Institut européen
des métiers de la **traduction** | IEMT
Université de Strasbourg

Web, corpus, traduction : exploitations

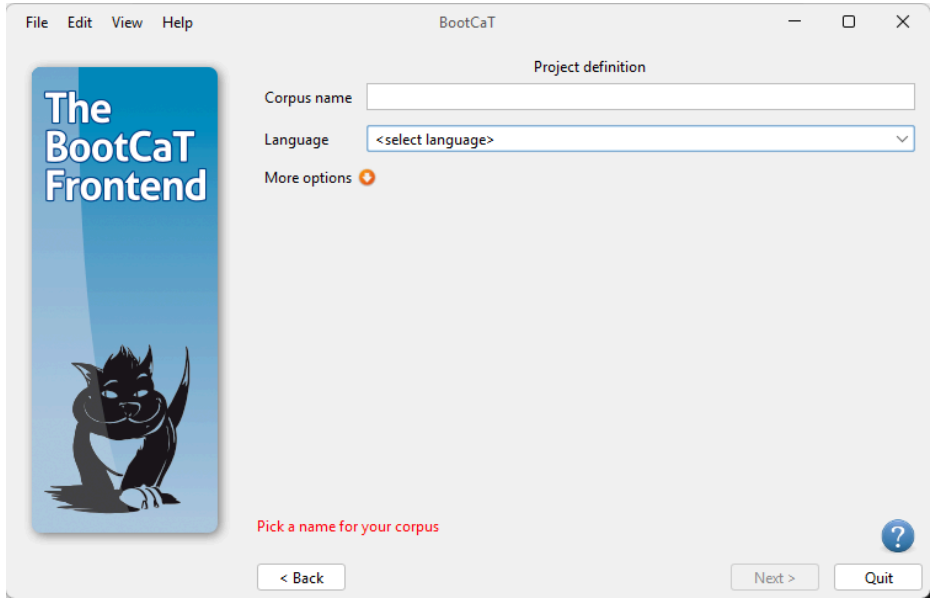
BootCaT

Enzo Doyen

2025 - M1

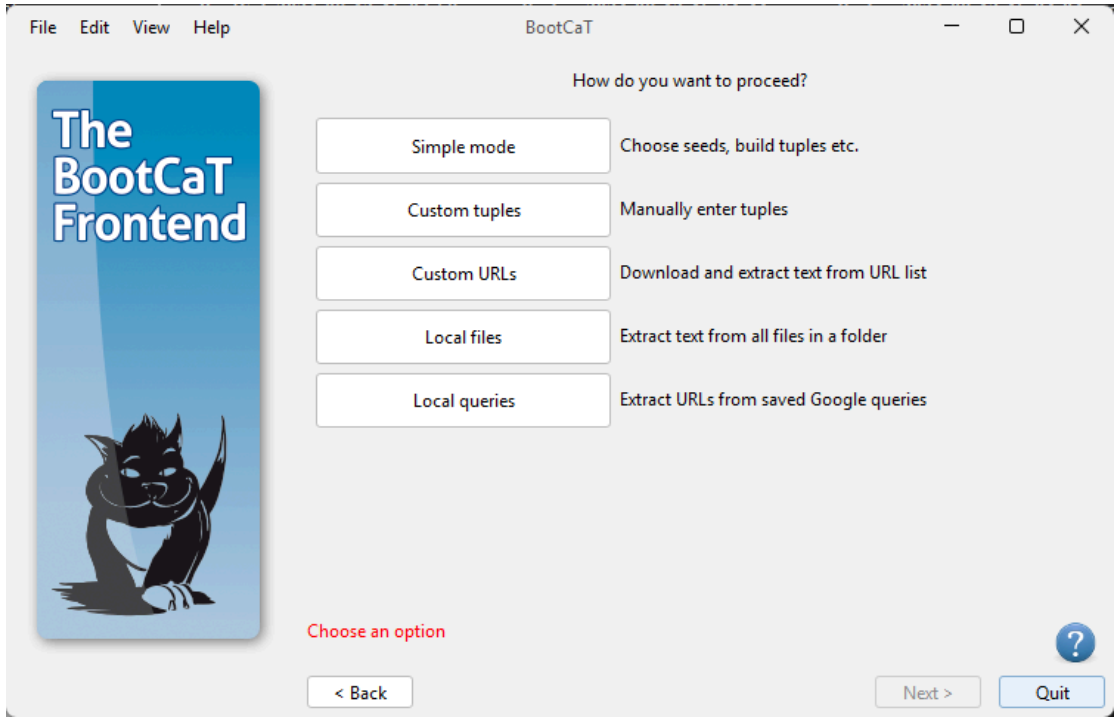
BootCaT : fenêtre de démarrage

Indiquez le nom du corpus et la langue.



BootCaT : modes

Cinq modes disponibles.



BootCaT : modes

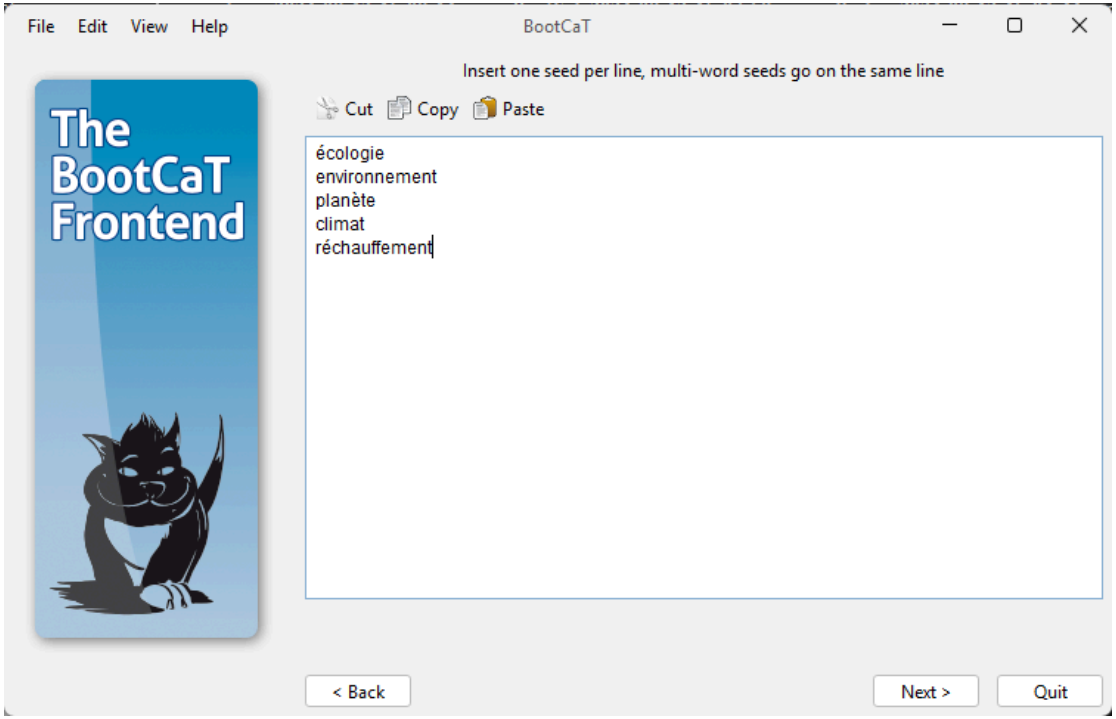
1. **Simple mode** : mode guidé (mode le plus simple et recommandé).
2. **Custom tuples** : définition manuelle des uplets de mots-clés.
3. **Custom URLs** : liste d'adresses URL des pages desquelles extraire le contenu, sous forme de fichier texte.
4. **Local files** : fichiers locaux desquels extraire le contenu (fichiers .txt, .html, .doc, .pdf…).
5. **Local queries** : fichiers HTML de recherches Google contenant les adresse URL des pages à extraire (ce que l'on fait également en *Simple mode*).

