

DÉVELOPPEMENT D'UN SYSTÈME RAG POUR LA RECOMMANDATION D'ÉVÉNEMENTS CULTURELS

Projet RAG — OpenAgenda + Mistral + FAISS + LangChain

Projet Data Engineer – KELLENI Antoine
[lien GitHub](#) 1



ROADMAP DE LA MISSION

1. Contexte & problématique
2. Objectifs
3. Architecture du pipeline RAG
4. Différentes étapes du pipeline RAG
5. Conclusion

Contexte & problématique

Contexte :

L'entreprise Puls-Events souhaite tester un assistant intelligent capable de recommander des événements culturels. Les données sources proviennent d'OpenAgenda (événements culturels publics).

Lien : "<https://openagenda.com/fr/ile-de-france>"

Problème métier :

Les utilisateurs souhaitent trouver rapidement des événements pertinents selon leurs préférences.

The screenshot shows the OpenAgenda website interface for the Île-de-France region. At the top, there's a navigation bar with links for 'Aide', 'Français', 'Se connecter', and 'S'inscrire'. Below the header, a banner for 'OpenAgenda en Île-de-France' is displayed, featuring a map icon and a checkmark. A search bar is present above the main content area. The main content area displays two event cards:

- Travailler en temps partagé, pourquoi pas ?**
Il y a environ 3 heures ago.
Illustration of people working around a large table.
Text: "Vous êtes cadre, connaissez-vous les atouts du temps partagé ?"
Details: Mercredi 3 décembre, 09h30, Cité des Métiers de Saint-Quentin-en-Yvelines, Trappes.
Buttons: Partager.
- MILLE SECRETS DE POUSSINS - COLLECTIF QUATRE AILES**
Il y a environ 2 heures ago.
Illustration for 'Mille secrets' by Collectif Quatre Ailes.
Text: "Collectif Quatre Ailes Mille secrets".

On the right side of the page, there's a map of the Île-de-France region with event locations marked by yellow circles of varying sizes, indicating the number of events. Below the map is a calendar for December 2025, and at the bottom, there are buttons for 'Aujourd'hui', 'Demain', 'Ce week-end', 'Cette semaine', and 'Ce mois'.



ROADMAP DE LA MISSION

1. Contexte & problématique
2. **Objectifs**
3. Architecture du pipeline RAG
4. Différentes étapes du pipeline RAG
5. Conclusion

Objectifs

Objectif :

Construire un POC complet utilisant un pipeline RAG (Retrieval Augmented Generation).

Objectifs techniques :

- Collecter automatiquement les événements via l'API OpenAgenda
- Nettoyer et structurer les données
- Découper les descriptions en chunks
- Générer des embeddings avec Mistral
- Indexation FAISS pour recherche rapide
- Construire un chatbot RAG capable de recommander des événements

Objectif final :

