



Institut européen

des métiers de la **traduction** | IEMT

Université de Strasbourg

Web, corpus, traduction : exploitations

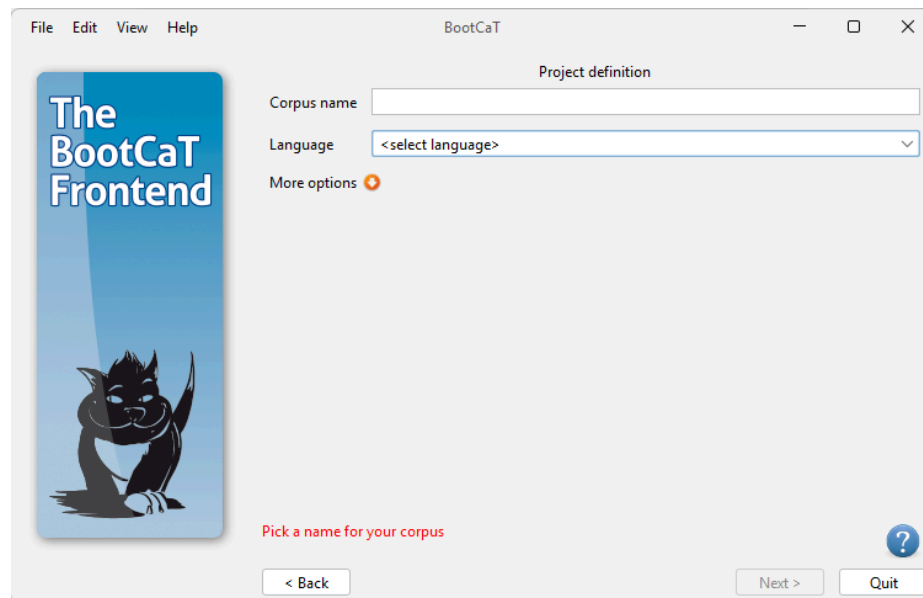
BootCaT

Enzo Doyen

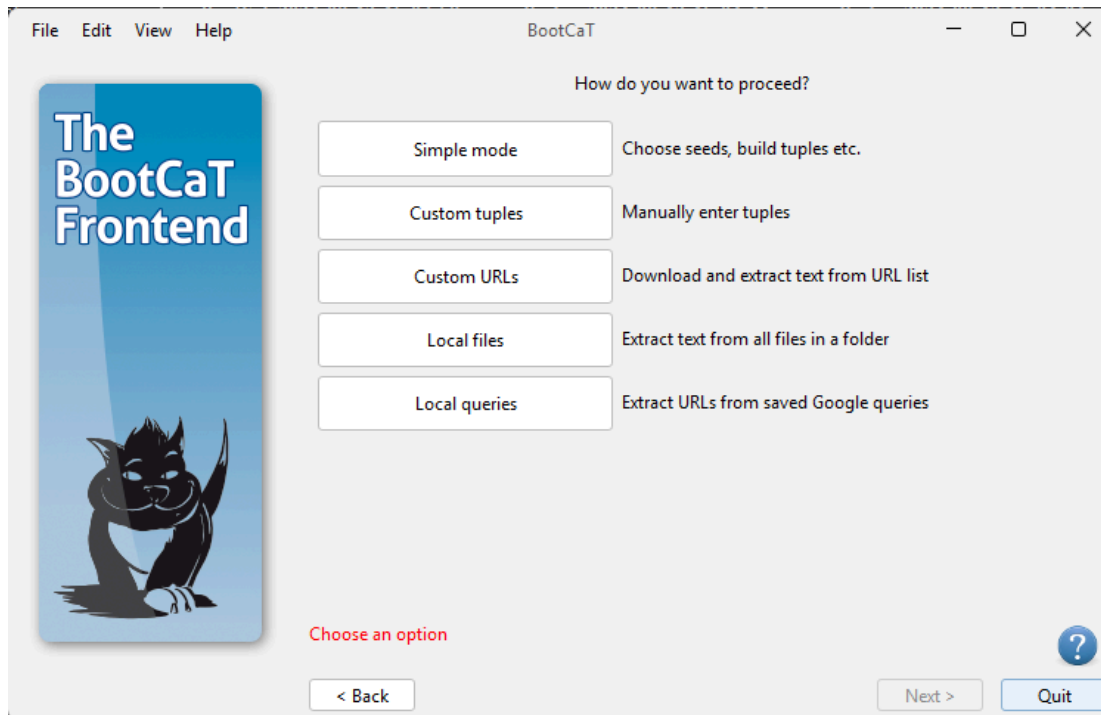
2025 - M1

Étapes de création d'un corpus personnalisé

Indiquez le nom du corpus et la langue.

