront être produites.

## **Les données :**

- Vous pouvez les utiliser gratuitement pour vos besoins internes, projets académiques, éducatifs ou personnels
- Vous pouvez les intégrer dans des applications non commerciales
- Vous pouvez les utiliser au sein de votre entreprise pour vos propres cas d'usage
- Vous êtes obligés d'utiliser une technologie RAG qui cite intégralement les sources
- Vous n'êtes pas autorisés à commercialiser des services ou produits qui utilisent directement ces données

# Fonctionnement Technique

# Comment ça fonctionne derrière ?

1 Utilisateur pose une question

La question est transformée en représentation sémantique

Et recherchée dans la **base de données vectorielles**

2 Eléments les plus pertinents

Utilisés pour créer un prompt

3 GPT répond avec une synthèse

Quelles sont les causes de la montée des eaux ?



[0.6442, 0.1337, [...],  
0.0279, 0.8995]



Page 7 - IPCC ARS  
"B11 Evidence [...]"

Page 7 - IPCC SR  
"Panel (b) Evidence [...]"



Voici la question utilisateur : [...]  
Et les documents que tu dois utiliser.

Les causes de la montée des  
eaux sont les suivantes : [...]

The screenshot shows a Q&A interface with a search bar containing the question "Quelles sont les causes de la montée des eaux ?". Below the search bar, there are several agent cards: "Agent Science - ready", "Agent Loi - ready", "Agent Organismes publics - ready", "Agent ADEME - ready", and "Spinoza - ready". The "Agent Science" card is selected and displays the question "What are the main causes of sea level rise?". Below the question, it lists the main causes of sea level rise: thermal expansion, loss of land ice, anthropogenic greenhouse gas emissions, anthropogenic subsidence, and redistribution of water. To the right of the list, there is a "Sources" section with two documents: "Doc 1 - The Ocean and Cryosphere in a Changing Climate - Page 341" and "Doc 2 - The Ocean and Cryosphere in a Changing Climate - Page 351". The relevance score for the sources is 87.7%.

# 1 – Réplication de l'espace

## **La philosophie :**

Répliquer votre propre interface en utilisant vos données à l'aide d'outils gratuits (nous vous recommandons donc de ne pas utiliser de données sensible ou confidentielles).

## **Les étapes :**

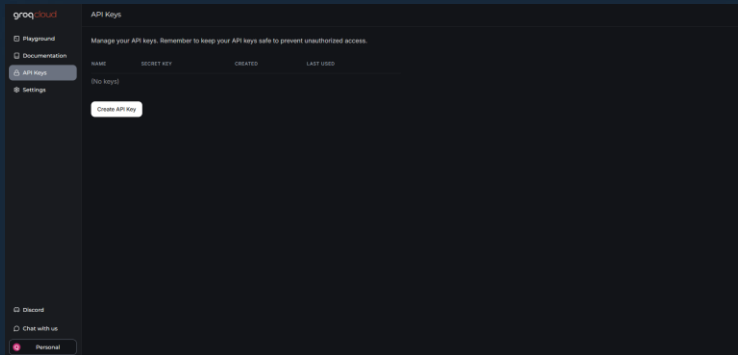
1. Créer un compte Groq (différent de Grok détenu par Elon Musk)
2. Enregistrer le token de développeur (il permettra d'accéder à des modèles gratuitement)
3. Créer un compte Hugging Face (si le compte est déjà existant sauter cette étape)
4. Créer un token d'accès avec les droits de lecture (il permettra à votre application de lire les bases de données que vous enregistrerez)
5. Dupliquez sur votre espace personnel l'application Spinoza (avec uniquement les données publiques).

# Créer un compte et enregistrer le token

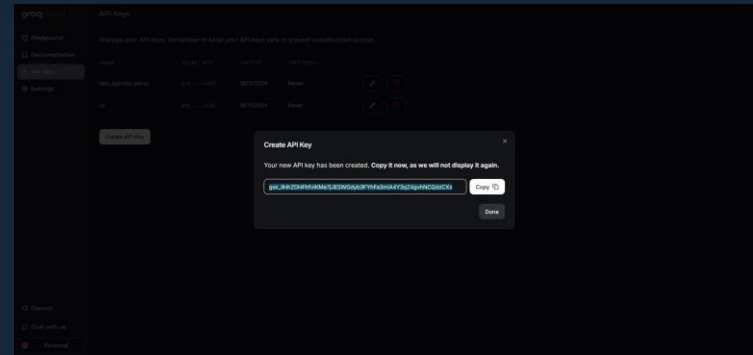
Se rendre sur la page et créer un compte : <https://console.groq.com/login>

=> Pour la première connexion il suffit de renseigner son email et un email de confirmation sera envoyé (pensez à vérifier les spams)

Une fois authentifiés, allez sur API Keys sur l'onglet de gauche



Cliquez sur créer une clef API (qu'importe le nom eg spinoza\_perso) et enregistrez là dans un fichier texte



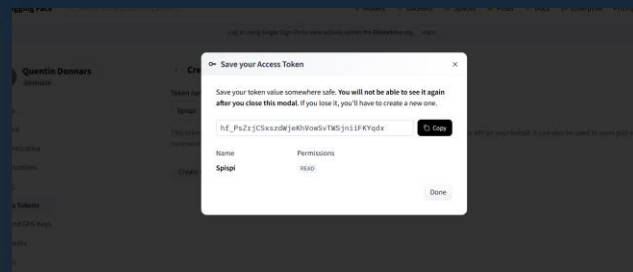
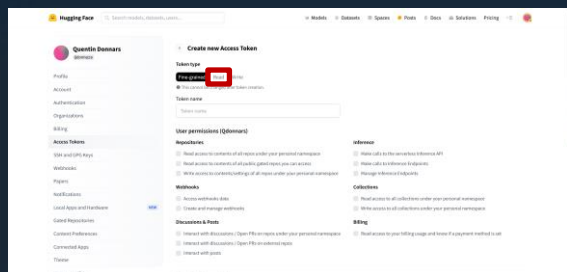
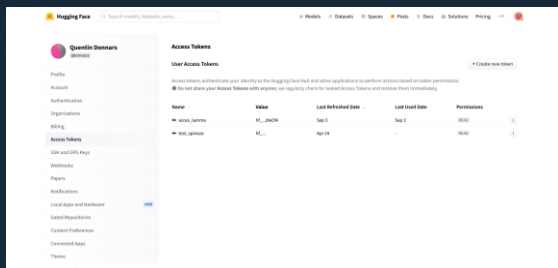
# Créez un compte hugging face et récupérez votre token d'accès :

