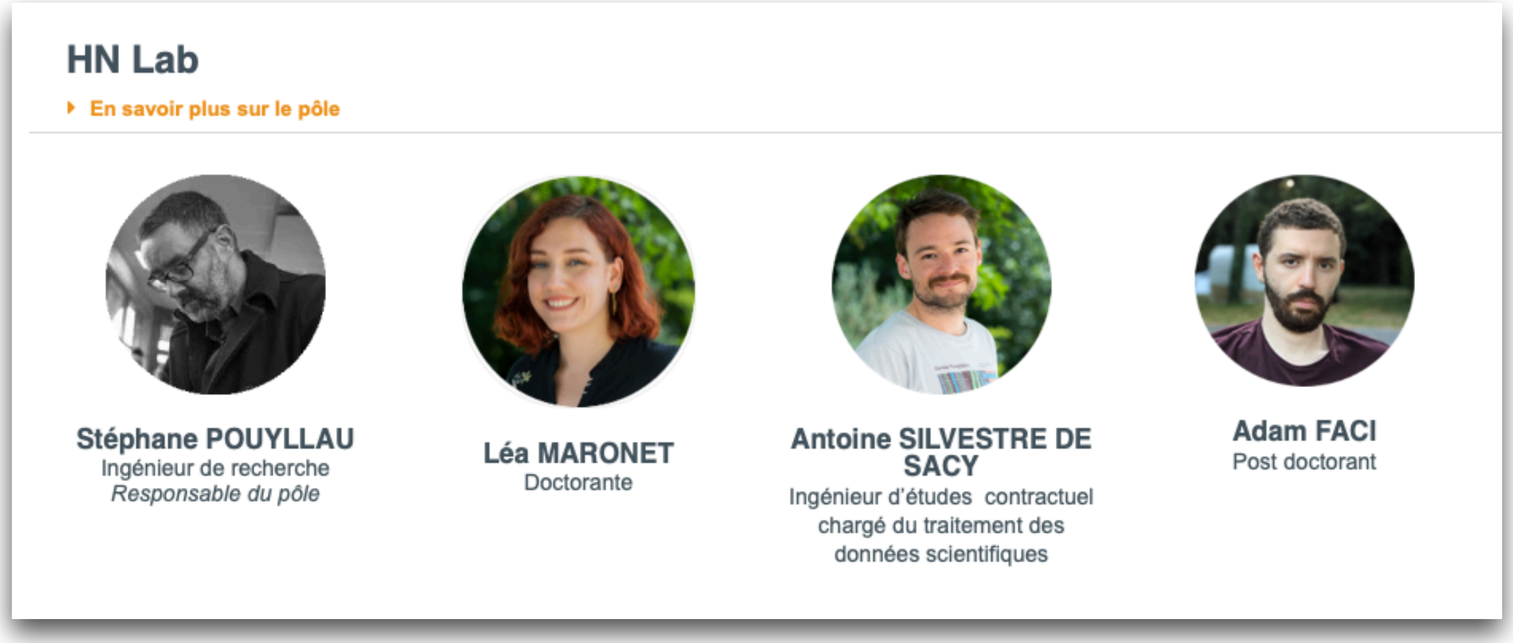
Pre-targeted-RAG

Retrieval Augmented Generation sur des groupes pré-ciblés de communautés d'articles de recherche