BootCaT : modes

1. **Simple mode** : mode guidé (mode le plus simple et recommandé).
2. **Custom tuples** : définition manuelle des uplets de mots-clés.
3. **Custom URLs** : liste d'adresses URL des pages desquelles extraire le contenu, sous forme de fichier texte.
4. **Local files** : fichiers locaux desquels extraire le contenu (fichiers .txt, .html, .doc, .pdf…).
5. **Local queries** : fichiers HTML de recherches Google contenant les adresse URL des pages à extraire (ce que l'on fait également en *Simple mode*).

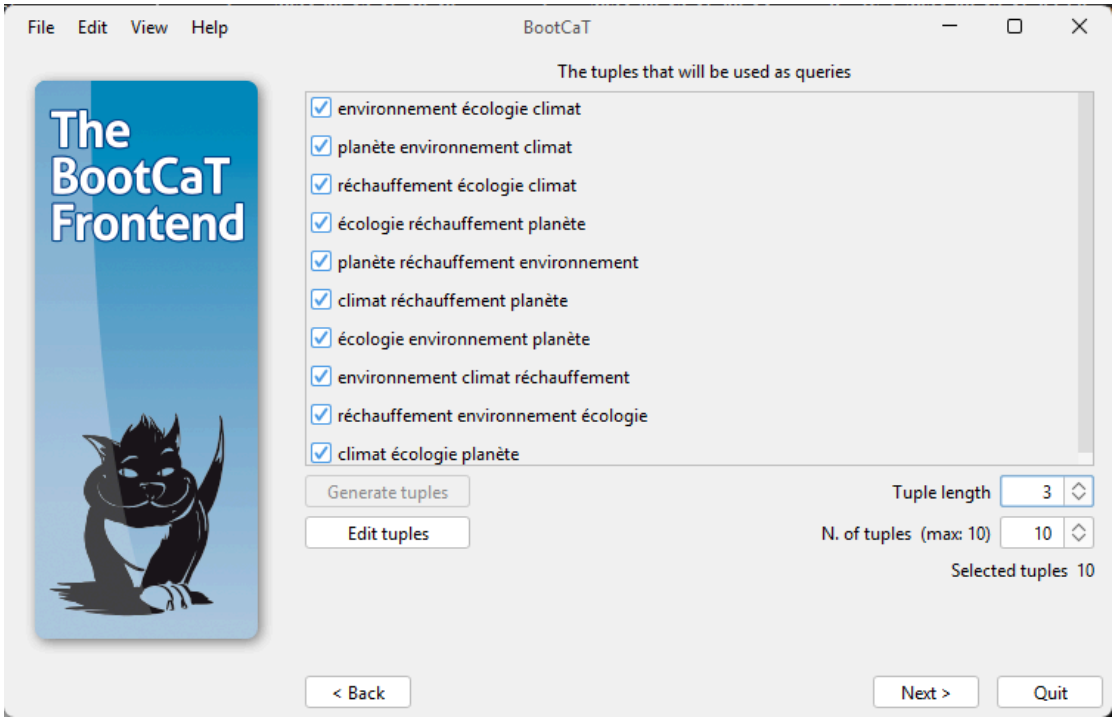
BootCaT : mode de base

Indiquez les mots-clés à rechercher (*seed*), un par ligne.



BootCaT : mode de base

Une fois les mots-clés fournis, des uplets (*tuple* en anglais) de mots-clés sont générés.



Notion de uplets (*tuple*)

Un **uplet** est une séquence ordonnée finie de n éléments.

$$t = (t_1, t_2, ..., t_n)$$

Un uplet de 2 éléments s'appelle un doublet, un uplet de 3 éléments s'appelle un triplet, etc.

Par exemple, $t = (\text{"environnement"}, \text{"écologie"}, \text{"climat"})$

Connaître le nombre d'uplets générés

On note n le nombre de mots-clés indiqués à BootCaT et l le nombre de mots-clés par uplet.

À travers la formule suivante, on veut savoir : combien il y a-t-il de combinaisons possibles de l parmi n ?

Coefficient binomial [?] $\binom{n}{l} = \frac{n!}{l!(n-l)!}$

(Dans notre cas, l'ordre des éléments n'est pas important.)

