

# Rapport de projet INF 301

14/12/2023

Laura ARLES, Marwa BENAMOR, Lou CLAEYSSSEN FABRIS

## Introduction

Le projet a pour but de développer un moteur de recherche intuitif spécialisé dans l'exploration d'un corpus d'abstracts de littérature scientifique.

## Scénario

Ce moteur de recherche doit permettre aux professionnels de santé de réaliser une recherche rapide dans un corpus d'articles scientifiques. Celle-ci doit retourner des articles pertinents à l'aide notamment d'une expansion de requête, et proposer des articles similaires afin d'approfondir le sujet.

Un scénario d'utilisation serait la recherche par un hématologue d'informations sur le traitement d'un type d'anémie : il peut requêter un corpus de résumés d'articles scientifiques portant sur l'anémie qui a été indexé selon 3 niveaux (mots simples, termes et concepts) afin de rechercher les articles les plus pertinents.

## Structure

Le projet prend la forme d'une application Java Springboot qui expose des API Rest. Suivant la structure d'un projet Java Springboot, il comprend différents packages décrits ci-dessous :

- un package **controller** qui comprend deux Controller (CorpusController et SearcherController) avec leurs API respectives, reliés par une configuration CORS
- un package **service** comprenant les logiques de code de la création de corpus et de l'indexation d'une part dans CorpusService, et de la recherche dans les index d'autre part dans SearcherService. Ces services appellent des méthodes situées dans des packages spécifiques décrit ci-après.
- un package **domain** comprenant un DocumentDomain qui représente un résultat PubMed comprenant un identifiant, un titre, un résumé et un lien URL

Les autres packages comprennent des méthodes propres aux différentes fonctionnalités :

- un package **abstractExtractor** qui comprend les méthodes qui vont requêter la base PubMed et extrait les résultats pour les écrire dans un fichier en format texte avec les informations suivantes : identifiant de l'article, titre, résumé et lien
- un package **indexerSearcher** qui comprend les méthodes d'indexation et les méthodes de recherche pour chaque niveau (mot simple, terme et concept), ainsi qu'une méthode créant un CharArraySet à partir d'un fichier de "stopwords" en anglais qui est utilisé par le StandardAnalyzer

- un package **ontologyLoader** qui comprend les méthodes pour charger une ontologie existante en local, récupérer l'URI d'un concept à partir de son label, et récupérer les termes liés à un terme

Enfin, la racine des packages du projet comprend la classe exécutable **Application** et les paramètres configurables dans la classe **Parameters**.

## Applications

Les API Rest exposées sont les suivantes :

### CorpusController

#### **generate** : création d'un corpus d'abstracts à partir d'une requête PubMed

La fonction **generate** permet d'extraire les résultats de la page HTML de résultats PubMed en fonction d'une requête spécifique.

Les paramètres d'entrée incluent le terme de recherche (**term**), la plage de dates (**dateRange**), et le nombre de résultats souhaité (**size**).

GET /api/corpus/generate Générer le corpus via PubMed

**Parameters** Cancel

Name	Description
<b>term</b> * required string (query)	Terme de recherche
<b>dateRange</b> * required string (query)	Plage de dates
<b>size</b> * required string (query)	Taille du corpus

anemia

2020-2023

100

Execute Clear

Les résultats sont ensuite stockés dans des fichiers texte (.txt) pour former le corpus d'abstracts (cf. capture d'écran ci-dessous).

Nom	Taille	Dernière modification
A Case of Fetal Familial Hemophagocytic Lymphohistiocytosis Type 5 caused by STXBP2 Gene Mutation.	1,2 ko	14:51 ☆
A Case of Lynch Syndrome-Associated Colorectal Adenocarcinoma in a 19-Year-Old Female Patient.	1,5 ko	14:51 ☆
A Case of Porto-Sinusoidal Vascular Disease.	2,3 ko	14:52 ☆
A Narrative Review of Maternal and Perinatal Outcomes of Dengue in Pregnancy.	1,5 ko	14:51 ☆
A Narrative Review on Effects of Maternal Bariatric Surgery on Offspring.	2,2 ko	14:52 ☆
Anemia burden in pregnancy and birth outcomes among women receiving antenatal care services from a secondary level hospital in South India: A record review.	2,7 ko	14:52 ☆
Anemia in liver transplant recipients: prevalence, severity, risk factors, and survival.	1,4 ko	14:52 ☆
A Rare Heterozygote with a Novel IVS-II-786 (T>A) Mutation on β-Globin Gene in a Patient with Thalassemia.	1,5 ko	14:51 ☆

## index-corpus : indexation d'un corpus de documents

La fonction **index-corpus** permet d'indexer le corpus créé précédemment selon trois niveaux : par mots simples, par termes, et par concepts.

