DOSSIER DE PRESSE

La BnL, pionnière de l'innovation numérique : l'intelligence artificielle au service du patrimoine imprimé luxembourgeois

Bibliothèque nationale
du Luxembourg

DATE :
Mercredi 04.10.2023 à 15 heures

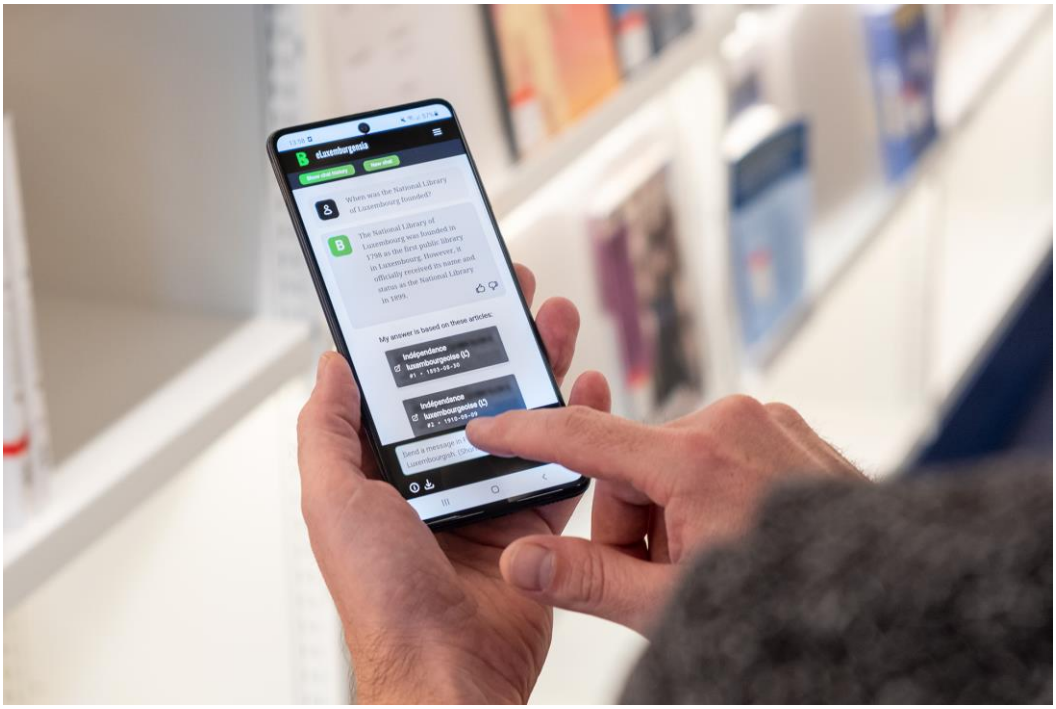
LIEU :
Bibliothèque nationale du Luxembourg
Amphithéâtre 3e étage
37D, avenue John F. Kennedy
L-1855 Luxembourg

Sam Tanson, ministre de la Culture, Claude D. Conter, directeur de la BnL, et Carlo Blum, directeur adjoint de la BnL, ont invité les membres de la presse et des médias à découvrir, en avant-première, le nouveau chatbot du portail eluxemburgensia.lu. Ce chatbot est un agent conversationnel à intelligence artificielle qui comprend le français, l'allemand et l'anglais et assiste les internautes dans l'exploration de l'histoire luxembourgeoise en proposant des réponses argumentées, basées sur des articles de presse historiques.

Sujets élucidés :

1. Le chatbot en pratique par Yves Maurer, responsable de la division Informatique et Innovation numérique

2. L'intelligence artificielle à BnL par Carlo Blum, directeur adjoint



1. LE CHATBOT EN PRATIQUE

Le nouveau chatbot du portail eluxemburgensia.lu comprend le français, l'allemand et l'anglais, il assiste les internautes dans l'exploration de l'histoire luxembourgeoise et propose des réponses argumentées en se basant sur des articles de presse historiques.