Adam Faci
Antoine Silvestre de Sacy

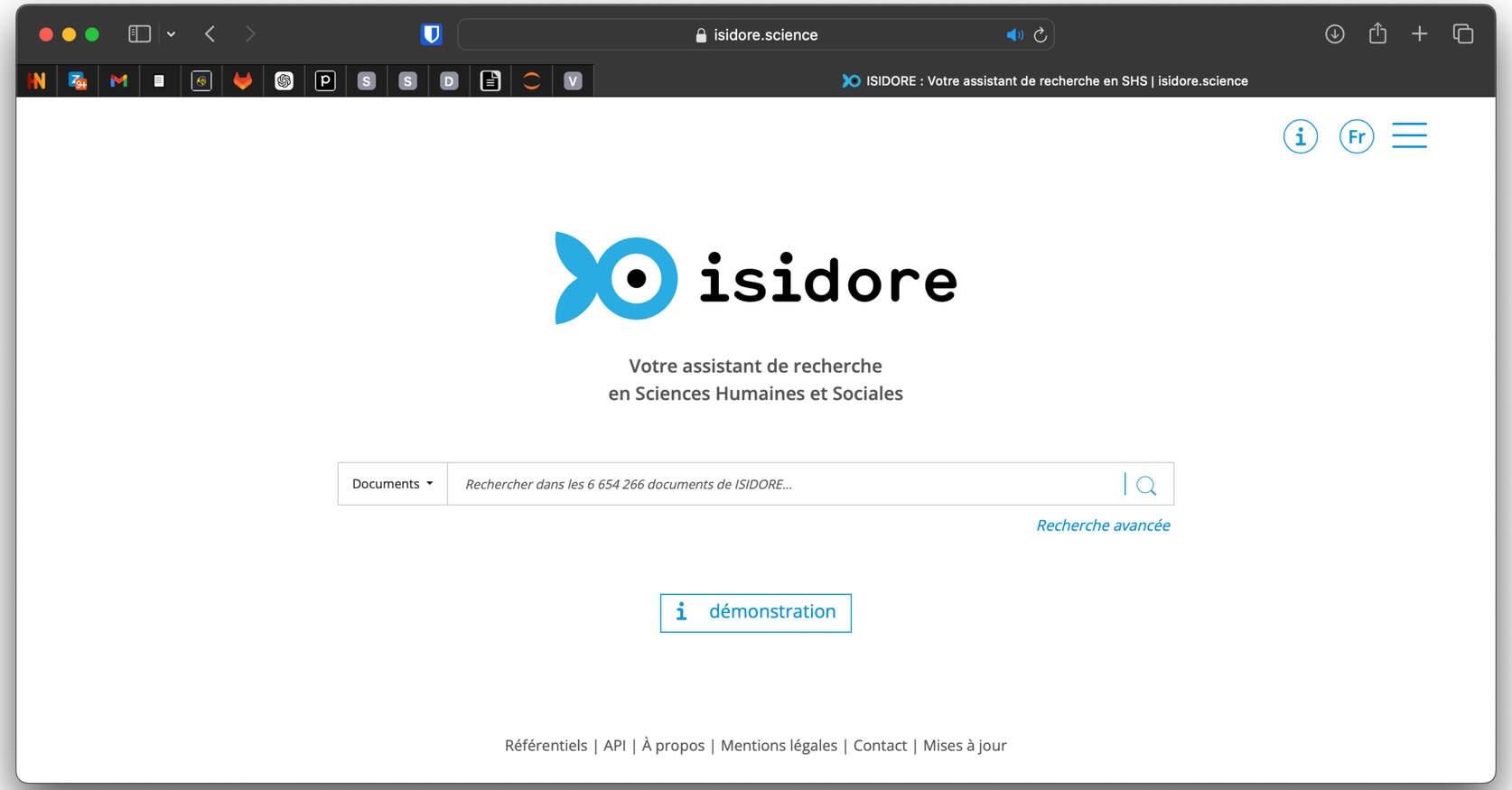
Huma-Num cnrs HN Lab

Département R&D de l'IR Huma-Num (UAR 3598), CNRS*



Introduction - Contexte

Refonte du moteur de recherche académique Isidore 2030



Stratégie HNLab

La tortue contre le lièvre

- Veille stratégique constante ;
- Expérimentations tous azimuts ;
- Développement de preuves de concept ;
- Liaison de relations avec des partenaires (académiques et industriels).

A hare and a tortoise in the run for artificial intelligence
© unofficial-hf-plugins/dalle2-image-generation

Voir : Pouyllau, S., Maronet, L., Silvestre de Sacy, A., & Faci, A. (2024). *Note sur l'expérience de l'IA au sein de l'Huma-Num Lab (huma-num version) (1.0)*. Zenodo. https://doi.org/10.5281/zenodo.10846773

Introduction - RAG

- Dans ce contexte, le RAG (*Retrieval Augmented Generation*) est un outil sur lequel nous nous concentrons, avec l'idée que le RAG pourrait être utilisé pour interroger des corpus spécifiques dans un domaine de recherche, ce qui aide à :
  - Contextualiser les requêtes sur des corpus spécifiques ;
  - Lutter contre les hallucinations ;
  - Augmenter la pertinence, la cohérence et surtout l'interprétabilité des générations ;
  - Contrôler les technologies d'IA tout en les spécialisant très fortement sur des problématiques de recherche spécifiques.

Introduction - Problématiques

Axe 1

Quelles sont les voies de recherches ouvertes par l'intégration de technologies d'IA maîtrisées et appliquées à des corpus de recherche en SHS ?

— Exploration et requêtage de corpus.

— Nouvelles vues et entrées sur les données (volumétrie, comparaisons, modélisations des champs).

— Fonctionnalités inédites.

Axe 2

Comment articuler ces méthodes d'IA à des méthodes computationnelles éprouvées depuis de nombreuses années ?

— Le RAG comme technologie en bout de chaîne de traitement, non comme fin en soi.

— Articulation de méthodes classiques de machine learning (classifications) au RAG.

Axe 3

Comment donner la main à l'utilisateur, au chercheur tout au long du processus de recherche ?

— Variation des prismes d'entrée sur les données.

— Contrôle des corpus d'étude.

— Retours aux textes et interprétabilité au coeur du système.

HN Lab

Problématique pour ce travail de recherche

Clustering + RAG pour l’analyse contrastive d’un corpus en histoire de l’art

- Des techniques de RAG proposant différentes pipelines pour maximiser la validité des réponses et la factualité
- Notre position:
  - Comment utiliser le RAG pour interroger la question de style dans un corpus de revues en histoire de l’art ?
  - Comment utiliser les techniques existantes pour donner un ensemble de réponses représentatives d’un corpus ?
  - Comment montrer des contrastes dans un corpus (entre époques, communautés, thèmes) ?

Méthodes

Ne réinventons pas l’eau tiède

- Mise en place d’un RAG simple pour l’instant.
- Expérimentations.
- Approfondissement et améliorations dans un second temps.