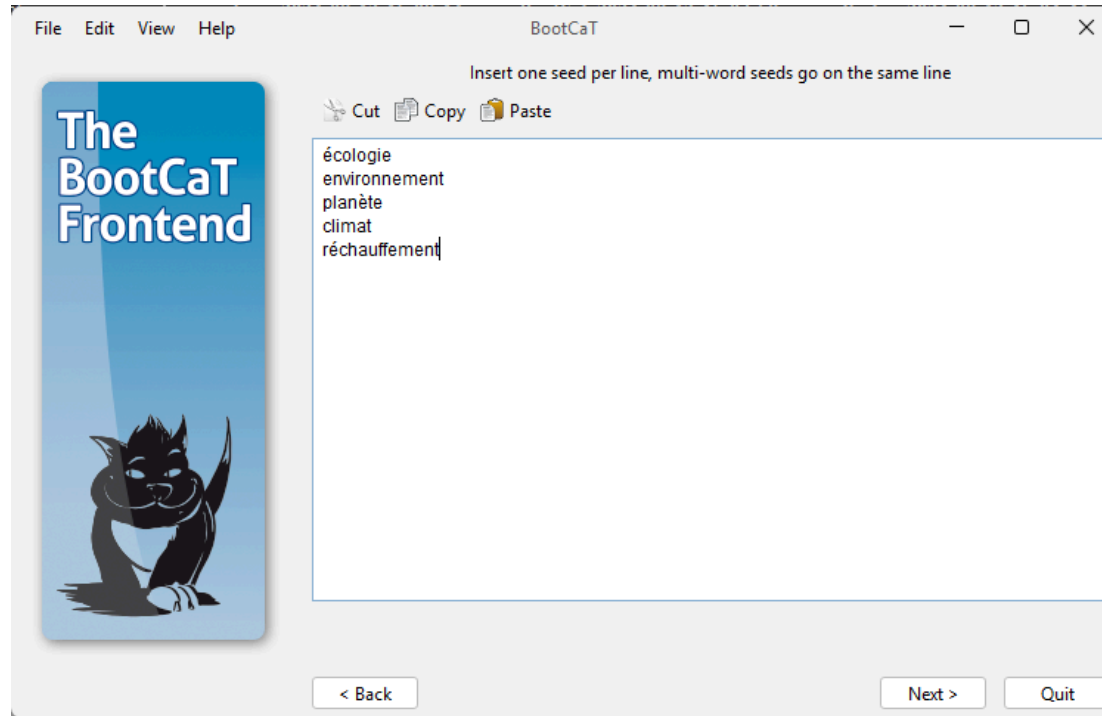


Cinq modes disponibles.