C'est grâce à une technologie en service chez ChatGPT, l'agent conversationnel à intelligence artificielle développé par OpenAI, que les experts de la BnL ont indexé les documents luxembourgeois numérisés et préparé une base de données performante, qui permet de réaliser des recherches sémantiques. Cette avancée marque un jalon majeur dans la mission de la BnL d'offrir un accès facilité et enrichi à ses ressources luxembourgeoises numérisées.

Notons que le chatbot est un outil expérimental gratuit et accessible à distance. Pour l'utiliser, il suffit de s'authentifier avec sa carte de lecteur ou un compte Google.

Premiers pas

Commencez simplement par poser une question en utilisant une ou plusieurs phrases complètes. Dans un premier temps, le chatbot identifie la langue de votre texte. Ensuite, il recherche des réponses potentielles parmi les huit millions d'articles des journaux numérisés sur eluxemburgensia.lu. Au cours de ce processus, il essaie de trouver des articles pertinents en allemand et en français, quelle que soit la langue dans laquelle vous avez rédigé votre question. Enfin, le chatbot se connecte à l'application ChatGPT. Cette dernière est un générateur de textes alimentée par l'intelligence artificielle. Le chatbot lui demande d'analyser ces articles et de proposer une réponse argumentée. Les articles sur lesquels le chatbot se base apparaissent en dessous de la réponse.

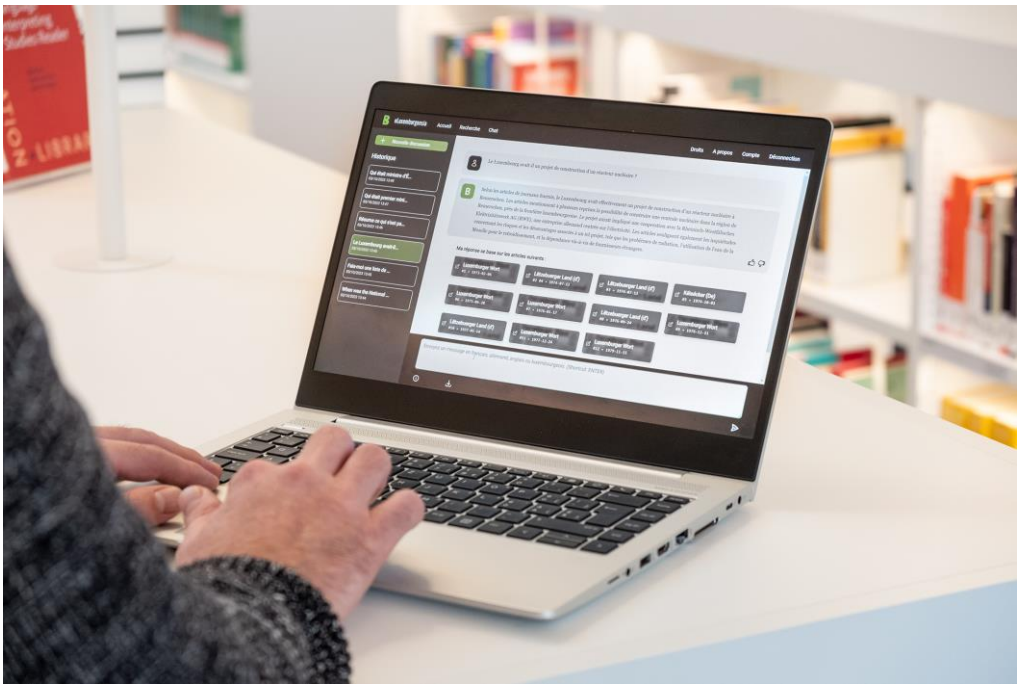
Vous pouvez poser des questions complémentaires, qui utilisent l'historique du dialogue en cours comme contexte.

En quoi le chatbot diffère-t-il de la barre de recherche ?