Methodologie

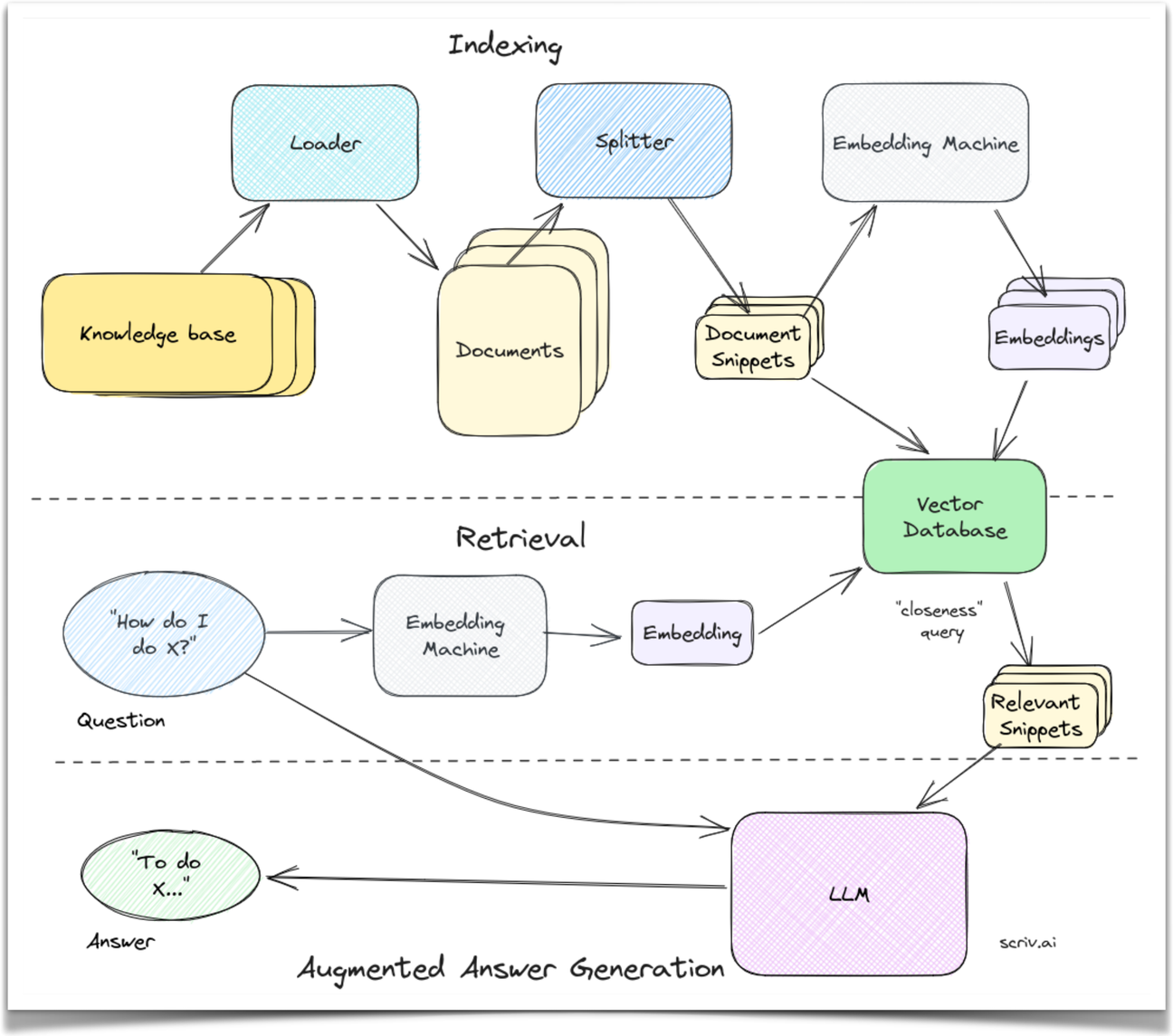
Presentation

- RAG (*Retrieval Augmented Generation*) combine deux modèles d'intelligence artificielle : un modèle de recherche d'informations et un modèle de génération de texte.

- Le modèle de recherche trouve les informations pertinentes dans une base de données, puis le modèle de génération les utilise pour créer un texte original et précis.

- Applications : réponse à des questions, chatbots, rédaction d'articles ou de résumés, traduction automatique…

Schéma détaillé



Retrieval Augmented Generation

Principes généraux :

- ***Indexation :*** Le modèle encode la requête et les documents choisis dans des vecteurs d'intégration par le biais de la projection dans l'espace sémantique.

- ***Récupération :*** Le modèle recherche les documents pertinents (mesure de similarité ou autre) dans une base de données en fonction de la requête de l'utilisateur.

- ***Génération :*** Le modèle utilise la représentation unifiée pour générer une réponse à la requête de l’utilisateur.