Si vous avez déjà un compte hugging face sautez ces étapes :

1. Se rendre sur la page et créer un compte : <https://huggingface.co/join>
2. Faire une demande d'accès à l'application Spinoza : <https://huggingface.co/login?next=%2FSpinozaProject>
3. Attendre que la demande soit acceptée par un Admin (Vincent ou un des encadrants du Hackathon)

Pour tout le monde : création d'un token développeur

1. Se rendre dans la console de gestion des tokens : <https://huggingface.co/settings/tokens>
2. Créez un nouveau Token
  1. (Create new token en haut à droite)
  2. Donnez les droits d'accès en lecture (read)
  3. Nommez-le comme vous le souhaitez (eg spinoza\_perso)
  4. Cliquez sur créer
  5. Enregistrez le dans un document



# Dupliquez votre répertoire Spinoza

Rendez vous sur la page du projet : <https://huggingface.co/spaces/SpinozaProject/spinoza>

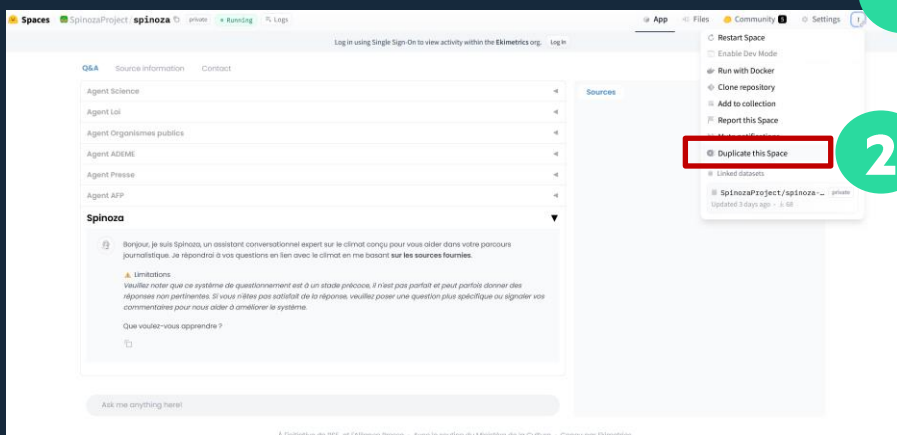
Dupliquez le projet

1

Cliquez sur les 3 petits points

2

Puis sur « duplicate this space »



# Remplissez les tokens nécessaires :

The screenshot shows the 'Duplicate this Space' dialog box. The 'Owner' field is set to 'Qdonnars' and the 'Space name' is 'spinoza\_public'. The 'Visibility' dropdown is set to 'Private'. A red circle with the number '1' is placed over the 'Space name' field, and another red circle with the number '2' is placed over the 'Visibility' dropdown.

1

Renseignez le nom souhaité du projet : `spinoza_perso`

2

Passez le répertoire en public (nécessaire pour certains services gratuits d'ingestions que nous allons utiliser)

3

Renseignez vos 2 tokens d'accès (HF = Hugging face, GROQ\_API\_KEY = token GROQ)

4

Cliquez sur « Duplicate this space »

The screenshot shows the 'Duplicate this Space' dialog box with the 'GROQ\_API\_KEY' and 'HF\_TOKEN' fields highlighted by a red rectangle. The 'Duplicate Space' button is also highlighted by a red rectangle. A red circle with the number '3' is placed over the 'GROQ\_API\_KEY' field, and another red circle with the number '4' is placed over the 'Duplicate Space' button.

3

4

# Vous avez désormais votre première interface Spinoza

Le projet porte est **hébergé de manière privée** sur votre espace personnel, et a le nom que vous lui avez donné

Seules les données  
acceptées seront  
disponibles

The screenshot displays the Spinoza interface within a user's personal space. At the top, the navigation bar includes 'Space', 'Qdonnars', and 'spinoza\_test' (highlighted with a red box), along with 'private', 'Running', and 'Logs' indicators. On the right, there are links for 'App', 'Files', 'Community', and 'Settings'. A 'Log in' button is also present. Below the navigation, a message prompts the user to log in using Single Sign-On to view activity within the Ekimetrics.org. The main content area is divided into two sections: 'Q&A' and 'Sources'. The 'Q&A' section is currently active, showing a list of sources: 'Agent Science', 'Agent Loi', 'Agent Organismes publics', and 'Agent ADEME', each with a right-pointing arrow. Below this list, the 'Spinoza' chat window is visible, featuring a greeting: 'Bonjour, je suis Spinoza, un assistant conversationnel expert sur le climat conçu pour vous aider dans votre parcours journalistique. Je répondrai à vos questions en lien avec le climat en me basant sur les sources fournies.' It also includes a 'Limitations' section with a warning icon and text: 'Veillez noter que ce système de questionnement est à un stade précoce, il n'est pas parfait et peut parfois donner des réponses non pertinentes. Si vous n'êtes pas satisfait de la réponse, veuillez poser une question plus spécifique ou signaler vos commentaires pour nous aider à améliorer le système.' At the bottom of the chat window, there is a prompt 'Que voulez-vous apprendre ?' and a small icon. A text input field at the bottom of the page contains the placeholder text 'Ask me anything here!'. The footer of the page reads: 'À l'initiative de RSF et l'Alliance Presse - Avec le soutien du Ministère de la Culture - Conçu par Ekimetrics'.