Connaître le nombre d'uplets générés

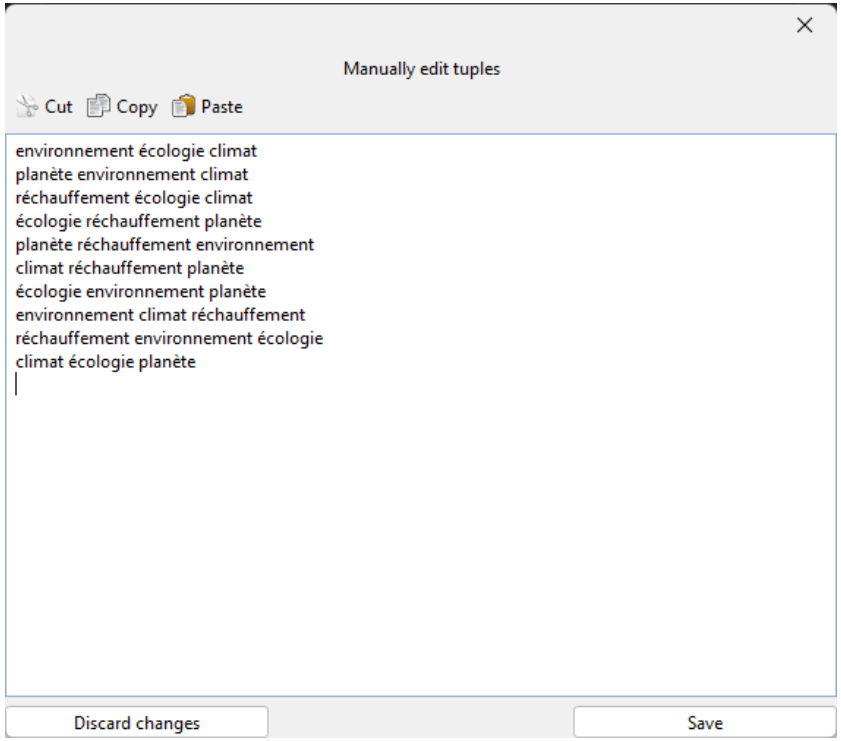
Exemple avec 5 mots-clés au total ($n = 5$) et 3 mots-clés par uplet ($l = 3$) :

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4}{2 \times 1} = 10$$

10 uplets de 3 mots-clés sont générés.

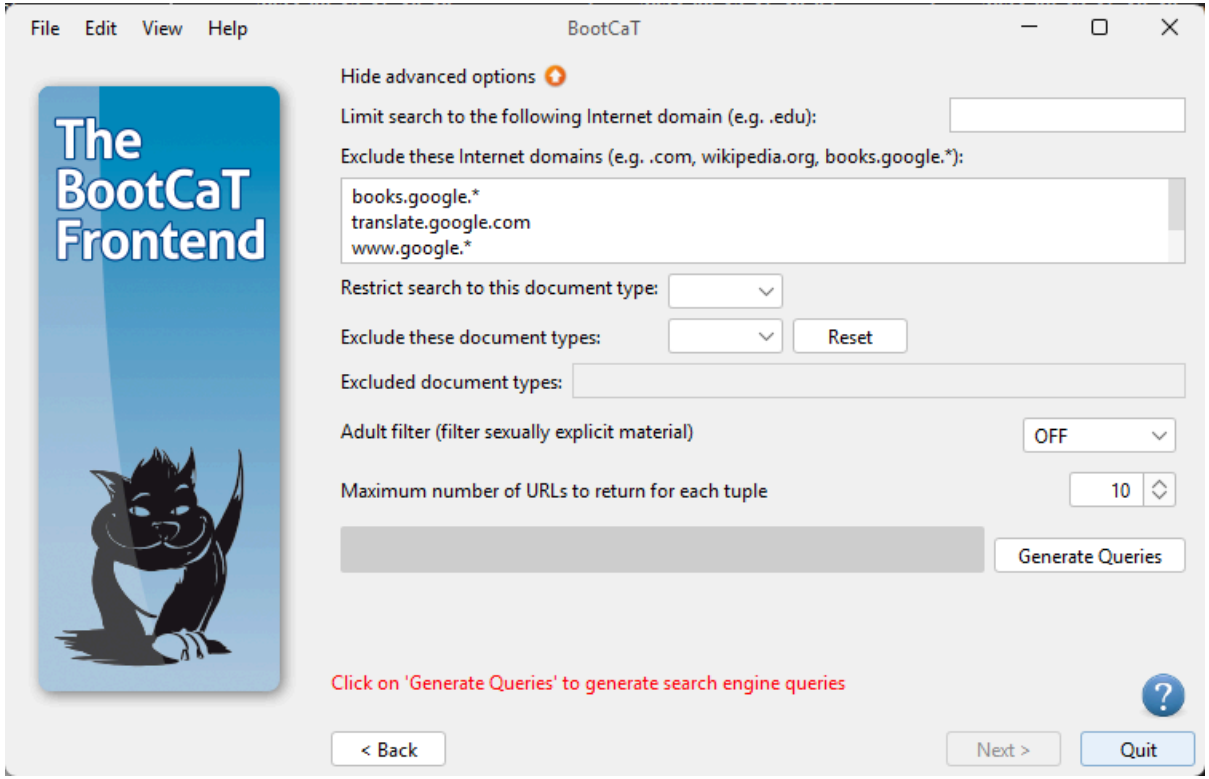
BootCaT : mode de base

On peut également modifier manuellement les uplets.