Okay,

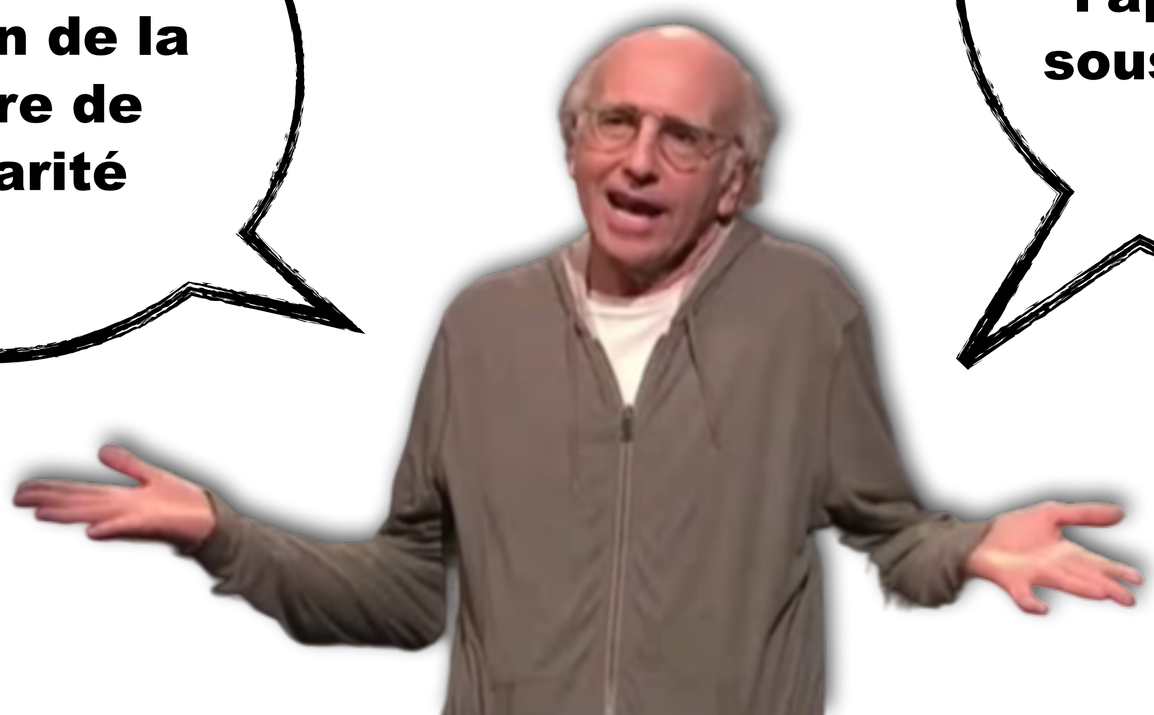
mais sera-t-il pertinent et précis lorsqu'il sera lancé sur

7 000 documents relatifs au même domaine de recherche ?

Eh

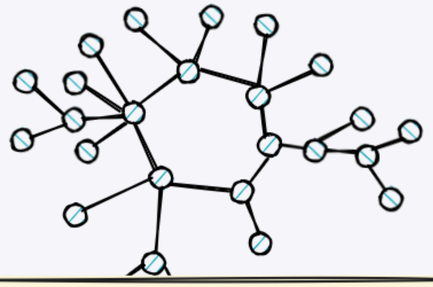
Peut-être, en fonction de la mesure de similarité

Mais ne serait-il pas pertinent de l'appliquer sur des sous-corpus ciblés ?



Pre-targeted-RAG

Retrieval Augmented Generation based on pre-targeted clusters of research papers communities



* Overall info about relationships between individuals in the corpus.
* Overall info on the weight of individuals.
* Overall info on the weight of relationships.

Main corpus

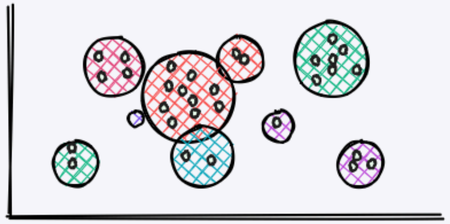
Community detection based on co-authorship

Retrieve connected authors

RAG on pre-targeted subcorpuses

Clusterisation based on texts' semantics

Retrieve TF-IDF score



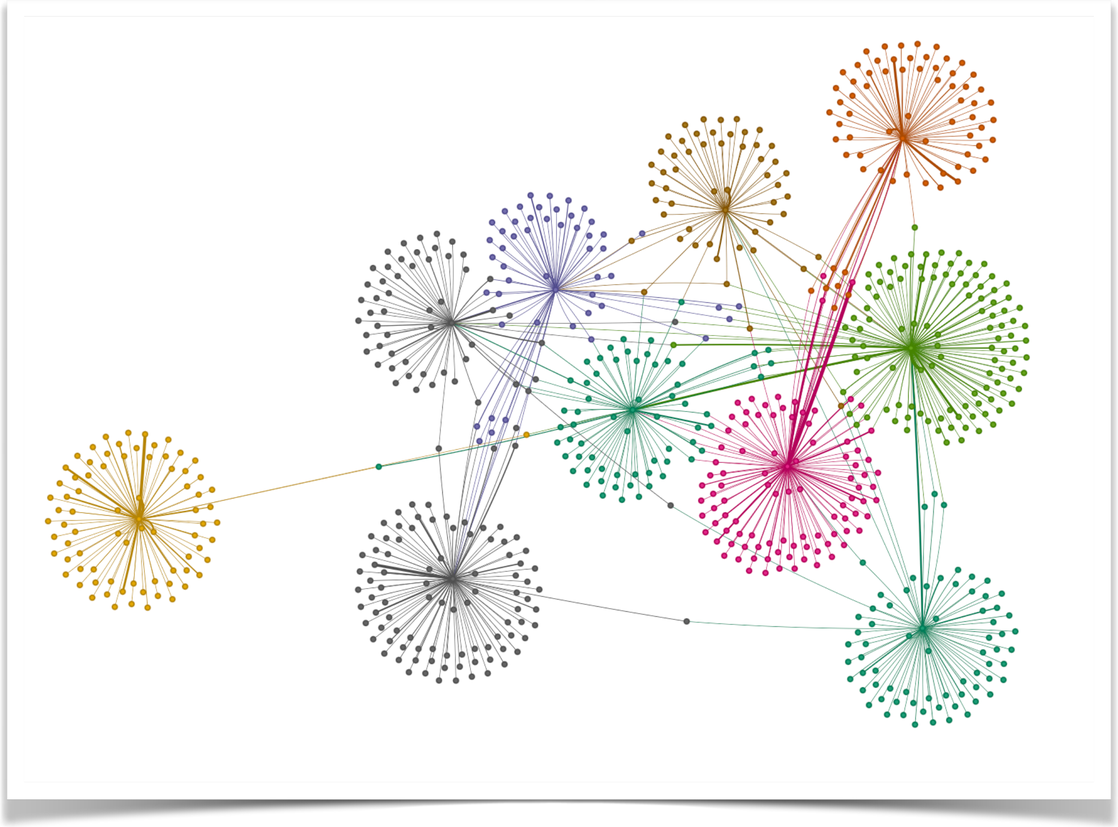
* Semantics clusters of communities
* Give informations about the general topic
* Extracted to be RAGged