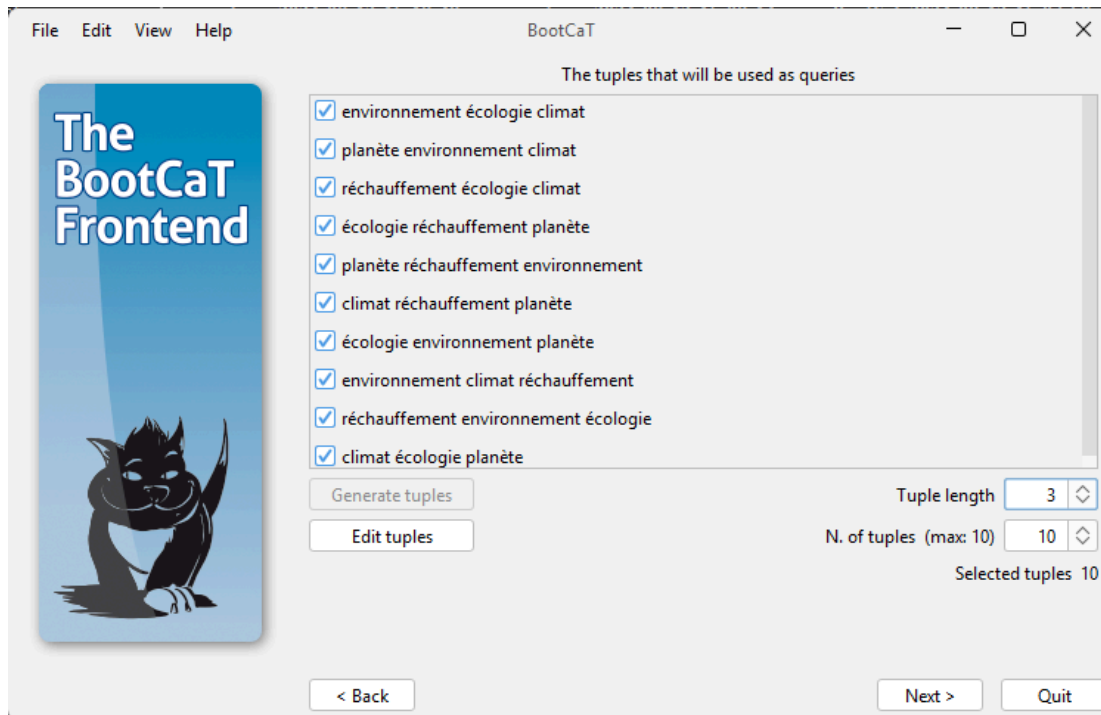


1. **Simple mode** : mode guidé.
2. **Custom tuples** : définition manuelle des uplets de mots-clés (on verra ça dans quelques minutes !).
3. **Custom URLs** : liste d'adresses URL des pages desquelles extraire le contenu, sous forme de fichier texte.
4. **Local files** : fichiers locaux desquels extraire 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*).

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



Des uplets (*tuple* en anglais) de mots-clés sont générés.



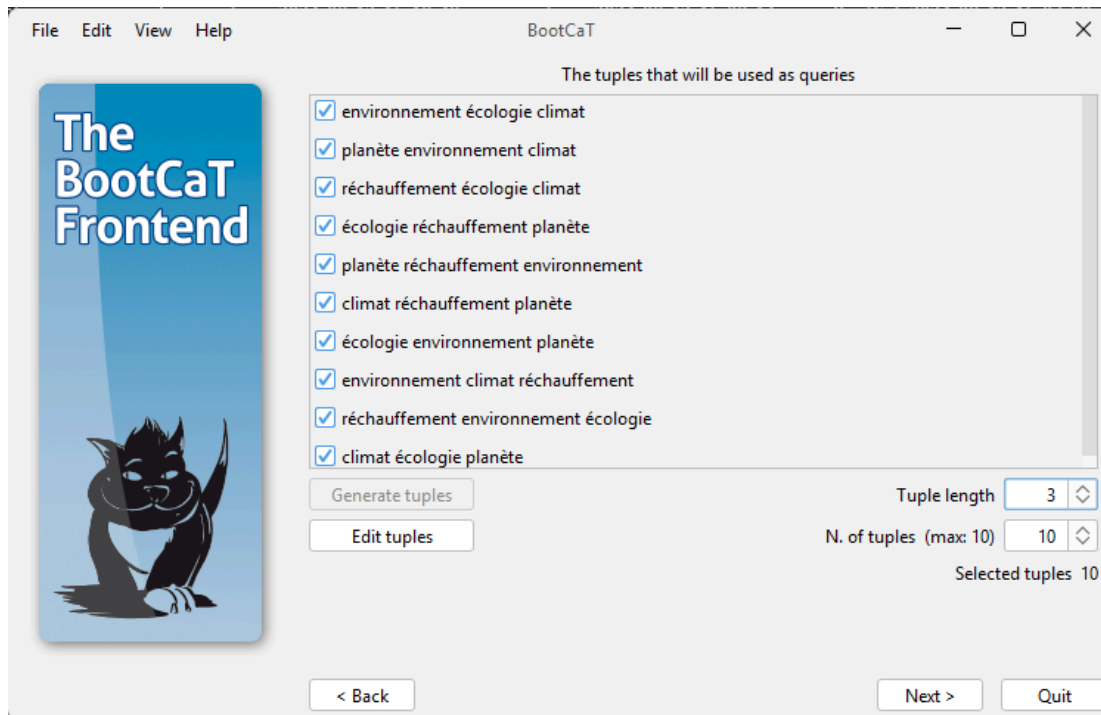
Notion de uplets (*tuple*)

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

$$t = (t_1, t_2, \dots, t_n)$$

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.