Le paramètre d'entrée correspond au niveau d'indexation souhaité : 1 pour l'indexation par mots simples, 2 pour l'indexation par termes et 3 pour l'indexation par concepts.

POST /api/corpus/index-corpus Indexation du corpus

Parameters

Name	Description
<b>indexLevel</b> * required integer (\$int32) (query)	Niveau d'indexation (1 pour SIMPLE_WORDS, 2 pour PHRASE, 3 pour CONCEPTS)

1

Execute Clear

Ci-dessous un exemple d'index créé suite à une indexation par termes.

Nom	Taille	Dernière modification
_0.cfe	447 octets	14:56 ☆
_0.cfs	86,2 ko	14:56 ☆
_0.si	344 octets	14:56 ☆
segments_1	154 octets	14:56 ☆
write.lock	0 octet	14:56 ☆

## SearcherController

### simple : recherche par mots simples dans le corpus de documents

La fonction **simple** réalise une recherche par mots simples dans le corpus de documents. Elle prend pour paramètre d'entrée un mot simple.

GET /api/simple

Parameters

Name	Description
<b>term</b> * required string (query)	

anemia

Execute Clear

Elle retourne le nombre total de résultats (**totalHits**) ainsi qu'une liste d'informations sur les documents correspondants, comprenant l'identifiant (**docId**), le titre (**docTitle**), le résumé (**docContent**), et le lien URL sur PubMed (**docLink**) : voir l'exemple ci-dessous.

**Responses**

**Curl**

```
curl -X 'GET' \
  'http://localhost:8083/api/simple?term=anemia' \
  -H 'accept: */*'
```

**Request URL**

```
http://localhost:8083/api/simple?term=anemia
```

**Server response**

**Code** **Details**

200

**Response body**

```
{
  "totalHits": 10,
  "documents": [
    {
      "docId": 10,
      "docTitle": "Clinical characteristics of hepatitis-associated aplastic anemia in children.",
      "docContent": "To understand the current situation of hepatitis-related aplastic anemia (HAAA) in children, we analyzed the patients with HAAA admitted to our hospital in the past 5 years to understand the disease characteristics and prognosis. The clinical data of patients with HAAA admitted to our hospital from February 2017 to May 2022 were retrospectively analyzed. A total of 81 patients with HAAA, 56 males and 25 females. The median onset age was 5.9 years. The median time from hepatitis to occurrence of hemocytopenia was 30 days, and the median follow-up time was 2.77 years. There were 23 cases (28.5%) of severe aplastic anemia (SAA), 50 cases of very severe aplastic anemia (VSAA), and 3 cases of non-severe aplastic anemia (NSAA). At the beginning of the disease, cytotoxic T lymphocyte (CTL) was higher than normal in 60% of patients, and the median CD4/CD8 ratio was 0.2. As of follow-up, 72 children survived, 4 were lost, and 5 died. Thirty-four cases were treated with immunosuppressive therapy (IST), with a median follow-up time of 0.97 years. The total reaction rate was 73.5% (25/34), the complete reaction rate was 67.6% (23/34), and the nonreaction rate was 26.5% (9/34). Multivariate analysis suggested that co-infection was an independent risk factor affecting the efficacy of IST at 6 months, with an OR value of 16.76, 95% CI (1.23, 227.95), P=0.034. No independent influencing factors were found at the end of follow-up. The proportion of CTL cells in peripheral blood of children with HAAA is relatively increased, and IST is effective in 73.5% of children. Co-infection may prolongs the time to response to IST.",
      "docLink": "https://pubmed.ncbi.nlm.nih.gov/38082101/"
    },
    {
      "docId": 20,
      "docTitle": "Effect of Stanazolol combined with Cyclosporine A on aplastic anemia.",
      "docContent": "Objective: To investigate the clinical efficacy of Stanazolol combined with Cyclosporine A for treatment of aplastic anemia and its influence on cytokine levels. Methods: This is a retrospective analysis of 90 patients with aplastic anemia treated in Department of Hematology, Shandong Provincial Third Hospital from July 2019 to July 2022. According to the different treatment methods, these patients were assigned into a control group and an observation group, with 45 cases in each group. Patients in the control group were treated with Stanazolol alone, while those in the observation group were treated with the combination of Stanazolol and Cyclosporine A. Patients in the observation group were treated for six months continuously. The indicators in terms of therapeutic effect, drug onset time, cytokine levels, quality of life, and adverse reactions were recorded."
    }
  ]
}
```

**Response headers**

```
connection: keep-alive
content-type: application/json
date: Thu, 14 Dec 2023 12:56:59 GMT
keep-alive: timeout=60
transfer-encoding: chunked
vary: Origin,Access-Control-Request-Method,Access-Control-Request-Headers
```

## **phrase** : recherche par termes dans le corpus de documents

La fonction **phrase** effectue une recherche par termes dans le corpus de documents. Elle prend en paramètre d'entrée un terme, et elle retourne le nombre total de résultats (**totalHits**) ainsi qu'une liste d'informations sur les documents correspondants, comprenant l'identifiant (**docId**), le titre (**docTitle**), le résumé (**docContent**), et le lien URL sur PubMed (**docLink**), tel que montrée pour l'API **simple**.