Methodologie - Multiplication des prismes d’entrée sur les données

- ***Relations entre auteurs*** *: Comment les auteurs sont-ils liés les uns aux autres au sein d'un domaine de recherche ?*
- ***Relations entre les articles :*** *Comment les articles de recherche sont-ils liés les uns aux autres ?*
- ***Détection de communautés :*** *Peut-on détecter des communautés de recherche ? Comment ?*
- ***Sémantique des articles :*** *Peut-on détecter la sémantique d'une communauté de recherche donnée ?*
- ***Diachronie*** *: Comment analyser un champ de recherche de façon diachronique ?*
- ***Lexie*** *: Comment analyser un terme spécifique au sein d’un domaine de recherche ?*

Prismes d’entrées : relations entre auteurs

Graphs

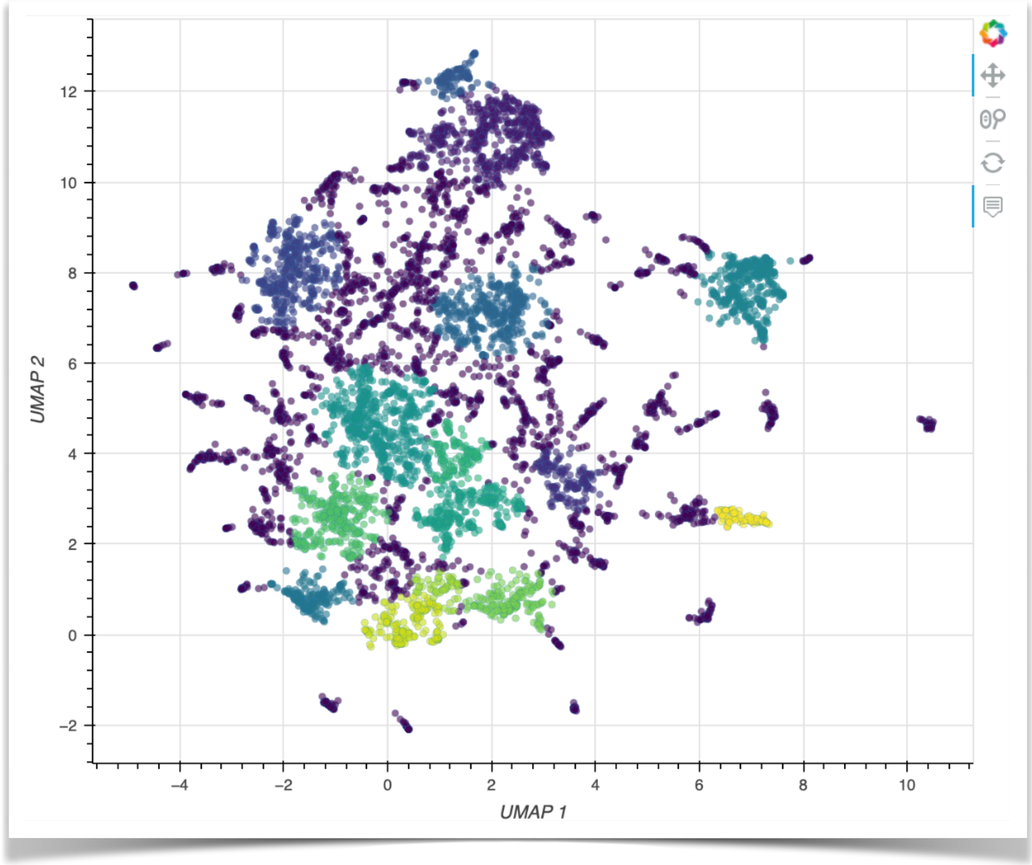


Co-authorship dynamic network on the 10th most connected authors

Louvain’s Method

Prismes d’entrées : par les thèmes du jeu de données

Clustering



Clustered communities of research papers

TF-IDF + UMAP dimensionality reduction + DBSCAN clustering

Sorties, exemples sur un corpus

Créer des parcours utilisateurs

3 profils

Explorateur

Chercheur d’or

Expert

Créer des parcours utilisateurs

Explorateur

- Corpus inconnu, besoin d’une vue globale/contrastive du corpus
- Questions génératrices de question + qui ne portent pas sur le corpus

Créer des parcours utilisateurs

Explorateur

- Corpus inconnu, besoin d’une vue globale/contrastive du corpus
- Questions génératrices de question + qui ne portent pas sur le corpus

Chercheur d’or

- Corpus connu, mais besoin d’analyser les articles connus, les comparer, de jauger l’évolution, de structurer

Créer des parcours utilisateurs

Explorateur

- Corpus inconnu, besoin d’une vue globale/contrastive du corpus
- Questions génératrices de question + qui ne portent pas sur le corpus

Chercheur d’or

- Corpus connu, mais besoin d’analyser les articles connus, les comparer, de jauger l’évolution, de structurer

Expert

- Question très précise sur un corpus, structuration

Typologie de questions

5 catégories

Ciblée

Exogènes

Problématique

Hors sujet

À tiroirs

Typologie de questions

Questions très précises sur le corpus.