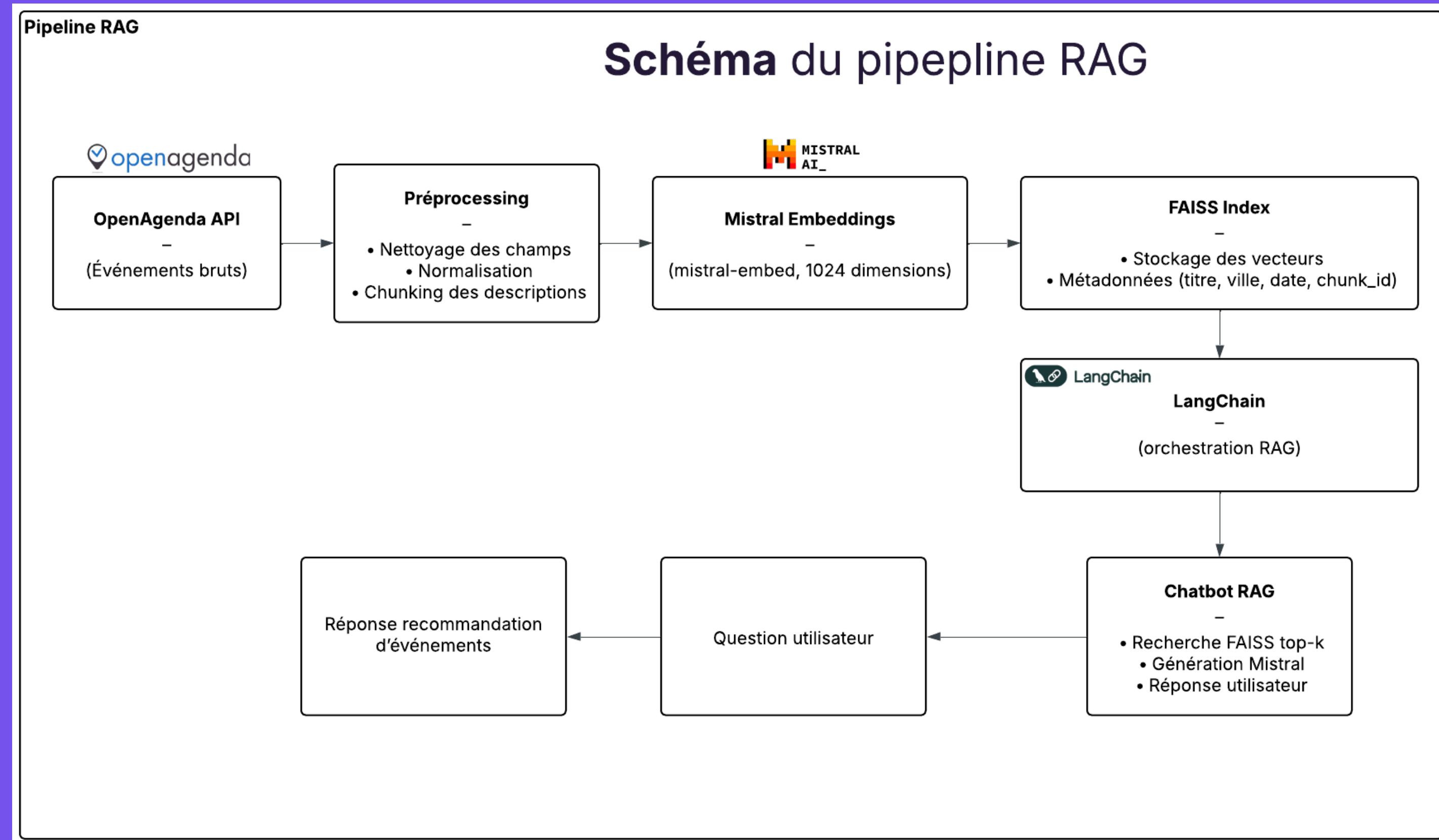
Démontrer la faisabilité d'un système RAG complet.

ROADMAP DE LA MISSION

1. Contexte & problématique
2. Objectifs
- 3. Architecture du pipeline RAG**
4. Différentes étapes du pipeline RAG
5. Conclusion

Architecture du pipeline RAG

Organisation du pipeline



ROADMAP DE LA MISSION

1. Contexte & problématique
2. Objectifs
3. Architecture du pipeline RAG
- 4. Différentes étapes du pipeline RAG**
5. Conclusion

Préparation de l'environnement de travail

```
● PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python --version
Python 3.11.9
● PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python -m venv .venv
● PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> .\.venv\Scripts\Activate.ps1
● (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python -m pip install --upgrade pip
Requirement already satisfied: pip in c:\users\antoi\onedrive\documents\openclassrooms\projet 11 concevoir deployer systeme rag\.venv\lib\site-packages (24.0)
Collecting pip
  Downloading pip-25.3-py3-none-any.whl.metadata (4.7 kB)
  Downloading pip-25.3-py3-none-any.whl (1.8 MB)
    1.8/1.8 MB 22.7 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 24.0
    Uninstalling pip-24.0:
      Successfully uninstalled pip-24.0
  Successfully installed pip-25.3
● (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> pip install langchain faiss-cpu pandas numpy requests python-dotenv mistralai
Collecting langchain
  Downloading langchain-1.1.0-py3-none-any.whl.metadata (4.9 kB)
  Downloading faiss_cpu-1.13.0-cp311-cp311-win_amd64.whl.metadata (7.7 kB)
Collecting pandas
  Using cached pandas-2.3.3-cp311-cp311-win_amd64.whl.metadata (19 kB)
```

```
● (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python test_install.py
OK Chargement .env
OK Début vérification des imports...
OK Import langchain
OK Import langchain_core
OK Import langchain_community
OK Import faiss
OK Import numpy
OK Import pandas
OK Import requests
OK Import dotenv
OK Import mistralai
OK Tous les imports sont corrects
OK Clé Mistral chargée
OK Appel API Mistral opérationnel
OK environnement prêt !
❖ (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG>
```

Création d'un environnement Python 3.11

Installation des dépendances (langchain, faiss-cpu, mistralai, pandas, etc.)