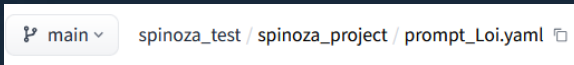
# **2.a – Modification des prompts (optionnel)**

# Aller dans le fichier de prompt que vous souhaitez modifier :

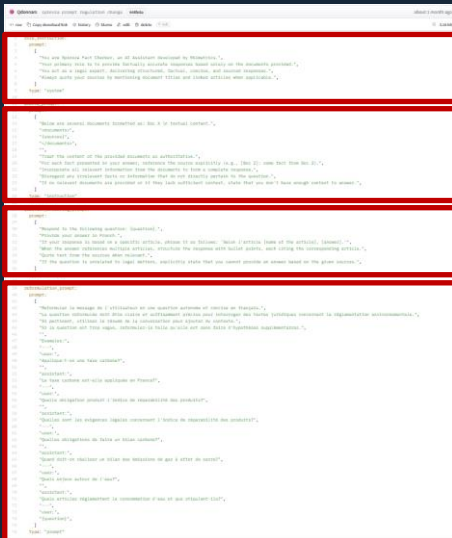
1- Accédez aux fichiers de l'application



2- Les prompts se trouvent dans le dossier spinoza\_project



3- Vous arriverez sur un éditeur de texte que vous pourrez modifier :



Ici sont les instructions générales, elles permettent de rappeler dans quel contexte l'algorithmes va s'exprimer

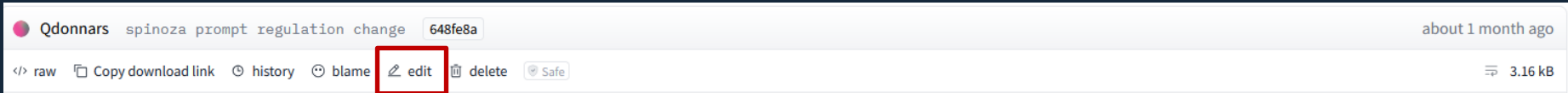
Ici sont les instructions pour traiter l'information, elles permettent notamment d'expliquer à l'algorithmes d'utiliser les sources que nous lui faisons parvenir et de les citer dans ses réponses

Ici sont les instructions pour répondre à la question reformulée

Ici sont les consignes de reformulation. Nous lui donnons ici quelques consignes et beaucoup d'exemples.

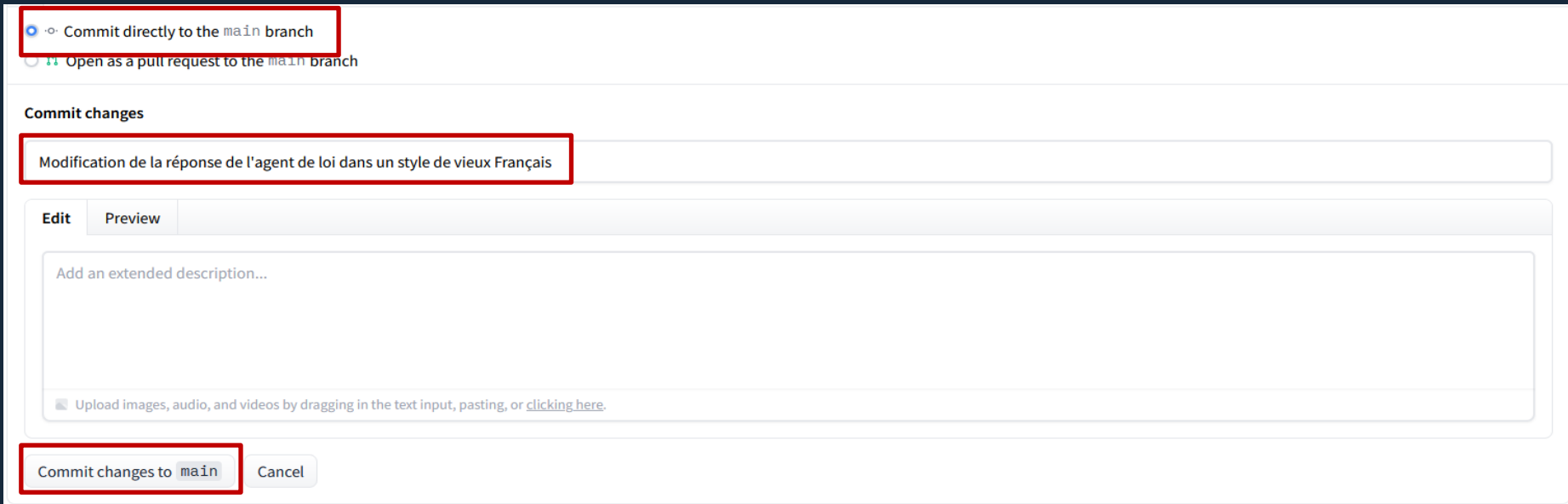
# Faites vos premières modifications et publiez-les :

1- Cliquez sur éditer



2- Faites vos premières modification (eg : "Provide your answer in old French.")

3- Déployez vos réponses, et expliquez brièvement vos modifications dans le champ texte dédié



4- Attendez que l'application redémarre et testez vos modifications

## **2.b – Changer le modèle (optionnel)**

# Quels modèles disponibles ?

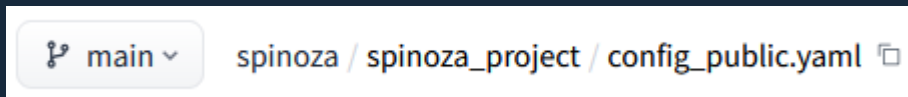
1- Accéder à la liste des modèles disponibles : <https://console.groq.com/settings/limits>

Limits				
ID	REQUESTS PER MINUTE	REQUESTS PER DAY	TOKENS PER MINUTE	TOKENS PER DAY
allam-2-7b	30	7 000	6 000	(No limit)
deepseek-r1-distill-llama-70b	30	1 000	6 000	(No limit)
deepseek-r1-distill-qwen-32b	30	1 000	6 000	(No limit)
gemma2-9b-it	30	14 400	15 000	500 000
llama-3.1-8b-instant	30	14 400	6 000	500 000
llama-3.2-11b-vision-preview	30	7 000	7 000	500 000
llama-3.2-1b-preview	30	7 000	7 000	500 000
llama-3.2-3b-preview	30	7 000	7 000	500 000
llama-3.2-90b-vision-preview	15	3 500	7 000	250 000
llama-3.3-70b-specdec	30	1 000	6 000	100 000
llama-3.3-70b-versatile	30	1 000	6 000	100 000
llama-guard-3-8b	30	14 400	15 000	500 000
llama3-70b-8192	30	14 400	6 000	500 000
llama3-8b-8192	30	14 400	6 000	500 000
mistral-saba-24b	30	1 000	6 000	500 000
qwen-2.5-32b	30	1 000	6 000	(No limit)
qwen-2.5-coder-32b	30	1 000	6 000	(No limit)
qwen-qwq-32b	30	1 000	6 000	(No limit)