Ciblée

- **Question:** What are the positions of **Daniel Buren** towards contemporary **exhibitions** ?

Typologie de questions

Questions très précises sur le corpus.

Ciblée

- **Q:** What are the positions of **Daniel Buren** towards contemporary **exhibitions** ?

- **Contexte:** The artist **Daniel Buren**, for example, complained on the occasion of documenta 5 in **1972** that **exhibitions** no longer tended to **exhibit artworks**, but rather themselves. In **2003** Jens Hoffmann, an **exhibition** organizer, initiated the project “The Next Documenta Should be Curated by an **Artist**”, supported by **artists** such as Marina Abramovic, Tino Sehgal, or Lawrence Weiner. In 2017 the critic, media and **art** scientist Stefan Heidenreich declared the curatorial practice to be undemocratic, authoritarian and corrupt.

HN Lab

Typologie de questions

Questions très précises sur le corpus.

Ciblée

- **Q:** What are the positions of **Daniel Buren** towards contemporary **exhibitions** ?

- **Ctxt:** The artist **Daniel Buren**, for example, complained on the occasion of documenta 5 in **1972** that **exhibitions** no longer tended to **exhibit artworks**, but rather themselves. In **2003** Jens Hoffmann, an **exhibition** organizer, initiated the project “The Next Documenta Should be Curated by an **Artist**”, supported by **artists** such as Marina Abramovic, Tino Sehgal, or Lawrence Weiner. In 2017 the critic, media and **art** scientist Stefan Heidenreich declared the curatorial practice to be undemocratic, authoritarian and corrupt.

- **Réponse:** According to the information provided, **Daniel Buren**’s position towards contemporary exhibitions is that he complained that **exhibitions no longer tend to exhibit artworks, but rather themselves.** This is mentioned in the context of Documenta 5 in 1972.

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus
Cluster + RAG simple

- **Q:** How does the notion of **style evolve** throughout the corpus?

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus
Cluster + RAG simple

- **Q:** How does the notion of **style evolve** throughout the corpus?

- **Ctxt:** Passages about evolution of style or evolution of art

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus
Cluster + RAG simple

- **Q:** How does the notion of **style evolve** throughout the corpus?

- **Ctxt:** Passages about evolution of style or evolution of art

- **R:** The notion of **style evolves** throughout the corpus in several ways:
  […] as mentioned in the quote: « Art itself will never be fully articulated **as a data structure** »
  […] However, as the discussion **progresses**
  […] need for a more inclusive and pluralistic understanding of **style**

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus

Partition temporelle + Cluster + RAG

- **Q:** What is the notion of **style** ? / What is **style** ? / How to define **style** ?

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus

Partition temporelle + Cluster + RAG

- **Q:** What is the notion of **style** ? / What is **style** ? / How to define **style** ?

- **Ctxt:** Style definition different in each partition, with new concepts appearing in recent papers.

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus

Partition temporelle + Cluster + RAG

- **Q:** What is the notion of **style** ? / What is **style** ? / How to define **style** ?

- **Ctxt:** Style definition different in each partition, with new concepts appearing in recent papers.

- **R: Style** as a cultural history of practices […]

2015

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus

Partition temporelle + Cluster + RAG

- **Q:** What is the notion of **style** ? / What is **style** ? / How to define **style** ?

- **Ctxt:** Style definition different in each partition, with new concepts appearing in recent papers.

- **R: Style** as a cultural history of practices […]

2015

- **R: Style** dependent on the multimodal output […] digital art redefined the notion of **style** by […]

2018

HN Lab

Typologie de questions

Exogènes

Questions surplombantes, exogènes sur les thèmes du corpus

Partition temporelle + Cluster + RAG

- **Q:** What is the notion of **style** ? / What is **style** ? / How to define **style** ?

- **Ctxt:** Style definition different in each partition, with new concepts appearing in recent papers.

- **R: Style** as a cultural history of practices […]

2015

- **R: Style** dependent on the multimodal output […] digital art redefined the notion of **style** by […]

2018

- **R: Style** determined by the algorithm […] according to the dataset […]

2023

Typologie de questions

Questions sur la problématique d'un article

Problématique

- **Ex (Titre d'un article):** Can you tell me about curatorial practices in the post-digital age and how online practices call into question those practices ?
- **Ctxt:** Problèmatiques d'articles ou questions dans le corps ou en intro
- **Problèmes potentiels:**
  - L'article ciblé n'est pas retrouvé
  - Passage précédé par une négation
  - Bruit produit par les passages d'autres articles

HN Lab

Typologie de questions

Questions qui ne portent pas sur le corpus

Hors sujet

- **Ex :** Can you tell me about Léa Maronet in Art history ?
- **R:** There is no information about Léa Maronet […] The papers mention various artists such as […]

Typologie de questions

Questions génératrices de questions

À tiroirs

- **Ex :** Au regard du corpus et de ses grands enjeux, quelles grandes questions de recherche penses-tu qu'il soulève ?
- **2 stratégies:**
  - Directement demander au modèle de génération
    - **R:** Réponse correcte (validée par une experte), mais pas nécessairement représentative du corpus
  - Échantillonnage équilibré du corpus à partir des N documents les plus pertinents
    - **Comment:** Récupérer plus de documents, et équilibrer selon différentes modalités (date, auteurs, journaux, thèmes)
    - **R:** Questions plus diverses que dans le premier cas, à valider manuellement par une étude manuelle du corpus