Vérification complète via test_install.py

Chargement de la clé API Mistral

Test API Mistral : OK

Préprocessing OpenAgenda

```
● (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deploier systeme RAG> python .\openagenda_preprocessing.py
OK Chargement .env
OK Variables d'environnement chargées
OK Début préprocessing OpenAgenda
OK Dataset complet déjà présent : data\openagenda_events_full_1an.csv
OK Chargé depuis le cache : 15168 événements
OK Filtre (multi-villes) : ['Paris', 'Versailles', 'Viroflay', 'Plaisir', 'Jouy-en-Josas', 'Roissy-en-France'] -> 4598 événements
OK DataFrame après nettoyage : 4598 lignes
OK Événements à venir uniquement : 607 lignes
OK Limitation événements : 100 sur 607 (les plus proches à venir)
OK DataFrame retenu pour chunking : 100 événements
OK Chunking : 100 événements -> 148 chunks
OK DataFrame final pour embeddings : 148 lignes
OK Calcul embeddings : 148 textes
OK Batch 0 -> 64
OK Batch 64 -> 128
OK Batch 128 -> 148
OK Embeddings générés
OK Données sauvegardées dans data\openagenda_events_preprocessed.csv
OK Embeddings sauvegardés dans data\openagenda_events_embeddings.npy
OK Préprocessing terminé
● (.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deploier systeme RAG> python .\build_faiss_index.py
```

- Récupération des événements sur 1 an : 15 168 événements
- Normalisation des champs : titre, description, longue description, ville, horaires
- Nettoyage des textes
- Production du fichier → openagenda_events_full_1an.csv

Points techniques :

- Gestion pagination
- Filtre sur les événements pertinents
- Logs complets des différentes étapes

Chunking

Pourquoi le chunking ?

1. Pour découper les descriptions trop longues
 2. Mieux représenter la sémantique
 3. Améliorer la recherche vectorielle FAISS

```
data > openagenda_events_preprocessed.csv > data
1 uid,slug,title,description_short,description_long,conditions,keywords,city,venue_name,first_begin,first_end,text_for_embedding,chunk_id,chunk_text
2 92816716,geremy-credeville-enchanteur-5746295,GÉRÉMY CRÉDEVILLE - Enchanteur,Humour,"Dans son nouveau one-man-show, Géraly Crédéville explore la thématique du choix avec cette question : « Pourquoi choisir ? ». Refusant de se cantonner à une seule discipline, il revisite les classiques et invente de nouvelles formes d'humour. Les conditions sont les suivantes : 49 € Carré or / 39 € Cat 1 / 35 € Cat 2.",0,"GÉRÉMY CRÉDEVILLE - Enchanteur
3
4 Humour
5
6 Dans son nouveau one-man-show, Géraly Crédéville explore la thématique du choix avec cette question : « Pourquoi choisir ? ». Refusant de se cantonner à une seule discipline, il revisite les classiques et invente de nouvelles formes d'humour. Les conditions sont les suivantes : 49 € Carré or / 39 € Cat 1 / 35 € Cat 2."
7
8 Conditions : 49 € Carré or / 39 € Cat 1 / 35 € Cat 2.",0,"GÉRÉMY CRÉDEVILLE - Enchanteur
9
10 Humour
11
12 Dans son nouveau one-man-show, Géraly Crédéville explore la thématique du choix avec cette question : « Pourquoi choisir ? ». Refusant de se cantonner à une seule discipline, il revisite les classiques et invente de nouvelles formes d'humour. Les conditions sont les suivantes : 49 € Carré or / 39 € Cat 1 / 35 € Cat 2."
13
14 Conditions : 49 € Carré or / 39 € Cat 1 / 35 € Cat 2."
15 83375880,nadiya-et-cest-re-parti-tour,Nâdiya - Et c'est re-parti Tour !,"Artiste majeure de la scène pop des années 2000 avec plus de 5 millions de disques vendus, NÂDIYA est l'une des artistes les plus populaires et énergiques de son époque. Ses chansons, caractérisées par leur rythme entraînant et leurs paroles directes, ont marqué plusieurs générations. NÂDIYA a également collaboré avec de nombreux artistes internationaux, contribuant ainsi à la diversité de la musique mondiale. Ses performances live sont légendaires, mêlant danse, chant et interaction avec le public. Aujourd'hui, NÂDIYA continue d'enchanter ses fans à travers le monde avec ses nouveaux projets et sa passion pour la musique.""
16
17 ### NÂDIYA - ET C'EST RE-PARTI TOUR !
18
19 Artiste majeure de la scène pop des années 2000 avec plus de 5 millions de disques vendus, NÂDIYA a marqué les années 2000 avec sa musique vivante, populaire et énergique. Icône de toute une génération, elle se distingue par sa musique iconique et vivante qui mélange pop et influences urbaines, créant un son distinctif qui la rend unique.
```