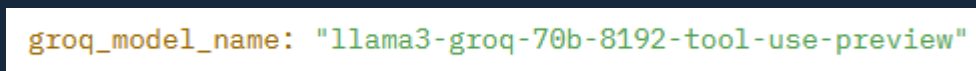
*Le modèle que nous recommandons à date*

# Comment changer le modèle (optionnel) ?

1- Revenez sur votre application et allez modifier le fichier `config_public.yaml` (il contient les éléments de configuration)



2- Modifiez ce paramètre avec le nom de du modèle souhaité et disponible dans l'app groq (cf slide precedente)



3- Déployez vos réponses comme indiqué dans cette slide (vous pouvez cliquer dessus c'est une redirection)

**Faites vos premières modifications et publiez-les :**

1- Cliquez sur éditer

A screenshot of a GitHub commit interface. The page title is 'Faites vos premières modifications et publiez-les :'. Below it, there are three numbered instructions: 1- Cliquez sur éditer, 2- Faites vos premières modification (eg : "Provide your answer in old French."), and 3- Déployez vos réponses, et expliquez brièvement vos modifications dans le champ texte dédié. The interface shows a commit message field with the text 'Modification de la réponse de l'agent de loi dans un style de vieux Français'. There are 'Edit' and 'Preview' tabs, and a 'Commit changes to main' button at the bottom. Red boxes highlight the 'edit' button in the top navigation, the commit message field, and the 'Commit changes to main' button.

2- Faites vos premières modification (eg : "Provide your answer in old French.")

3- Déployez vos réponses, et expliquez brièvement vos modifications dans le champ texte dédié

Commit changes

Modification de la réponse de l'agent de loi dans un style de vieux Français

Edit Preview

Add an extended description...

Commit changes to main Cancel

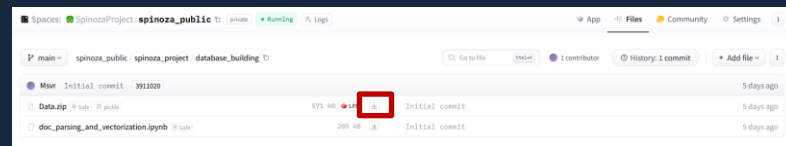
4- Attendez que l'application redémarre et testez vos modifications

# 3.a – Intégrer vos premières données

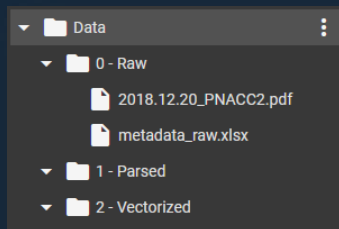
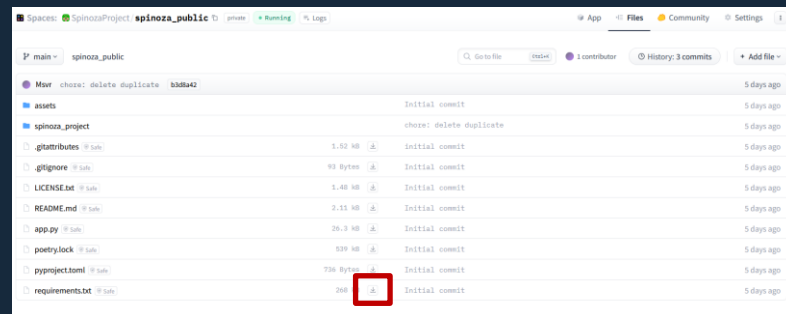
# Télécharger les données de test

Vous pouvez aussi accéder à ces données aux emplacements suivants :

1 – Téléchargez les données de test pour le hackathon (le plan national d'adaptation au changement climatique) : [lien](#)



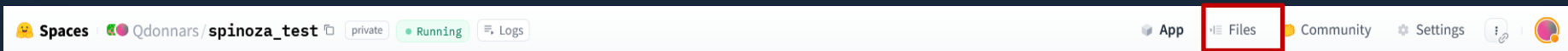
2 – Téléchargez les requirements.txt : [lien](#)



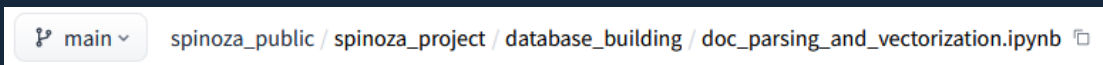
Info : Ce fichier ne contient que un document d'exemple pour le parsing ainsi qu'un fichier excel qui permet de lister les données supplémentaires

# Ouvrir le notebook de parsing dans Colab (compte google requis)

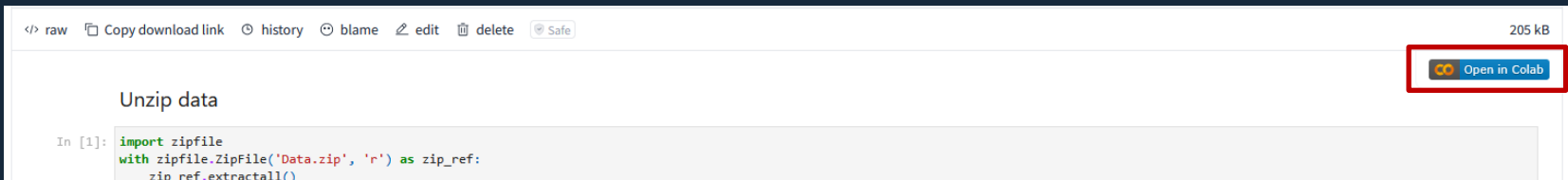
0- Déplacez vous dans les fichiers de **votre répertoire**



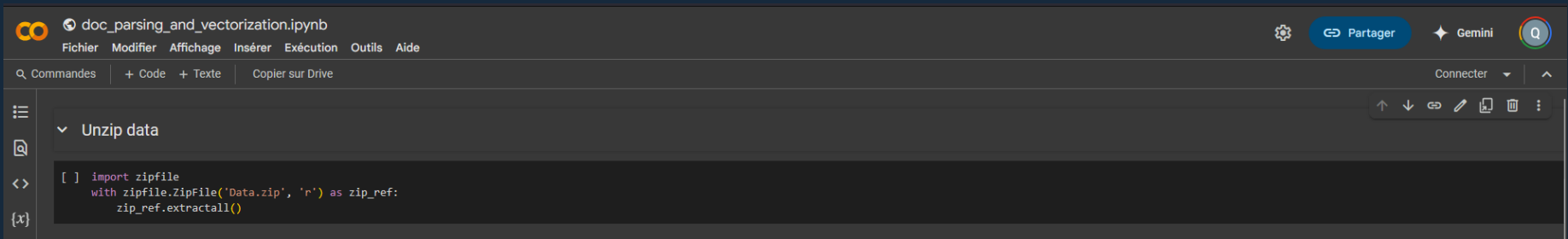
1- Accédez au notebook dans le projet :