## **concept** : recherche par concepts dans le corpus de documents

La fonction **concept** permet une recherche par concepts dans le corpus de documents. Elle prend en paramètre d'entrée un concept, et elle retourne le nombre total de résultats (**totalHits**) ainsi qu'une liste d'informations sur les documents correspondants, comprenant l'identifiant (**docId**), le titre (**docTitle**), le résumé (**docContent**), et le lien URL sur PubMed (**docLink**), cf exemple ci-dessous.

GET /api/concept

Parameters

Cancel

Name	Description
term <span>required</span> string (query)	<input type="text" value="Growth and Build"/>

ExecuteClear

Responses

Curl

```
curl -X 'GET' \
  'http://localhost:8083/api/concept?term=Growth%20and%20Build' \
  -H 'accept: */*'
```

Request URL

```
http://localhost:8083/api/concept?term=Growth%20and%20Build
```

Server response

Code	Details
200	<div>Response body</div> <div><pre>{   "totalHits": 10,   "documents": [     {       "docId": 32,       "docTitle": "Association between anaemia and long-term prognosis in patients with non-ST segment elevation myocardial infarction.",       "docContent": "Background: The majority of existing studies examining the association between anaemia and the prognosis of patients with acute coronary syndrome (ACS) have focused on all patients with ACS without further categorisation. As a result, there is a dearth of research specifically exploring the relationship between anaemia and the long-term prognosis of patients with non-ST segment elevation myocardial infarction (NSTEMI). To address this gap, this study aimed to investigate the correlation between anaemia and the long-term prognosis of NSTEMI patients. Methods: This study included 482 NSTEMI patients who underwent percutaneous coronary intervention (PCI) at the First Affiliated Hospital of Chongqing Medical University from September 1, 2016 to May 31, 2022, and the patients were classified into the major adverse cardiovascular events (MACE) group and non-MACE group according to whether or not they had developed MACE as of February 28, 2023 at follow-up. COX regression analysis was used to assess whether anaemia was an independent factor influencing MACE occurrence in patients with NSTEMI. Receiver operating characteristic (ROC) curve analysis was conducted to determine if haemoglobin levels could enhance the predictive capacity of the Global Registry of Acute Coronary Events (GRACE) score for the prognosis of NSTEMI patients. Haemoglobin levels were categorised into two groups based on the optimal cut-off value and transformed into binary data. The log-rank test was performed to compare the two groups, and a risk function was plotted. Results: During a median follow-up period of 31 months, 124 (25.7%) MACE were identified. Univariate and multivariate COX regression analyses revealed that sex, age, smoking history, diabetes, creatinine, erythrocyte count, and haemoglobin level were independent risk factors that significantly influenced survival time. Subsequently, ROC curve analysis was performed to evaluate the predictive accuracy of specific variables. When the cut-off value for the decline ratio of haemoglobin was set at 123.50, the area under the curve (AUC) was determined to be 0.694, with a sensitivity of 0.403 and a specificity of 0.771. Similarly, setting the cut-off value for the reduction ratio of the GRACE score at 141.5 yielded an AUC of 0.700, with a sensitivity of 0.645 and a specificity of 0.709. Furthermore, when the cut-off value for the predicted probability of haemoglobin combined with the GRACE score was 0.270, the AUC was calculated as 0.702, with a sensitivity of 0.677 and a specificity of 0.696. Conclusion: Haemoglobin levels were identified as an independent factor influencing the survival duration of patients with NSTEMI."     },     {       "docId": 330, </pre></div> <div>Download</div>

Response headers

```
connection: keep-alive
content-type: application/json
date: Thu, 14 Dec 2023 13:00:10 GMT
keep-alive: timeout=60
transfer-encoding: chunked
vary: Origin,Access-Control-Request-Method,Access-Control-Request-Headers
```

## Guide d'utilisation

Le guide d'utilisation est disponible dans le [README](#) à la racine du projet.

## Difficultés rencontrées

Le projet ne comprend pas de techniques de proposition d'articles similaires, ni de système permettant d'indiquer la performance du système.

L'indexation par termes se fait à l'aide du WhitespaceAnalyzer plutôt que les concepts compris dans l'ontologie.