Embeddings Mistral

Modèle utilisé : mistral-embed

Dimension : 1024

Calcul par batchs de 64

Résultat obtenu → openagenda_events_embeddings.npy

Contraintes rencontrées :

- Quota Mistral → limitation à 200 embeddings
- Gestion des erreurs API (code 429)

```
OK Batch 0 -> 64
OK Batch 64 -> 128
OK Batch 128 -> 148
OK Embeddings générés
OK Données sauvegardées dans data\openagenda_events_preprocessed.csv
OK Embeddings sauvegardés dans data\openagenda_events_embeddings.npy
OK Préprocessing terminé
```

```
raise models.SDKError("API error occurred", http_res, http_res_text)
mistralai.models.sdkerror.SDKError: API error occurred: Status 429. Body: {"object": "error", "message": "Service tier capacity exceeded for this model."}
```

Construction de l'index FAISS

Avec LangChain + FAISS :

1. Chargement des embeddings
2. Création de l'index IndexFlatL2
3. Stockage des métadonnées : titre, ville, date, chunk_id
4. Sauvegarde locale → /data/faiss_openagenda_lc/

Test interne FAISS :

Requête : "concert à Paris"

```
● (.venv) PS C:\Users\antois\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python .\build_faiss_index.py
OK Chargement .env
OK DataFrame prétraité : 148 lignes (chunks)
OK Embeddings : shape = (148, 1024)
OK Construction du vector store FAISS (LangChain) à partir des embeddings pré-calculés...
OK Vector store FAISS créé avec 148 vecteurs
OK Vector store sauvegardé dans data\faiss_openagenda_lc
OK Test de recherche : 'concert à Paris' ...
--- Résultat 1 ---
Titre : DU SPECTACLE
Ville : Paris
Début : 2025-12-05T20:30:00+01:00
Chunk id : 0
Extrait chunk : DU SPECTACLE Nous sommes Te Beijo, Pierre Mahier et Vaslo. Nous avons créé Du Spectacle pour faire voyager la musique de notre amie artiste Emmanuelle Cadoret le plus loin et le plus longtemps possible
--- Résultat 2 ---
Titre : STÉPHANE DERADDI
Ville : Paris
Début : 2025-12-12T20:30:00+01:00
Chunk id : 0
Extrait chunk : STÉPHANE DERADDI Concert qui sera l'occasion de célébrer la sortie du deuxième album « Toujours à gauche ». La grande occasion aussi d'avoir des invités sur scène ayant participé ou pas à l'enregistrement
--- Résultat 3 ---
Titre : Voix-là #2 : concert de Bára Zmeková
Ville : Paris
Début : 2025-12-10T20:00:00+01:00
Chunk id : 0
Extrait chunk : Voix-là #2 : concert de Bára Zmeková Pour le deuxième concert de la série Voix-là, le Centre tchèque accueillera la chanteuse et compositrice Bára Zmeková ! Voix-là: concert au féminin -----
OK Étape 3 terminée : index FAISS + métadonnées prêts pour l'étape 4.
✿ (.venv) PS C:\Users\antois\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python .\chatbot_rag.py
```



Chatbot intelligent

CHATBOT RAG (FAISS + MISTRAL)

Pipeline RAG :

1. Embedding de la requête utilisateur
2. Recherche FAISS top-k
3. Sélection du contexte pertinent
4. Génération Mistral
5. Réponse naturelle et contextualisée