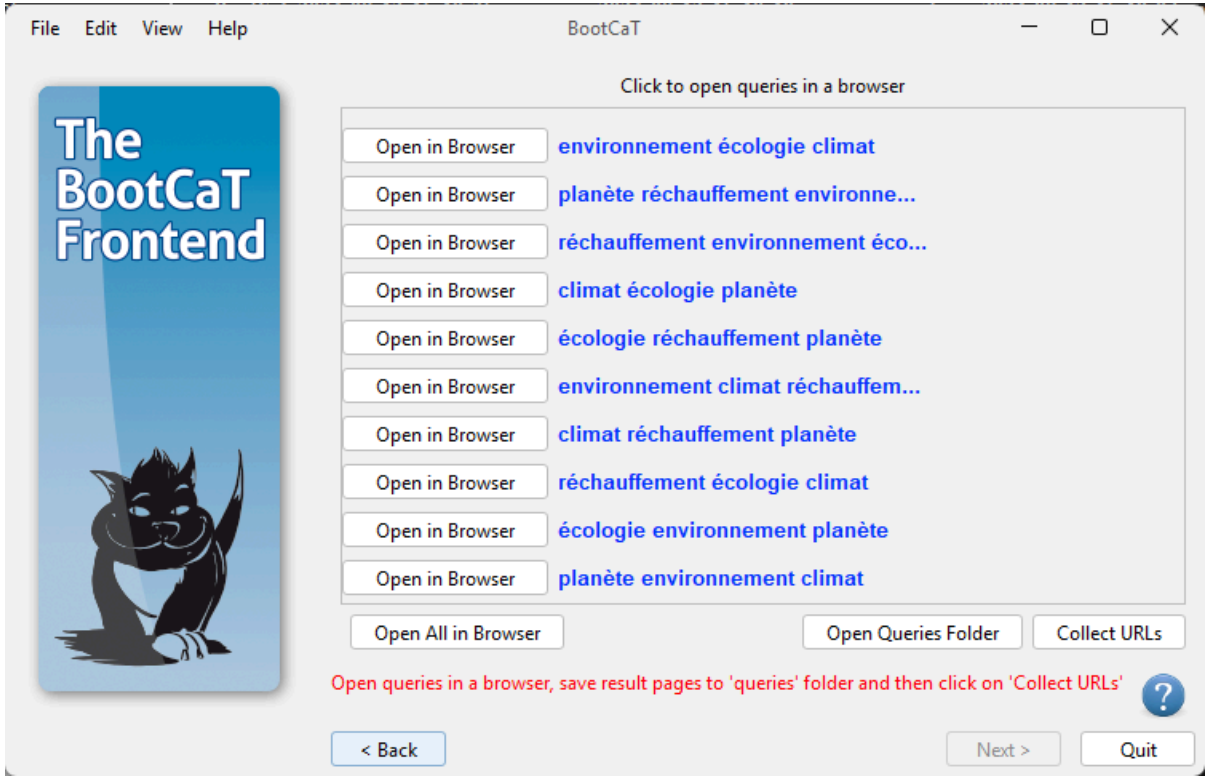
BootCaT : mode de base

Possibilité de filtrer le contenu à extraire.



BootCaT : mode de base

Voici les requêtes récupérées.



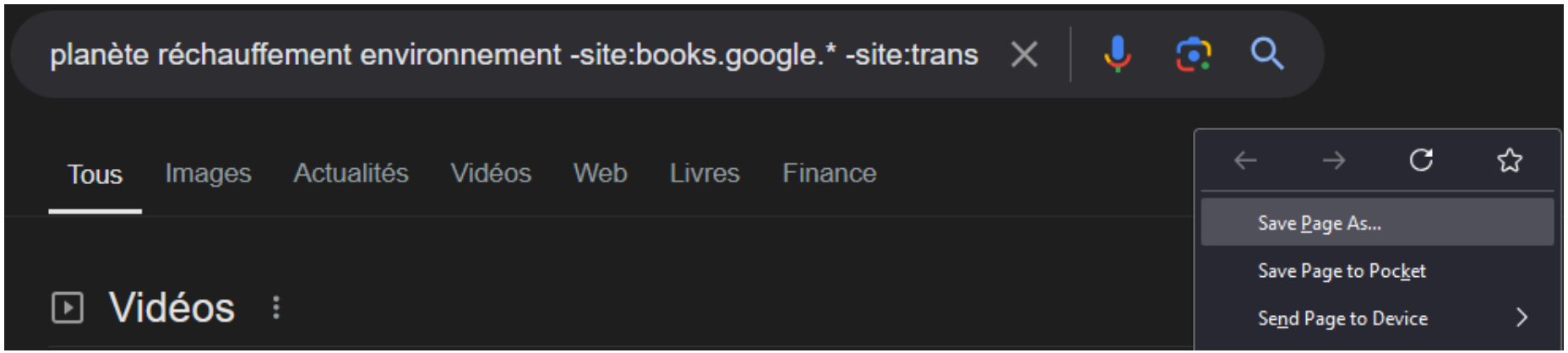
Sauvegarde des pages de requêtes

Pour éviter tout blocage de la part de Google, il nous est demandé de sauvegarder nous-même les pages de résultats.

On peut cliquer sur « **Open All in Browser** » pour ouvrir toutes les pages de requêtes Google en même temps : cela ouvrira plusieurs onglets dans votre navigateur.

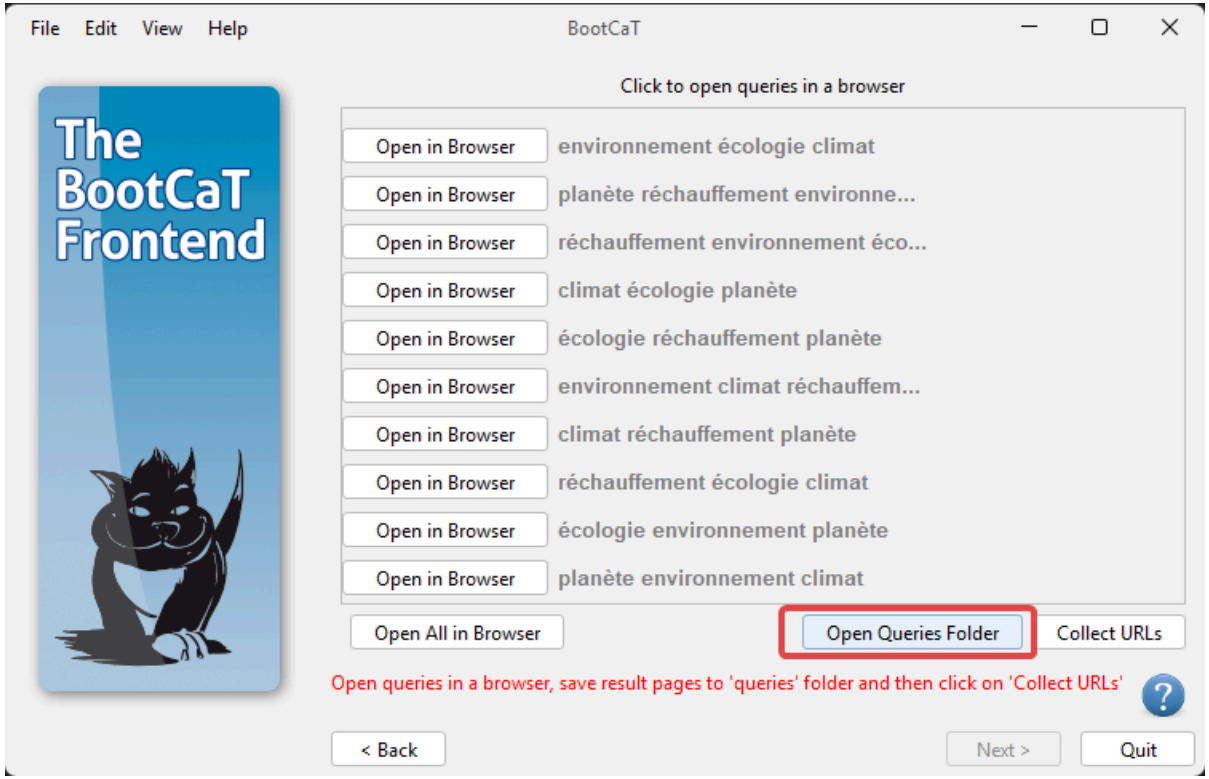
BootCaT : sauvegarde des pages de requêtes

Par la suite, vous pouvez appuyer sur CTRL + S pour sauvegarder chaque page, ou bien faire un clic droit → **Enregistrer sous**.