L'interaction avec le chatbot ne remplace pas la recherche classique par mot-clé ; il s'agit plutôt d'un service complémentaire et expérimental, qui répond à votre requête en se basant sur un nombre prédéfini d'articles. Le chatbot n'est donc pas aussi exhaustif qu'une recherche classique. Il sélectionne les articles en fonction de leur sujet, et non pas selon des mots-clés ou des orthographes spécifiques. Il génère une réponse structurée sur base d'une sélection d'articles et vous évite ainsi de faire un premier tri et facilite la suite de votre recherche. En revanche, la barre de recherche vous permet de trouver les mots-clés souhaités dans l'ensemble de la collection d'articles. Cela peut par exemple s'avérer important si vous recherchez toutes les mentions de certaines entités, telles que des personnes ou des lieux. Elle vous permet également de rechercher une orthographe spécifique, une combinaison exacte de mots ou de dates.

Quelles sont les limites du chatbot pour la recherche ?

Les questions qui ne peuvent pas être résolues en lisant simplement quelques articles comme le ferait un humain sont également hors de portée pour le chatbot. Par exemple, il ne lui serait pas possible d'énumérer toutes les dates d'inauguration de la fête foraine annuelle *Schueberfouer*, à moins qu'un seul article ne contienne la réponse exacte.



Pour plus d'informations sur le fonctionnement et l'utilisation du chatbot, veuillez consulter l'aide en ligne : chat.eluxemburgensia.lu/info

2. L'INTELLIGENCE ARTIFICIELLE A LA BNL

La Bibliothèque nationale du Luxembourg, toujours soucieuse de proposer de meilleurs services à ses usagers, a lancé fin 2019 un ensemble d'initiatives et d'expérimentations destinées à optimiser l'exploration visuelle de nos fonds numériques et numérisés sur le site eluxemburgensia.lu. Pour réaliser ceci, nous avons misé sur les technologies de l'intelligence artificielle (IA) développées par des experts au sein de notre division informatique et de l'innovation numérique.

L'utilisation de techniques d'intelligence artificielle permet une analyse plus rapide et plus approfondie des collections. Elle facilite l'indexation et la classification des documents tout comme la recherche et l'accès à nos ressources numérisées.

– Carlo Blum, directeur adjoint

Exemples de développement réalisé et en voie de réalisation

Amélioration de la qualité de l'OCR

La numérisation des journaux historiques avec reconnaissance optique de caractères (OCR) a commencé à la BnL en 2006. Souvent la mauvaise qualité du support papier, des imperfections de l'impression et la dégradation des originaux due à l'usure du temps font que l'OCR n'a pas pu identifier correctement toutes les lettres des originaux. Les lettres qui n'ont pas pu être identifiées correctement font que beaucoup de mots présents sur les supports papier ne peuvent pas être trouvés par le moteur de recherche de eluxemburgensia.lu.

Le logiciel Nautilus-OCR améliore la reconnaissance de textes (OCR). Il propose plusieurs modules utilisant des techniques d'apprentissage automatique et permet de détecter automatiquement les lignes de texte, de classer des polices de caractères, d'estimer l'amélioration potentielle de la qualité, et de reconnaître des caractères individuels.

BEFORE NAUTILUS-OCR

1) SScrgebenS wirb gemeine Klugheit ung rathen, ung
2) felbft 3« beberrfd;en unb rubigju bleiben, in bitten
3) biefer îâ'ufdjungen, wk ter 2Beife, »on welchem
4) Horaj fpriebt; tiejer Sîatfi hilft nid)ts. D6ne ein
5) antereS ScbeitalStaS materielle, »ermögen wir nid)t
6) im kennen einzuhalten; eS ift nidjt tte Brte, wag
7) ung diul)t auf Grten gewährt. Sag reltgi&fe Cc«
8) ben allein »rbeipt uns ©efül;le, tie fähig ftnb, uns