Fonctionnalités :

- Recommandation d'événements
- Filtrage implicite par ville / type
- Reformulation intelligente

```
(.venv) PS C:\Users\antoi\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir depoyer systeme RAG> python .\chatbot_rag.py
OK Chatbot RAG prêt (FAISS + Mistral). Tape 'quit' pour sortir.

Vous: Combien avez vous en possession d'événements ? Peux tu me détailler le nombre par ville aussi

Bot:
- Jouy-en-Josas : 2 événements
- Paris : 27 événements
- Plaisir : 2 événements
- Roissy-en-France : 3 événements
- Versailles : 61 événements
- Viroflay : 5 événements

Total général d'événements : 100

Vous: J'aimerai sortir le weekend à versailles, as tu des recommandations à me proposer pour une sortie en amoureux ?

Bot: Pour une sortie en amoureux à Versailles ce week-end, je te recommande les événements suivants :

1. **Journée découverte au Centre de musique baroque de Versailles** (le 6 décembre)
   - **Pourquoi ?** : Cette journée propose une variété d'activités comme des ateliers, des spectacles, des visites guidées et un jeu de piste numérique. C'est une excellente occasion de passer un moment culturel et ludique ensemble. Le jeu de piste "La saga Papillon à Versailles" est particulièrement adapté pour une sortie en couple, car il permet de découvrir Versailles de manière interactive et amusante.

2. **Soirée dansante AFMTELÉTHON - Versailles swing Danse** (le 6 décembre)
   - **Pourquoi ?** : Si vous aimez danser, cette soirée organisée par Versailles Swing Danse pour le Téléthon est une excellente option. L'ambiance festive et solidaire peut rendre votre sortie encore plus spéciale. De plus, le prix d'entrée est accessible (5 € minimum), ce qui permet de profiter de la soirée sans se ruiner.

Malheureusement, la braderie de jouets ne semble pas adaptée pour une sortie en amoureux.

Vous: pour cette fin d'année, je compte passer 2 jours à Paris, peut tu me proposer 3 activités à faire en famille avec mes enfants ?

Bot: Voici trois activités adaptées pour une sortie en famille à Paris pendant les fêtes de fin d'année :

1. **Contes et histoires : Noël en famille**
   - **Lieu** : Hôtel de la Marine, Paris
   - **Date** : 5 décembre 2025 à 10h30
   - **Description** : Cet événement propose des contes et des histoires pour célébrer Noël en famille dans un cadre scintillant. C'est une activité idéale pour les enfants, qui pourront profiter de l'ambiance festive tout en écoutant des récits captivants.

2. **Sortie loisirs : Escape Game**
   - **Lieu** : Paris
   - **Date** : 13 décembre 2025 à 15h00
   - **Description** : Une sortie en escape game pour les jeunes de 11 à 17 ans. Les enfants pourront résoudre des énigmes, fouiller pour trouver des indices et vivre une aventure ludique. Cette activité est parfaite pour une journée en famille, où les enfants pourront tester leurs capacités de résolution de problèmes et de travail d'équipe.

3. **Sortie loisirs : Escape Game**
   - **Lieu** : Paris
   - **Date** : 13 décembre 2025 à 15h00
   - **Description** : Une autre option d'escape game, cette fois dans une salle de 600m². Les enfants pourront vivre le frisson d'une chasse au trésor et résoudre des énigmes dans un environnement immersif. Cette activité est idéale pour une journée en famille, où les enfants pourront développer leurs compétences en communication et en gestion du temps.
```

Chatbot intelligent

CHATBOT RAG (FAISS + MISTRAL)

chatbot_rag.py

Vous: à partir du 06 decembre, je suis en vacances, j'ai posé 4 semaines de vacances, peut tu me donner une liste d'activité à faire durant les jours à venir ?