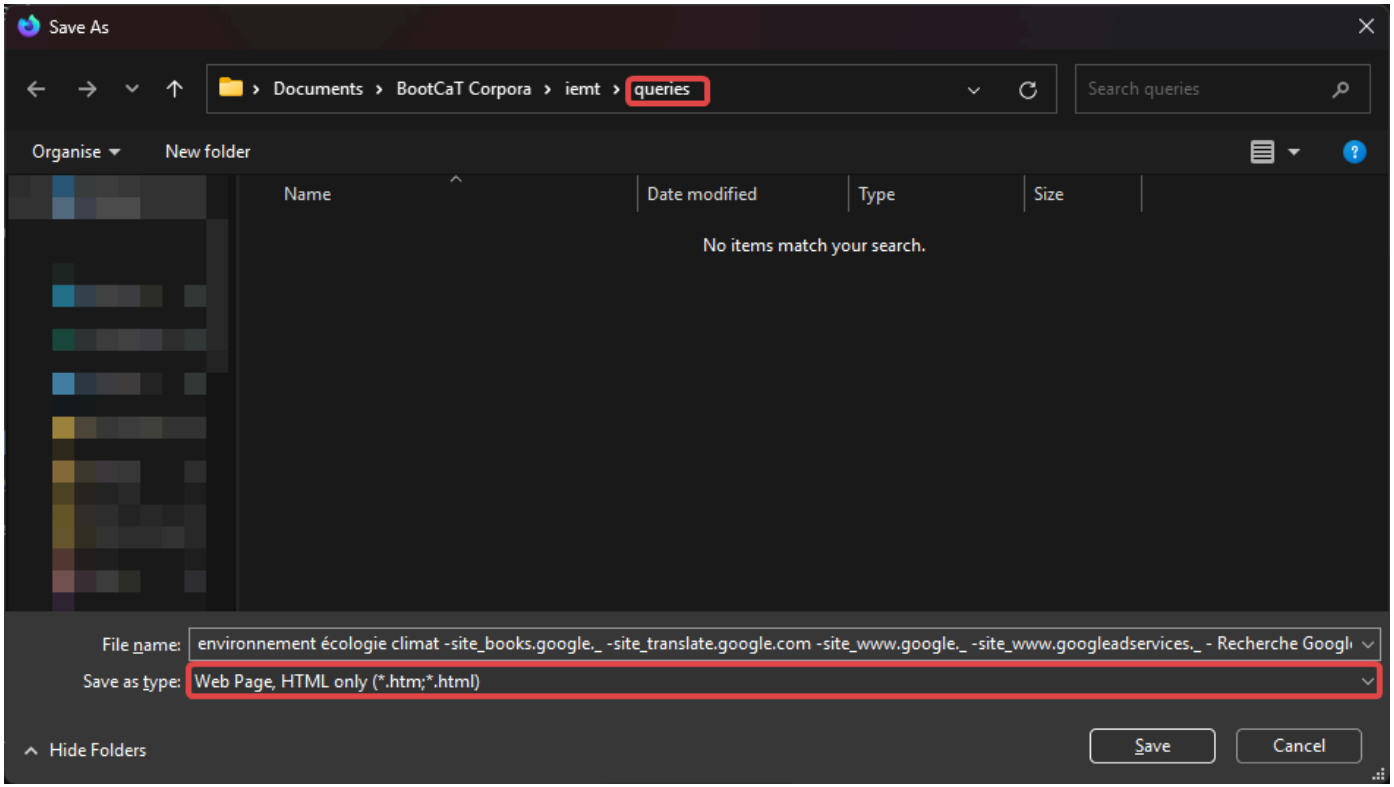
BootCaT : sauvegarde des pages de requêtes

Les pages doivent être sauvegardées dans le dossier « queries » :



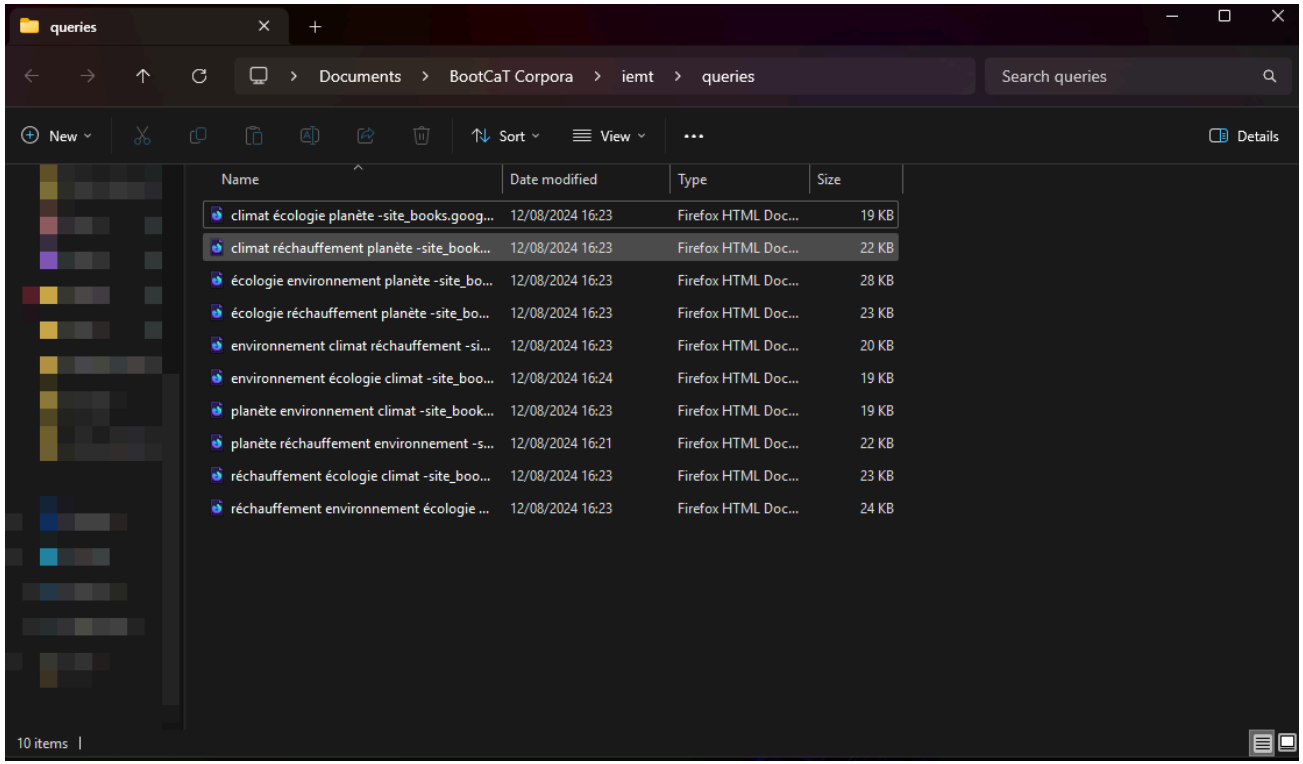
BootCaT : sauvegarde des pages de requêtes