WITH NAUTILUS-OCR

1) Vergebens wird gemeine Klugheit uns rathen, uns
2) selbst zu beherrschen und ruhig zu bleiben, in Mitten
3) dieser Täuschungen, wie der Weise, von welchem
4) Horaz spricht: dieser Rath hilft nichts. Ohne ein
5) anderes Leben als das materielle, vermögen wir nicht
6) im Rennen einzuhalten; es ist nscht die Erde, was
7) uns Ruhe auf Erden gewährt. Das religiöse Le-
8) ben gllein verheißt uns Gefühle, die fähig sind, un

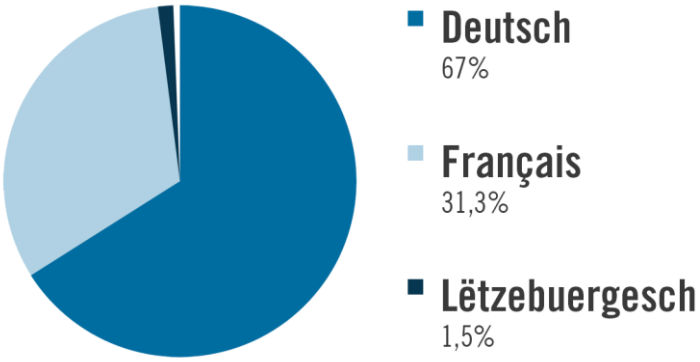
Un aspect important de Nautilus-OCR est qu'il travaille avec le format de données METS/ALTO qui est largement utilisé dans le domaine des bibliothèques pour les documents numérisés. Il prend comme entrée un jeu de données en METS/ALTO, améliore la qualité de l'OCR et produit un jeu de données METS/ALTO amélioré. Ainsi l'application de l'outil dans d'autres bibliothèques et institutions qui disposent d'un programme de numérisation est facilitée.

La BnL a appliqué Nautilus-OCR sur l'entièreté de la collection de journaux historiques, numérisée entre 2006 et 2021. Ceci a entraîné une amélioration de 28 millions de lignes de texte.

Recherche par langue

Les journaux, livres et revues historiques numérisées par la BnL témoignent de la culture de multilinguisme présente au Luxembourg et rassemblent souvent différentes langues sur une même page. En avril 2023, le portail eluxemburgensia.lu a été d'une nouvelle option de filtrage par langue. Celle-ci permet à un utilisateur, qui préfère lire des articles dans une langue spécifique, de filtrer l'ensemble des contenus et de cibler d'avantage sa recherche.

Les langues principales sur le portail sont l'allemand, qui compte pour 67% du total, suivi du français à hauteur de 31,3%. Le luxembourgeois est représenté dans une proportion nettement inférieure, avec un peu plus de 118.123 articles (soit 1,5%), tandis que l'anglais, avec 10.149 articles (soit 0,1%), est considéré comme minoritaire. 14 autres langues ont été détectées dont le latin, l'italien, le portugais et le polonais.



Détection d'images similaires

Pour faciliter la recherche d'illustrations, de photos, de portraits ou de caricatures, une preuve de concept a été élaborée pour classer des images par genre et détecter celles avec forte ressemblance. En utilisant un ensemble de 600 000 illustrations tirées de journaux numérisés de 1844 à 2007, les diverses propriétés de ces images ont

été extraites par des techniques d'IA. Ces propriétés ont ensuite servi à repérer des similitudes préexistantes. Une option de dédoublage a aussi été implémentée pour éliminer des illustrations identiques.

En outre, l'IA est également utilisée pour déterminer le contenu des images. Ce développement simplifie l'exploration des illustrations de nos collections numérisées en ajoutant des mots-clés aux images.