2- Ouvrez le dans Colab :

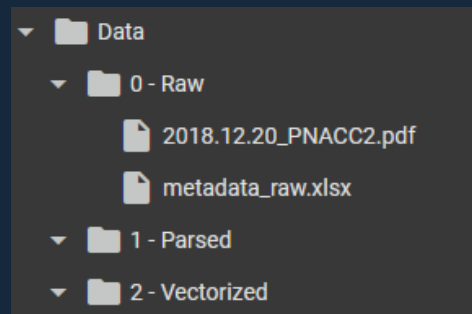
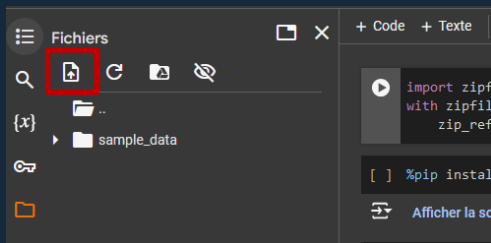


3- Vous devez maintenant voir une page web similaire (potentiellement sur fond blanc)

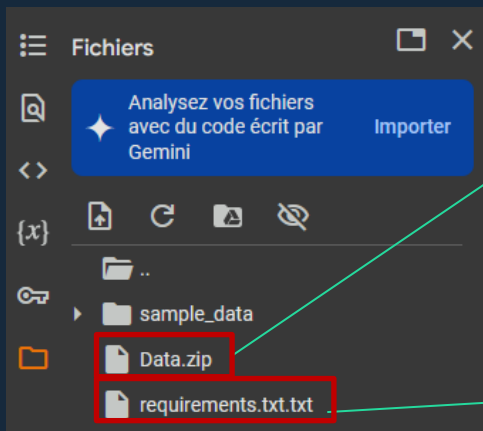


# Intégrez les données pour votre premier run

0- Importez vos données (data.zip & requirements.txt)



1- Votre répertoire doit donc ressembler à ça



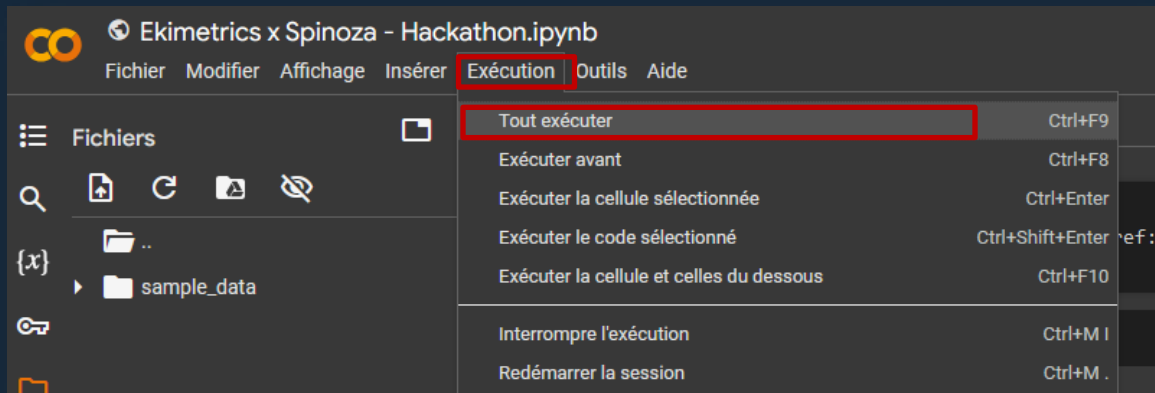
Ce fichier contient :

- Raw : les données (au format pdf) et les informations relatives à chacune de ces données
- Parser & vectorized : de fichier vides où seront stockées des données de traitement intermédiaires

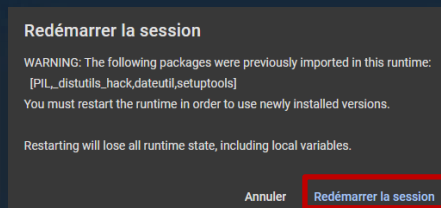
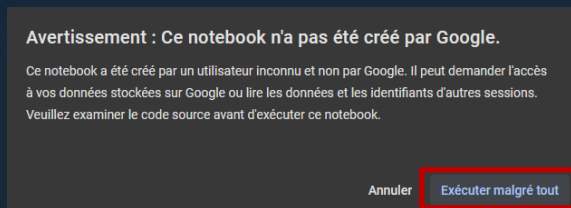
Ce fichier contient la liste des bibliothèques open source nécessaire à l'exécution du programme

# Lancez votre premier run

1- Exécutez toutes les cellules (cela peut prendre environ 10mn):



2- Ces deux messages pourraient apparaître, réalisez les actions suivantes :

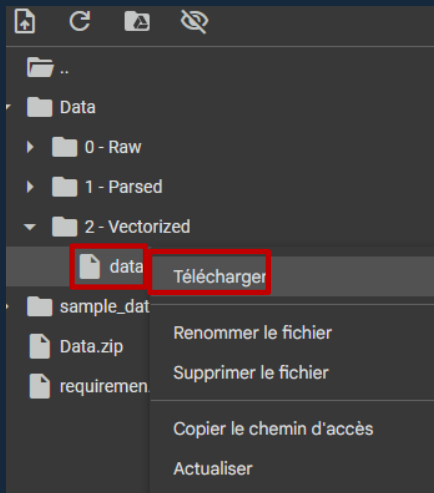


Armez vous de patience, le run va durer une dizaine de minutes

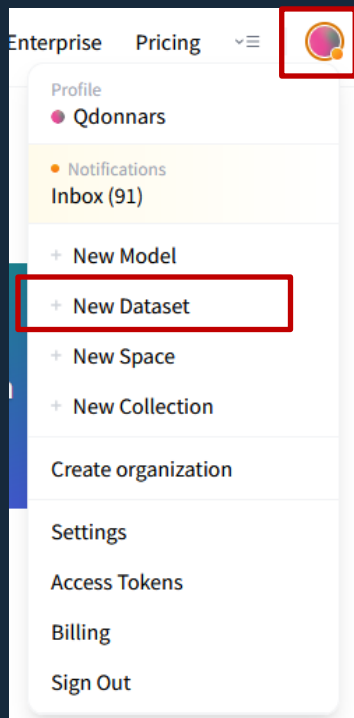
# **3.b – Intégration de nouvelles données**