HN Lab

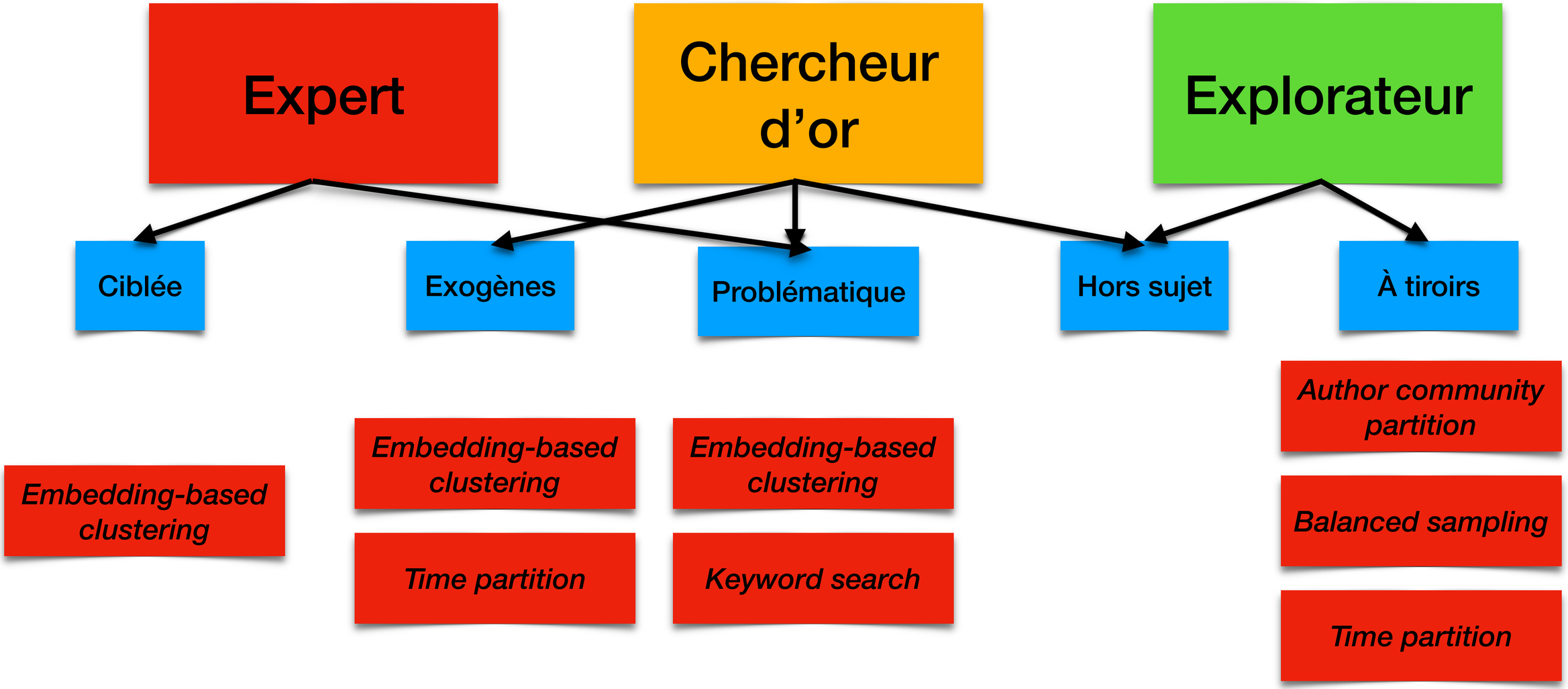
Techniques

- *Clustering* basé sur les plongements de mots (*embeddings*)
- Partition temporelle
- Recherche de mots-clés
- Partition auctoriale
- Échantillonnage équilibré

Association Questions - Techniques

Ciblée	Exogènes	Problématique	Hors sujet	À tiroirs
Embedding-based clustering	*Embedding-based clustering*	*Embedding-based clustering*		*Author community partition*
	Time partition	*Keyword search*		*Balanced sampling*
				Time partition

Association profils - Techniques



Conclusion

Suite des travaux

Conclusion

Applications futures

- Intégration du RAG dans un micro-service
- Expérimentation sur l'impact du *fine-tuning* sur le domaine du corpus exploré
- Permettre à l'utilisateur d'avoir la main et de naviguer, à la manière de *ResearchRabbit*
- Design UI/UX pour une intégration dans ISIDORE 2030