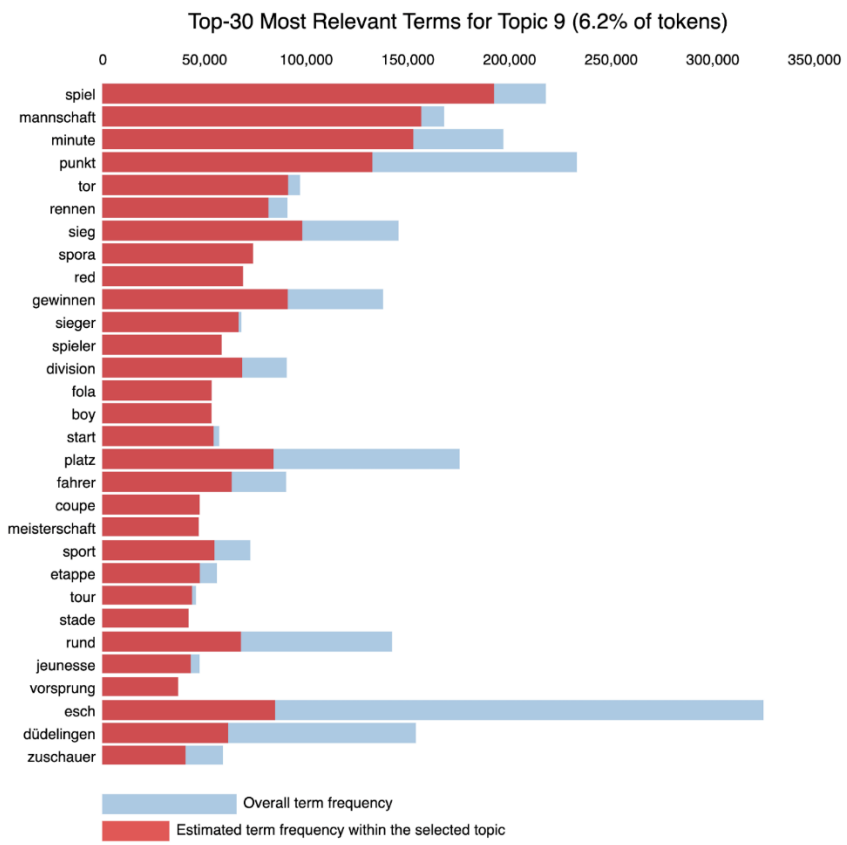
Exemple de douze images contenant au moins un parapluie comme objet.

Des modèles d'apprentissage automatique ont été utilisés pour déterminer pour chaque illustration son type (photo, dessin, etc.) et son contenu général (portrait, groupe, architecture, carte, caricature, musique, etc.). L'IA permet également d'identifier les objets présents ainsi que les sujets décrivant l'illustration.

Le développement et le raffinement de ces fonctions reposant sur l'IA sont toujours en cours, mais une première mise en œuvre sur eluxemburgensia.lu, accessible au public, est prévue au cours de l'année prochaine.

Identification des sujets d'articles de journaux

Une autre initiative, rendue possible grâce au projet Nautilus-OCR (amélioration de la qualité des textes reconnus par reconnaissance automatique de caractères), vise à identifier les sujets d'articles de journaux pour avoir une meilleure vue d'ensemble sur la collection et pour proposer aux utilisateurs une recherche par sujet. Il s'agit d'une phase préliminaire pour classifier les articles et pour procéder à l'entraînement des modèles d'apprentissage automatique, spécifiques à la BnL. Un modèle statistique (LDA, allocation de Dirichlet latente) a été appliqué sur 3.5 millions d'articles en langue allemande pour regrouper des contenus similaires. Des expérimentations ont été faites avec un nombre de groupes allant de 5 (thèmes plus généraux) à 100 (thèmes plus spécifiques). Ces groupes peuvent être visualisés à travers les termes les plus pertinents. Tous les groupes ont été analysés et les plus adéquats ont été retenus pour servir de base à une classification par sujets.


Top-30 Most Relevant Terms for Topic 9 (6.2% of tokens)
0
50,000
100,000
150,000
200,000
250,000
300,000
350,000
spiel
mannschaft
minute
punkt
tor
rennen
sieg
spora
red
gewinnen
sieger
spieler
division
fola
boy
start
platz
fahrer
coupe
meisterschaft
sport
etappe
tour
stade
rund
jeunesse
vorsprung
esch
düdelingen
zuschauer
Overall term frequency
Estimated term frequency within the selected topic