# Récupérez vos données dans Colab et créez un dataset dans Hugging Face

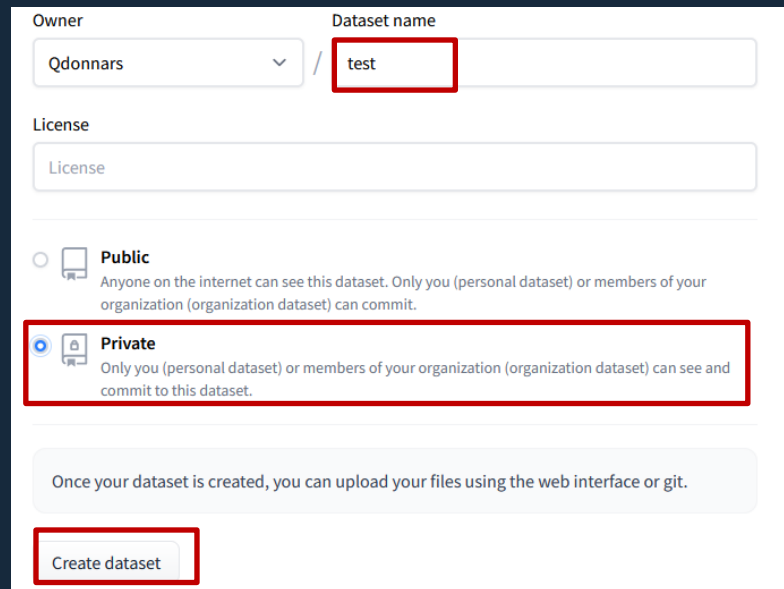
1- Téléchargez le fichier pkl dans le dossier 2 - Vectorized



2- Allez sur votre répertoire hugging face et créez un dataset

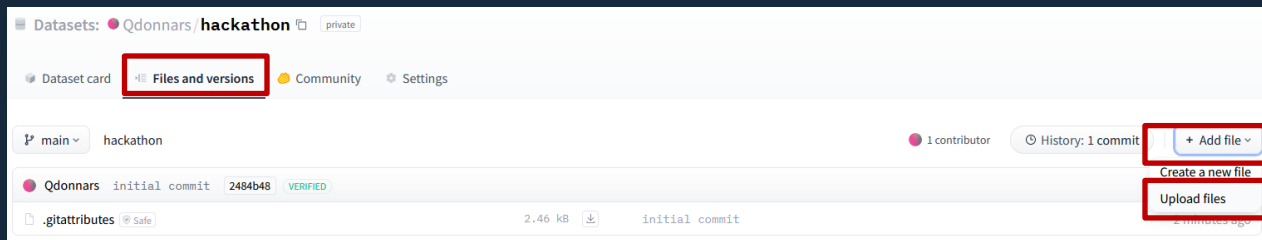


3- Donnez un nom à ce dataset, rendez le privé, et créez ce dataset



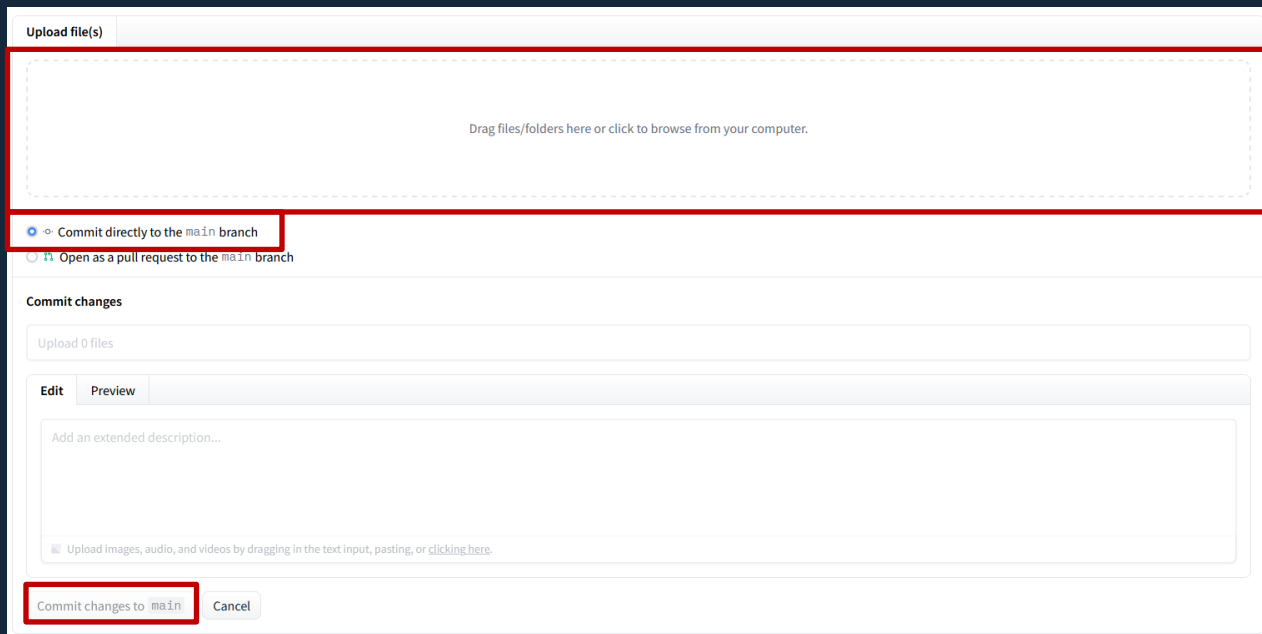
# Téléversez les données que vous avez téléchargées dans ce dataset

1- Allez dans l'onglet de téléversement de fichier sur votre « dataset »



The screenshot shows the Datasets interface for a dataset named 'hackathon'. The 'Files and versions' tab is selected and highlighted with a red box. In the top right corner, the '+ Add file' button is also highlighted with a red box. Below it, the options 'Create a new file' and 'Upload files' are visible.