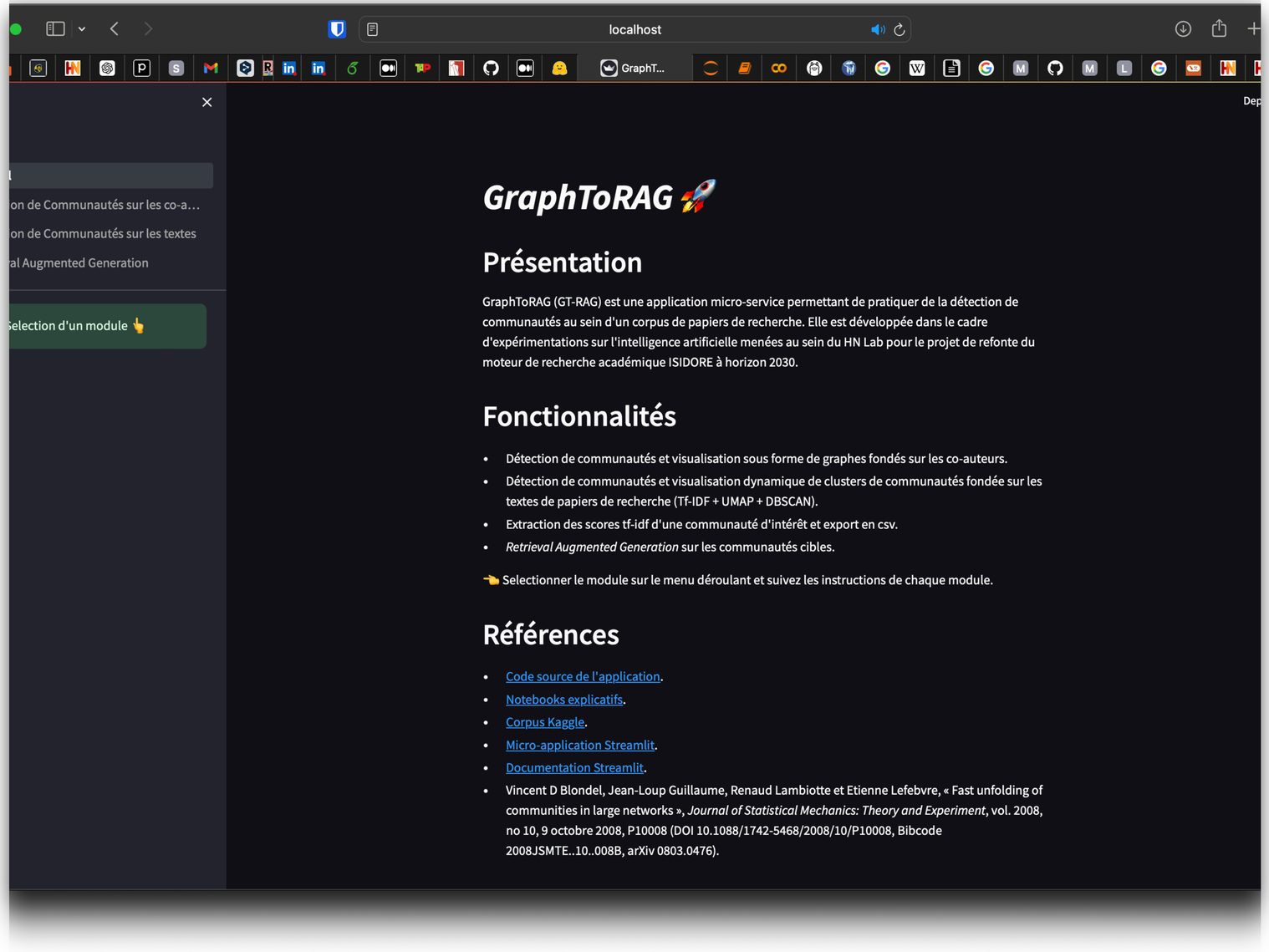
Opportunités

- Entrainement de modèle pour recommendation de stratégies en fonction de la question
- Comparaison des différentes stratégies de *clustering*
- Techniques existantes pertinentes pour notre cas:
  - Résumé hiérarchique (cf RAPTOR) / Fusion hiérarchique (cf Akesson CRAG)
    -> Réponse à différents niveaux
  - RAG à têtes multiples/spéculatif -> Différents aspects de la question
  - Chain-of-Thought / Recursive / Yan CRAG -> Explication de la réponse
  - Utilisation d'une base externe d'entités nommées -> qualité de l'article, auteur, etc.

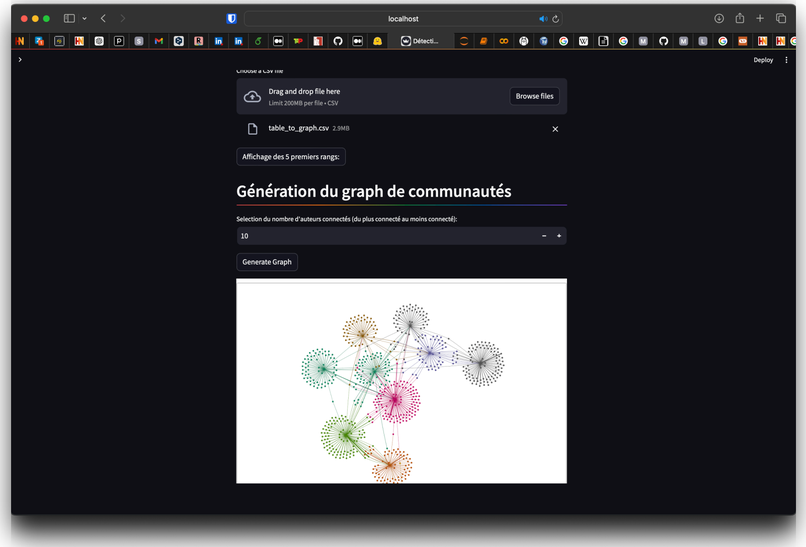
Suite de travail de recherche

- Plan d’expérience avec les techniques de l’état de l’art
- Interface Homme-Machine et test sur une communauté de chercheurs
- Généralisation à d’autres domaines (ex: style en littérature)
- Article en écriture pour une revue HN
- Article à prévoir en SI (IA appliquée / nouvelle chaine RAG)

Intégration micro-service



GraphToRAG micro-application using Streamlit framework



Community network of co-authors

Papers' clustering

Merci pour votre attention !