

ISIDORE 2030

ISIDORE 2030 est un programme de recherche et d'ingénierie en sciences humaines et sociales (SHS) visant à renouveler les fonctionnalités du moteur et assistant de recherche académique isidore.science. Lancé en 2010, isidore.science, comme les autres outils de découverte, est en évolution permanente. La généralisation des intelligences artificielles (IA) génératives et sélectives impacte fortement les instruments de recherche, qui doivent ainsi évoluer tout en restant fonctionnels pour les communautés qu'ils desservent. Après 13 ans d'indexation et d'enrichissement documentaire, isidore.science a vieilli et se trouve à la croisée des chemins : comment le renouveler et vers quoi l'orienter tout en continuant à l'exploiter ? L'intervention tracera les principaux chantiers d'ISIDORE 2030 tout en revenant sur les étapes franchies ces dernières années.

L'IA dans ISIDORE : de la classification au *Retrieval Augmented Generation*

ISIDORE utilise depuis 2010 de l'IA, principalement pour classer, catégoriser et enrichir des données et des métadonnées. Développé avec la société Antidot SA, les différentes IA ont évolué et ont permis à l'équipe d'Huma-Num d'acquérir une solide expérience dans le domaine des IA pour les SHS (Silvestre de Sacy et al., 2024).

Depuis 2023, au sein du HN Lab d'Huma-Num, plusieurs travaux ont été initiés pour inclure dans ISIDORE le domaine des IA génératives pondérées. Il s'agit de travailler sur la conceptualisation et la mise en œuvre de la refonte du moteur de recherche à horizon 2030. Cette refonte vise à améliorer les fonctionnalités actuelles du moteur tout en intégrant les dernières technologies offertes par l'arrivée et la démocratisation des grands modèles de langue (Pouyllau et al., 2024).

Dans ce contexte, l'équipe du HN Lab travaille depuis plusieurs mois sur une fonctionnalité précise, le RAG (pour *Retrieval Augmented Generation*). Le RAG est une technique innovante qui combine les capacités de génération de langage naturel et de recherche d'information pour améliorer les performances et l'interprétabilité des grands modèles de langage. Cette technique tente de minimiser leurs aspects boîte noire en s'appuyant sur un corpus spécifique fourni par l'utilisateur pour générer ses réponses (Silvestre de Sacy et al., 2024).

Lorsqu'une requête est soumise, le système récupère des informations pertinentes depuis la base de données fournie par l'utilisateur, les combine avec la requête de l'utilisateur et passe ensuite la requête augmentée au GML (Grands Modèles de Langue) pour générer une réponse fondée sur le contexte.

Cette approche vise à minimiser des problèmes tels que les hallucinations, les connaissances obsolètes et les limitations de raisonnement dans les GML. En intégrant des sources de connaissances externes, le RAG améliore la qualité globale, l'exactitude des réponses des GML et fournit du contexte interprétatif à la réponse tout en maintenant à jour un corpus de données qui peut être enrichi (Maronet & Truc, 2024).

Fonctionnalités attendues

- Analyse des contenus et rapprochement de corpus ;
- Tableaux de bord permettant de créer des propositions d'états de l'art sur des questions scientifiques ;
- Proposition de résumés automatiques et synthèses de documents ;

- Traduction et analyses de qualité de traduction (langues rares, anciennes, etc.) ;
- Exploration de communautés scientifiques ;
- Détection de thématiques et de communautés émergentes ;
- Veille scientifique ;
- Détection de thématiques de longue traîne ;
- Amélioration qualitative de métadonnées ;
- ...

Bibliographie

Silvestre de Sacy, A., Faci, A., Pouyllau, S., & Maronet, L. (2024, octobre 18). Pre-targeted-RAG - Retrieval Augmented Generation sur des groupes pré-ciblés de communautés d'articles de recherche. ColDoc, Université Paris-Nanterre. Zenodo. <https://doi.org/10.5281/zenodo.13950650>

Pouyllau, S. (2024, octobre 4). ISIDORE 2030 : adapter les IA aux besoins de la recherche de documents et de données en SHS. Conférence au GF2i (GF2i), PARIS. HN LAB. <https://doi.org/10.5281/zenodo.13892964>

Pouyllau, S., Silvestre de Sacy, A., Maronet, L., & Faci, A. (2024). Capitalizing on experience to experiment and innovate: feedback and reflection on the future of the Huma-Num research infrastructure. (1.0). Digital Humanities in the Nordic and Baltic Countries (DHNB), Reykjavik. HN Lab. <https://doi.org/10.5281/zenodo.13889742>

Maronet, L., & Truc, A. (2024). Improving workflows in digital art history: sharing annotations for patrimonial images segmentation and object detection. Transformations, A DARIAH Journal, 1. <https://doi.org/10.5281/zenodo.13947909>

Maronet, L., & Truc, A. (2024, juin 16). Improving workflows in digital art history: the usefulness of patrimonial images segmentation. Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities, Lisbon, Portugal. Zenodo. <https://doi.org/10.5281/zenodo.11863661>

Silvestre de Sacy, A., Faci, A., Maronet, L., & Pouyllau, S. (2024). Note sur l'expérience de l'IA au sein de l'Huma-Num Lab (huma-num version) (1.1). ACFAS 2024 (ACFAS), Ottawa. HN Lab. <https://doi.org/10.5281/zenodo.10846773>

Références supplémentaires

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Lewis, P., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Stoyanov, V. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Ces références supplémentaires fournissent un contexte théorique et pratique pour les techniques de génération de langage naturel et de recherche d'information, ainsi que pour les modèles de langage utilisés dans le cadre du projet ISIDORE 2030.