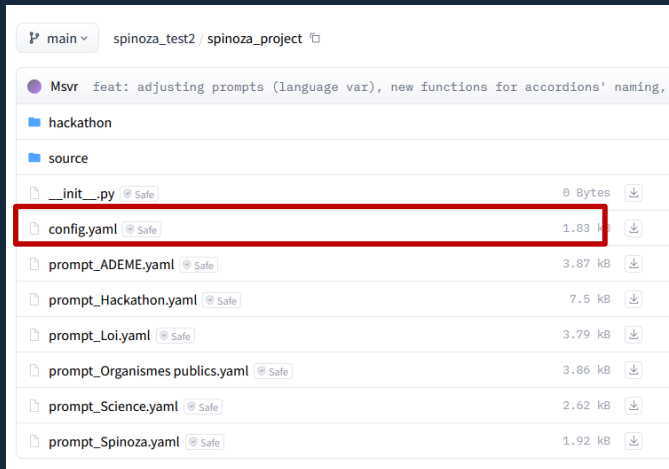
2- Glissez déposez votre fichier Database.Pickle



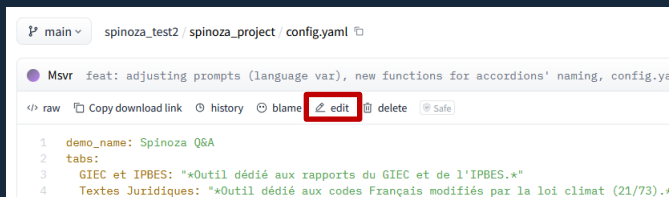
The screenshot shows the file upload interface. A large dashed red box highlights the drag-and-drop area with the text 'Drag files/folders here or click to browse from your computer.' Below this, the 'Commit directly to the main branch' option is selected and highlighted with a red box. The 'Commit changes' section is visible, showing 'Upload 0 files' and a text input field for an extended description. At the bottom, the 'Commit changes to main' button is highlighted with a red box.

# Paramétrer le nouvel onglet & reliez votre nouveau dataset à l'interface

1- Déplacez vous sur le fichier `spinoza_project/config.yaml`



2- Editez le fichier



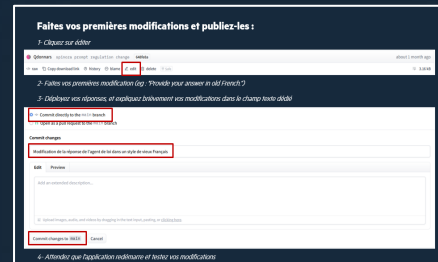
2- Ajoutez les informations relatives au nouvel onglet (`new_agent`) et données, ainsi que le chemin d'accès de la base de données que vous avez créés dans la slide précédente



2- Ainsi que dans les clefs:

- `tabs`
- `en_description`
- `source_mapping`
- `en_names`

3- Déployez vos réponses comme indiqué dans cette slide (lien de redirection en cliquant)



# Bravo vous pouvez désormais profiter de votre interface

The screenshot displays the Spinoza Q&A interface. At the top, the browser address bar shows "Spaces" and "Qdonnars spinoza\_public". The main navigation bar includes "Q&A", "Paramètres des Agents", "Détails des Données", and "À propos & Contact". A text input field prompts the user to "Posez votre question ici !". Below this is a list of agents: "Agent Science", "Agent Loi", "Agent Organismes publics", "Agent ADEME", and "Agent PNACC", which is highlighted with a red box. The "Spinoza" agent is selected, showing a chat window with the following text:

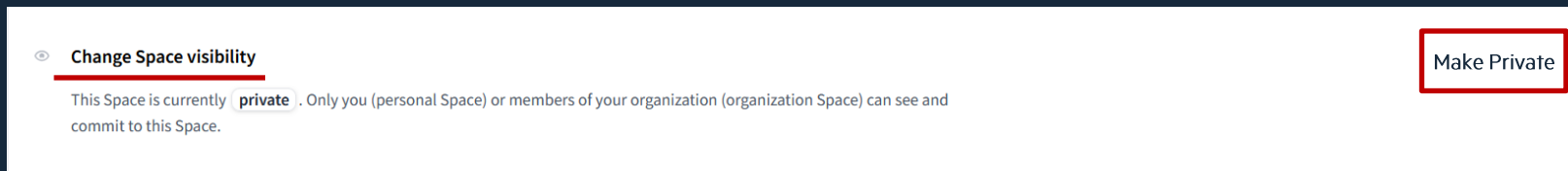
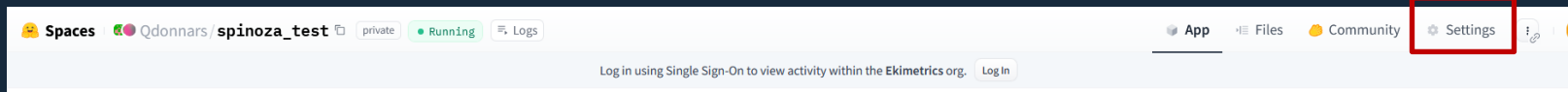
Bonjour, je suis Spinoza, un assistant conversationnel expert sur le climat conçu pour vous aider dans votre parcours journalistique. Je répondrai à vos questions en lien avec le climat en me basant sur **les sources fournies**.

**Limitations**  
Veuillez noter que ce système de questionnement est à un stade précoce, il n'est pas parfait et peut parfois donner des réponses non pertinentes. Si vous n'êtes pas satisfait de la réponse, veuillez poser une question plus spécifique ou signaler vos commentaires pour nous aider à améliorer le système.

Que voulez-vous apprendre ?