Attention à bien sauvegarder les fichiers au format HTML !



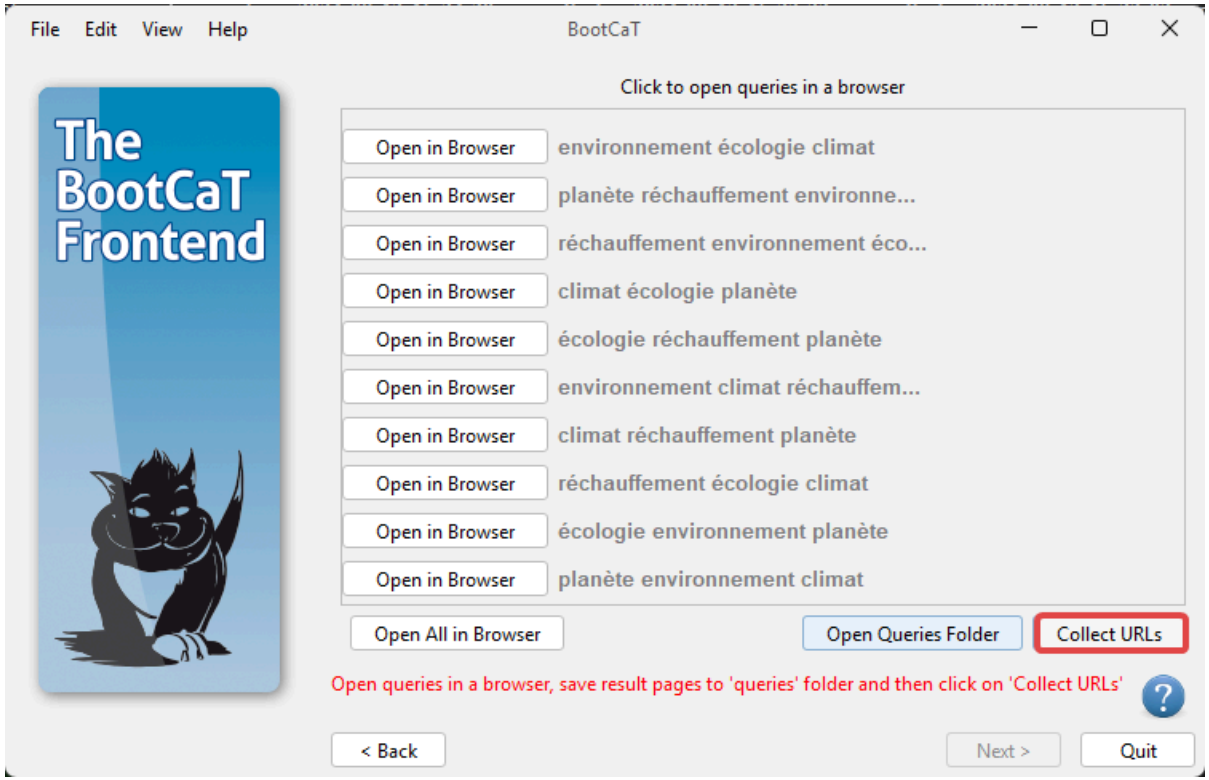
BootCaT : sauvegarde des pages de requêtes

Voici ce que l'on obtient :



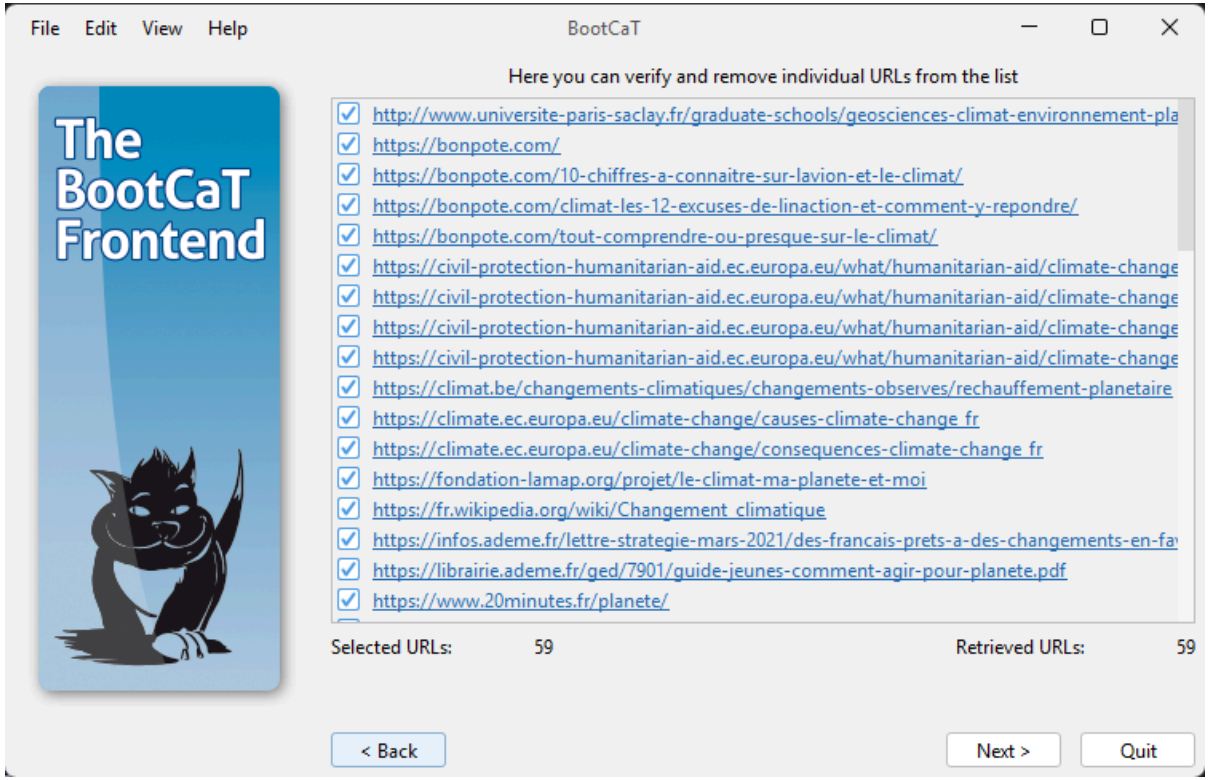
BootCaT : sauvegarde des pages de requêtes