Bot: Voici une sélection d'activités culturelles en île-de-France pour les jours à venir, en fonction de vos vacances à partir du 6 décembre :

1. **Café-philo à Roissy-en-France** (6 décembre)
- **Pourquoi y aller ?** : Un événement intellectuel et convivial pour réfléchir sur des thèmes importants comme l'altérité et le vivre-ensemble. Parfait pour une après-midi stimulante.
2. **Journée découverte à Versailles** (6 décembre)
- **Pourquoi y aller ?** : Une journée riche en activités avec des ateliers, spectacles, visites guidées et un jeu de piste numérique. Idéal pour une sortie culturelle et ludique.
3. **Marché de Noël à Jouy-en-Josas** (6 décembre)
- **Pourquoi y aller ?** : Un marché de Noël traditionnel avec des créateurs, des animations et des visites. Parfait pour entrer dans l'ambiance des fêtes de fin d'année.
4. **Marché de Noël à Roissy-en-France** (12-14 décembre)
- **Pourquoi y aller ?** : Un marché de Noël avec des ateliers manuels, des déambulations et une grande parade féérique. Idéal pour une sortie en famille avec des enfants.
5. **"La Fabrique de Noël" à Versailles** (13 décembre)
- **Pourquoi y aller ?** : Un atelier créatif pour enfants où ils pourront fabriquer des cadeaux et des décorations de Noël. Parfait pour une activité manuelle et festive.

Ces événements offrent une variété d'activités culturelles et festives pour profiter de vos vacances en île-de-France.

Vous: as tu une activité sportif à me proposer ?

Bot: D'après les événements fournis, il n'y a pas d'activité sportive spécifique proposée. Cependant, voici une option qui pourrait t'intéresser :

- **Sortie loisirs : Escape Game** à Paris le 13 décembre 2025. Bien que ce ne soit pas une activité sportive traditionnelle, l'escape game implique une activité physique légère, de la résolution d'éénigmes et une bonne communication entre les participants. Il se déroule dans une salle de 600m², ce qui ajoute une dimension physique à l'expérience.

Validation fonctionnelle

TESTS UNITAIRES (PYTEST)

Tests réalisés :

- Vérification du chunking
- Vérification de la normalisation
- Alignement CSV ↔ embeddings
- Index FAISS minimal
- Validation du chargement des fichiers

Résultat global :

```
keyour@keyour-OptiPlex-5090:~/.venv PS C:\Users\antoine\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python -m pytest tests\test_faiss_index.py
===== test session starts =====
platform win32 -- Python 3.11.9, pytest-9.0.1, pluggy-1.6.0
rootdir: C:\Users\antoine\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG
plugins: anyio-4.11.0, langsmith-0.4.47
collected 1 item

tests\test_faiss_index.py . [100%]

===== warnings summary =====
tests/test_faiss_index.py::test_faiss_index_construction_and_search
<frozen importlib._bootstrap>:241: DeprecationWarning: builtin type SwigPyPacked has no __module__ attribute

tests/test_faiss_index.py::test_faiss_index_construction_and_search
<frozen importlib._bootstrap>:241: DeprecationWarning: builtin type SwigPyObject has no __module__ attribute

tests/test_faiss_index.py::test_faiss_index_construction_and_search
<frozen importlib._bootstrap>:241: DeprecationWarning: builtin type swigvarlink has no __module__ attribute

-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
===== 1 passed, 3 warnings in 0.64s =====
● (.venv) PS C:\Users\antoine\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG> python -m pytest tests\test_preprocessing.py
===== test session starts =====
platform win32 -- Python 3.11.9, pytest-9.0.1, pluggy-1.6.0
rootdir: C:\Users\antoine\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG
plugins: anyio-4.11.0, langsmith-0.4.47
collected 3 items

tests\test_preprocessing.py ... [100%]

===== 3 passed in 0.53s =====
● (.venv) PS C:\Users\antoine\OneDrive\Documents\OPENCLASSROOMS\Projet 11 Concevoir deployer systeme RAG>
```



ROADMAP DE LA MISSION

1. Contexte & objectifs
2. Données sources (Input Files)
3. Architecture d'automatisation/data lineage
4. Différentes étapes du Workflow
- 5. Conclusion**

Limites du POC

Limites identifiées :

- Quota limité de Mistral
- Données OpenAgenda très volumineuses
- Pas encore d'interface web

Perspectives d'amélioration

Améliorations possibles :

1. UI Streamlit pour une interface utilisateur
2. Gestion multi-agendas OpenAgenda
3. Déploiement Azure ou GCP
4. Ajout d'un historique conversationnel

Conclusion

Pipeline RAG complet
Chatbot fonctionnel et pertinent
Indexation FAISS efficace
Architecture facilement extensible
Objectifs 100% atteints

MERCI

19