Le système RICS™

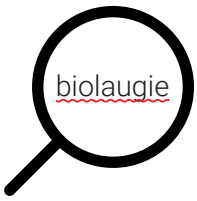
Recherche et indexation d'un corpus scientifique

RICS™, en quelques chiffres :

326 documents indexés

+13K termes indexés

18 rubriques indexés



Tolérance aux fautes

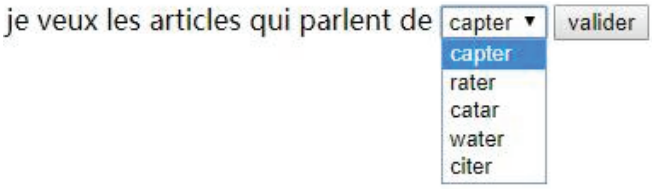
Plus jamais de fautes !

Correction des fautes en amont de la recherche

biolaugie ⟶ *biologie*

Suggestion de mots similaires en cas d'ambiguïté

je veux les articles qui parlent de carter

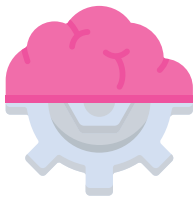


Parlez à RICS™ naturellement :

Quels articles parlent de biologie et de sport ?

Combien d'auteurs ont écrits sur les nanotechnologies ?

Quels sont les bulletins qui traitent de climat en juin 2016 ?



Reconnaissance du langage naturel

Recherche multi-critères

Filtrez selon :

Auteur

Date de parution

Identifiant

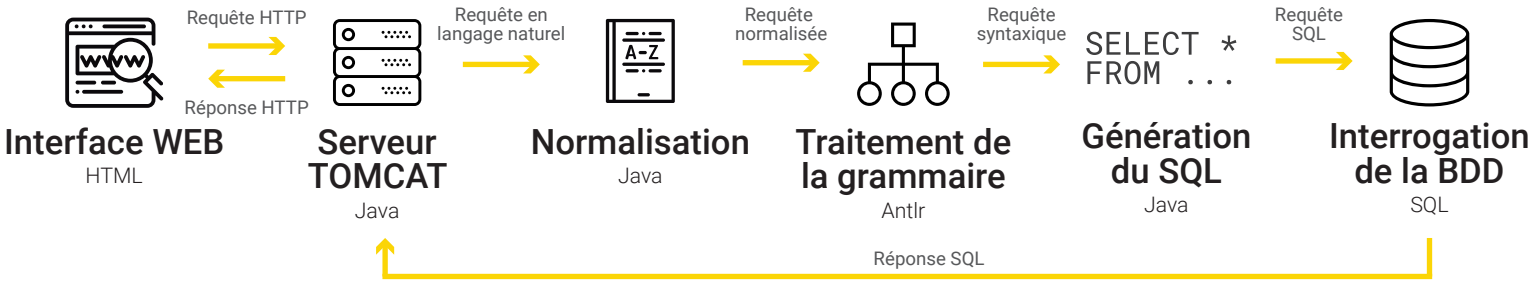
Titre

Rubrique

Contenu

RICS™, son fonctionnement technique

Schéma de fonctionnement



En détail

Normalisation

Application d'une correction orthographique, si nécessaire, par :
- la **distance de Levenshtein**
- une **recherche par préfixe**

Puis, **lemmatisation**, basé sur le lexique

Création du lexique

Calcul du $\mathbf{tf \times idf}$ sur le corpus afin de créer une **stop-list**.

Puis utilisation de l'**algorithme de troncation** pour créer les lemmes pour notre table inverse

Reconnaissance de la grammaire

Construction d'une grammaire grâce à ANTLRWorks. L'**arbre syntaxique** generé par la grammaire représente la requête en langage naturel.

L'arbre va ensuite être **convertit en SQL.**

Nous appliquons un **post-traitement sur le SQL** pour certains cas particuliers.

Serveur et inteface WEB

Le serveur est basé sur **TOMCAT** et permet de renvoyer les requêtes de l'utilisateur vers le système de normalisation.

Limites du système et améliorations

Conclusion

Système de recherche fonctionnel permettant de traiter des requêtes en langage naturel de différentes formes et tolérant aux fautes.

Limites

Certains mots non présents dans le corpus sont remplacés à tort.

Il n'est pas possible de rechercher sur une période (avant/après le X)

Améliorations

➔ Retourner plusieurs colonnes

➔ Améliorer la recherche sur les dates (différents formats)

➔ Améliorer la correction orthographique
→ Proximité des caractères du clavier

➔ Reconnaitre davantage de types de requêtes

➔ Classer les résultats par pertinence

➔ Ajuster les paramètres du correcteur (avec un algorithme de machine learning par exemple)