À l'initiative de RSF et l'Alliance Presse · Avec le soutien du Ministère de la Culture · Conçu par Ekimetrics

## Optionnel : repassez votre répertoire en privé



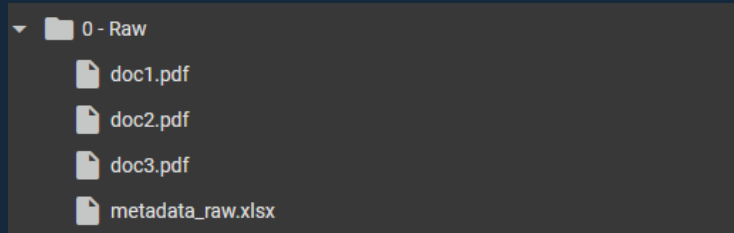
Quelles conséquences si le répertoire est public :

- Tout le monde peut accéder à l'application (et donc vos journalistes pour cet exercice)
- Vos données peuvent rester privées (cependant accessibles via l'app)

# **3.c – Intégrez votre premier corpus**

# Définissez les données de votre nouvelle base de données

1- Ajoutez l'ensemble des documents désirés dans le fichier 0 - Raw



2- Documentez les informations sur l'ensemble des ces documents dans le fichier metadata.raw.xlsx

file_source_type	file_name	file_title	file_url	file_data_publishing	file_filtering_modality
test	doc1	titre1	<a href="https://url1.com">https://url1.com</a>	2025	Document Public
test	doc2	titre2	<a href="https://url2.com">https://url2.com</a>	2024	Document Privé
test	doc3	titre3	<a href="https://url3.com">https://url3.com</a>	2023	Document Privé

Dans cette colonne, veuillez renseigner le **nom de la base de données** de laquelle provient le document.  
**ATTENTION** : Le nom de la database attendu est celui de la valeur renseignée dans la section "source\_mapping" du fichier config.yaml modifié au préalable

Dans cette colonne, veuillez renseigner les **noms** des documents PDF utilisés

Dans cette colonne, veuillez renseigner les **titres** des documents

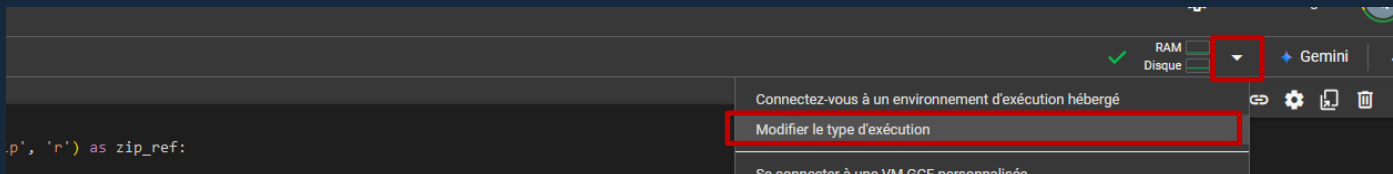
Dans cette colonne, veuillez renseigner les **URLs** des documents qui permettra à l'outil de rediriger vers les bon liens (quand vous cliquerez sur la source)

Dans cette colonne, veuillez renseigner les **dates de publication** des documents

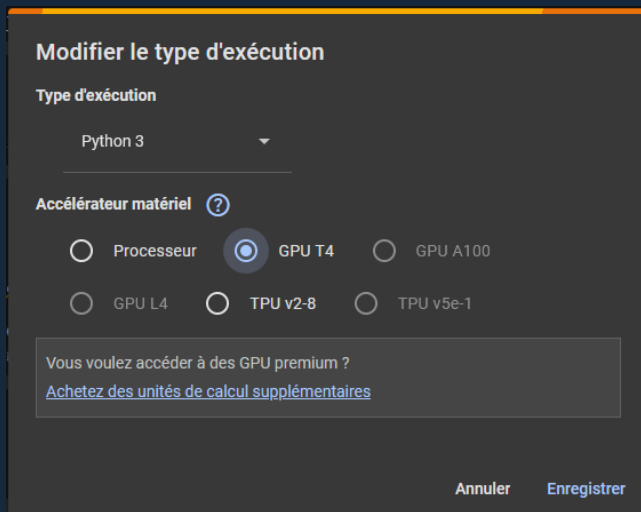
Dans cette colonne, veuillez renseigner les **catégories** auxquels les documents appartiennent et sur lesquels ils seront **FILTRÉS**

# Initiez votre premier run

1- Associez la bonne ressource :



2- Associez la bonne ressource :T4 GPU (4h max / jours) pour vos données (nous vous conseillons tout de même dans ce cadre d'apporter moins de 4000 pages de documents)



# Changer le nom de vos données

## 1- Nommez Votre nouveau jeu de données

```
[ ] # ici on charge les embeddings via HuggingFace, on peut aussi utiliser d'autres embeddings
embed_model=embeddings = HuggingFaceEmbeddings(
    model_name="intfloat/multilingual-e5-base"
)

preprompt="passage: "
```

## 2- Importez le puis définissez ses paramètres dans le fichier config comme indiqué dans ces slides (redirection en cliquant)

**Récupérez vos données dans Colab et créez un dataset dans Hugging Face**

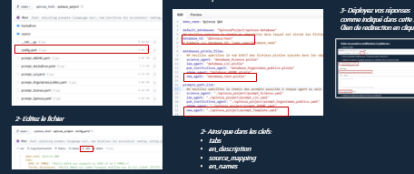
- 1- Téléchargez le fichier pdf dans le fichier 2 - voir terminal
- 2- Allez sur votre répertoire Hugging Face et créez un dataset
- 3- Cliquez sur nom à ce dataset, rendez le privé, et créez un dataset



The screenshot shows the Hugging Face web interface. On the left, the 'Datasets' menu is highlighted. In the center, the 'New Dataset' dialog is open, with the 'name' field set to 'Test' and the 'visibility' dropdown set to 'Private'. A red box highlights the 'name' field. At the bottom, the 'Create dataset' button is highlighted with a red box.

**Paramétrer le nouvel onglet & relier votre nouveau dataset à l'interface**

- 1- Déplacez vous sur le fichier `python_app.py` et modifiez le fichier
- 2- Ajoutez les informations relatives au nouvel onglet `name`, `vector_name`, ainsi que le chemin absolu de la base de données que vous avez créée dans la slide précédente
- 3- Déployez une réponse comme toujours dans cette slide (lien de redirection en cliquant)



The screenshot shows a code editor with the `python_app.py` file. The `config` dictionary is updated with the following values: `name = "Test"`, `vector_name = "Test"`, and `db_path = "/home/colab/.cache/huggingface/datasets/Test"`. A red box highlights the `name` assignment. Below the code, a list of variables is shown: `name`, `vector_name`, `db_path`, `embed_model`, `embeddings`, `preprompt`, and `embed_and_dump`. A red box highlights the `name` variable in this list.

An aerial photograph of a circular pond surrounded by a dense forest. The pond is a light blue color, and the surrounding trees are a deep green. The word "Merci!" is written in white, bold, serif font in the center of the pond.

**Merci !**

**Ekimetrics.**