Ensuite, on récupère les adresses URL :



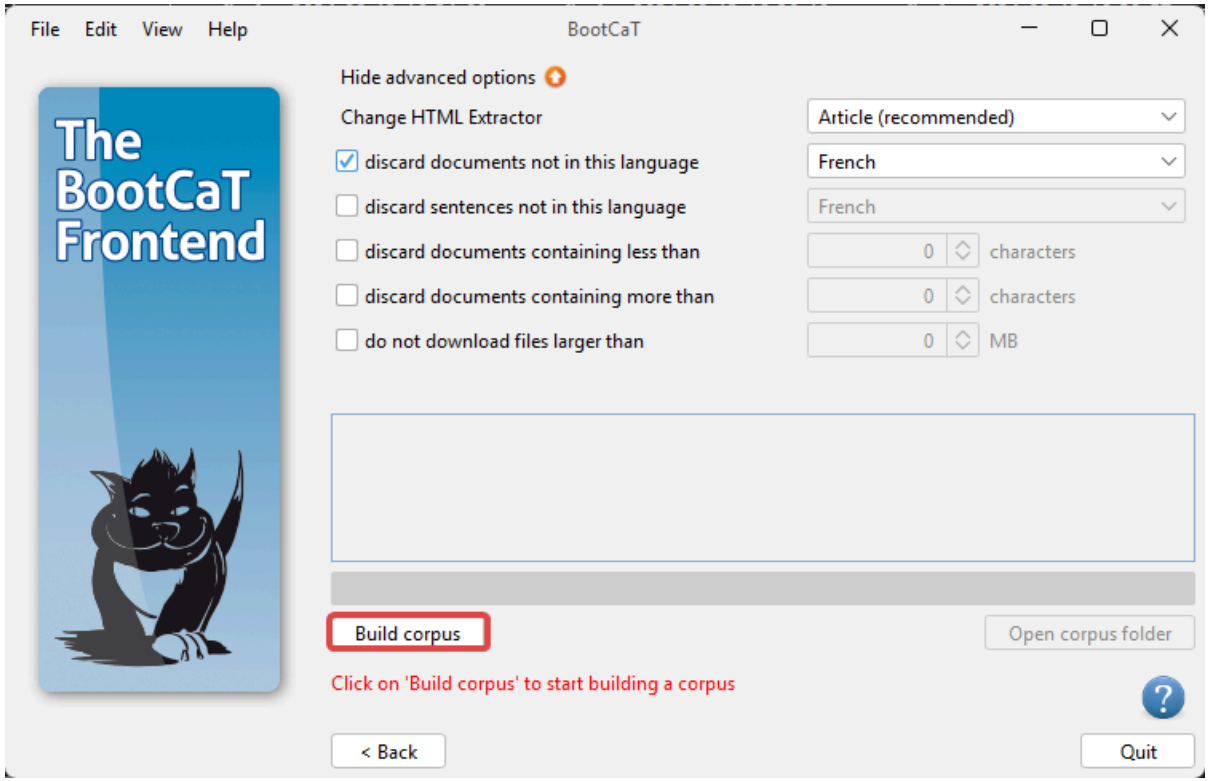
BootCaT : sauvegarde des pages de requêtes

On peut filtrer les adresses URL :



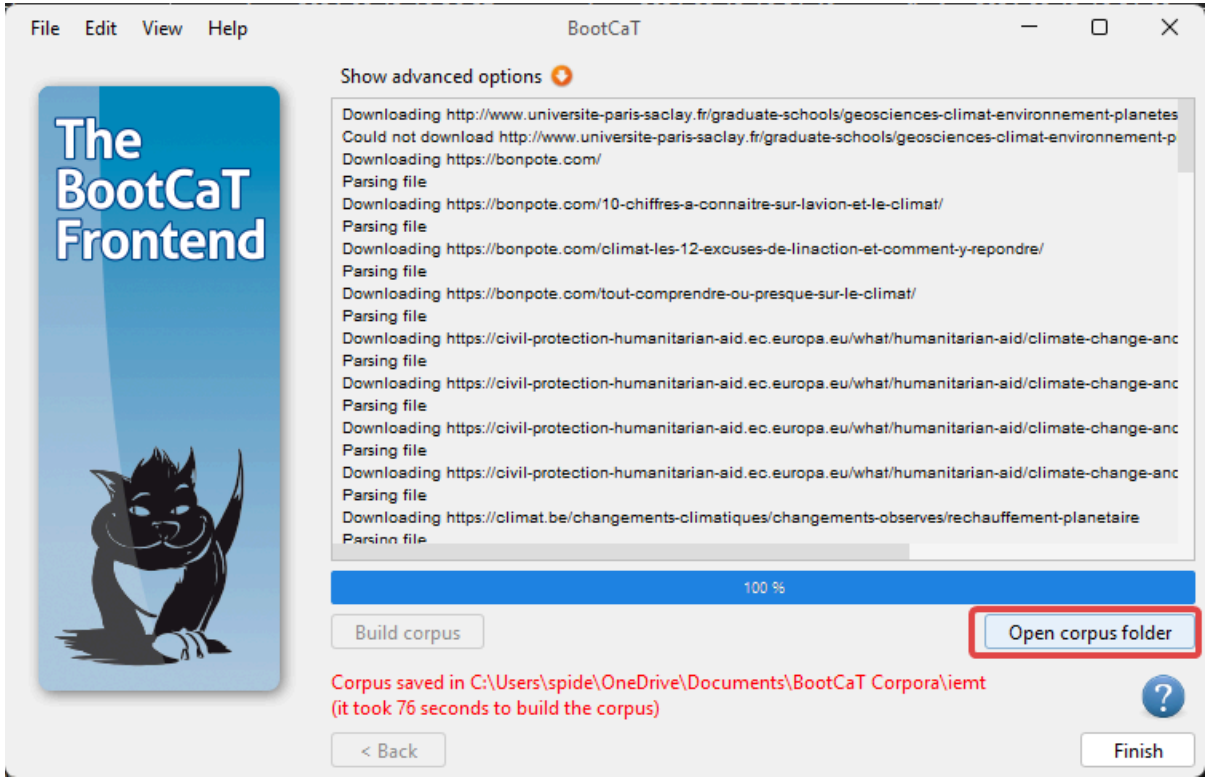
BootCaT : sauvegarde des pages de requêtes

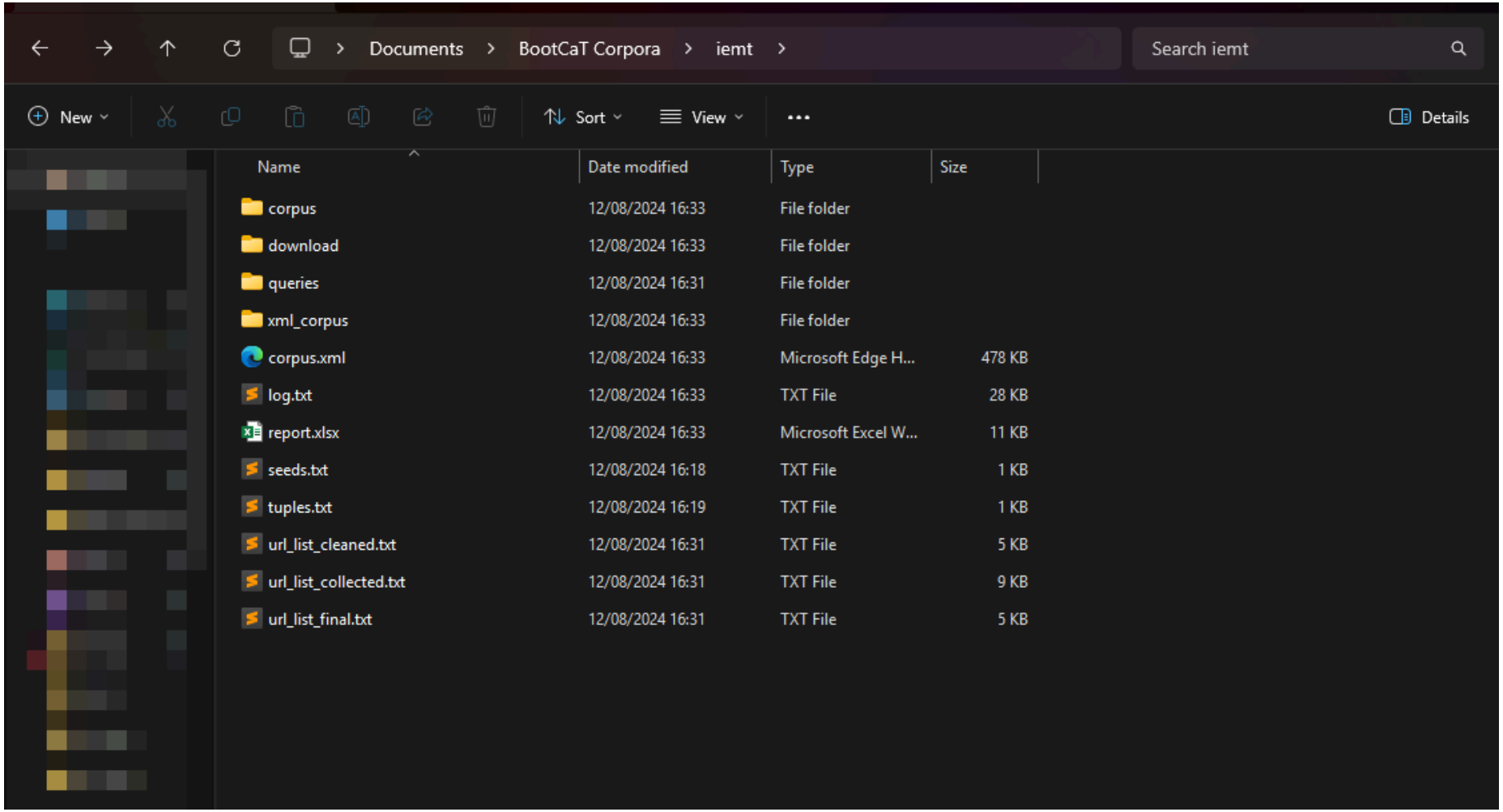
Il n'y a plus qu'à compiler le corpus !



BootCaT : sauvegarde des pages de requêtes

Enfin, on ouvre le dossier dans lequel le corpus a été compilé.




Documents
BootCaT Corpora
iemt
Search iemt
New
Sort
View
Details
Name
Date modified
Type
Size
corpus 12/08/2024 16:33 File folder
download 12/08/2024 16:33 File folder
queries 12/08/2024 16:31 File folder
xml_corpus 12/08/2024 16:33 File folder
corpus.xml 12/08/2024 16:33 Microsoft Edge H... 478 KB
log.txt 12/08/2024 16:33 TXT File 28 KB
report.xlsx 12/08/2024 16:33 Microsoft Excel W... 11 KB
seeds.txt 12/08/2024 16:18 TXT File 1 KB
tuples.txt 12/08/2024 16:19 TXT File 1 KB
url_list_cleaned.txt 12/08/2024 16:31 TXT File 5 KB
url_list_collected.txt 12/08/2024 16:31 TXT File 9 KB
url_list_final.txt 12/08/2024 16:31 TXT File 5 KB

BootCaT : sauvegarde des pages de requêtes

- Le dossier **corpus** contient une liste de fichiers .txt qui répertorient le texte récupéré pour chaque page.
- Le fichier **seeds.txt** contient les mots-clés du corpus.
- Le fichier **tuples.txt** contient les uplets de mots-clés.
- Les fichiers avec **url** dans le nom contiennent les adresses URL des pages d'où proviennent les textes.