$$\binom{n}{l} = \frac{n!}{l!(n-l)!}$$

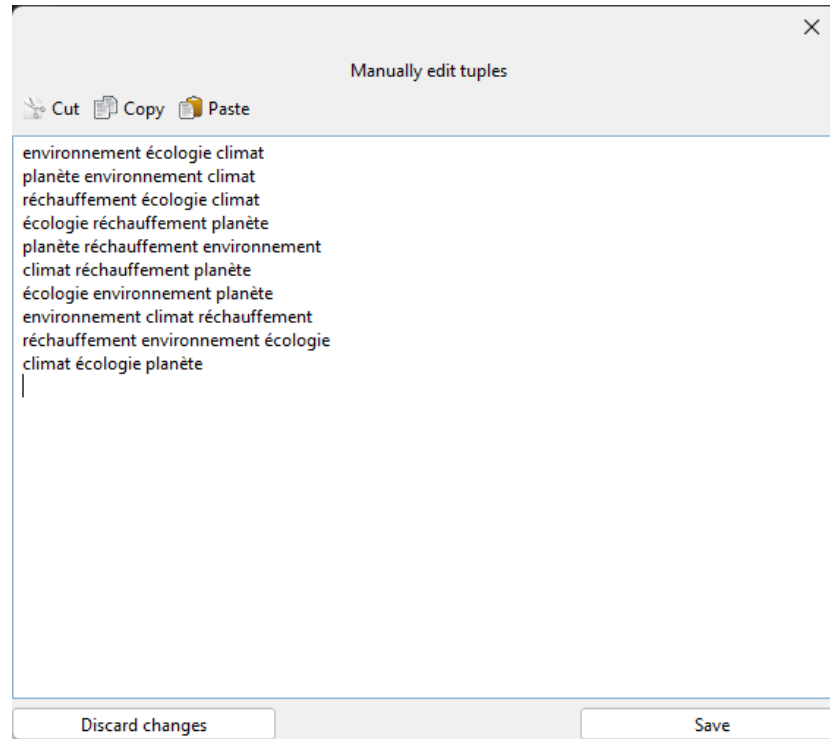
(Dans notre cas, l'ordre des éléments n'est pas important.)

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

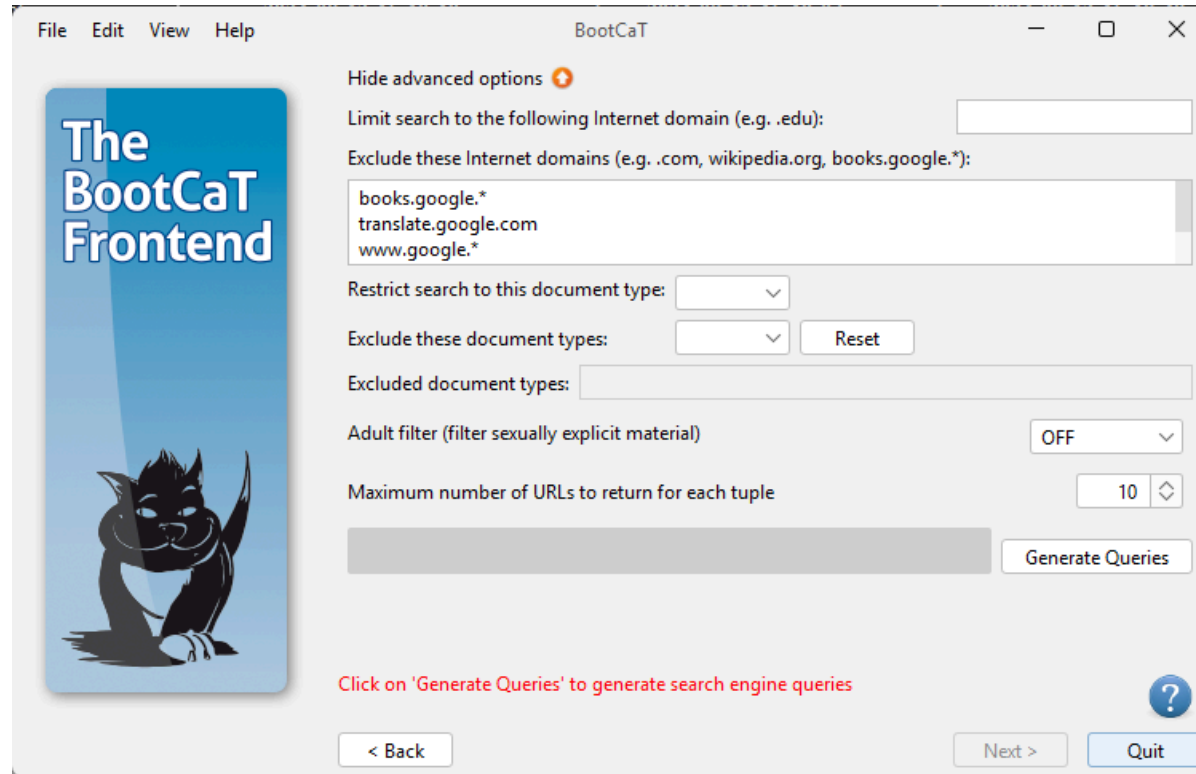
$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4}{2 \times 1} = 10$$

10 uplets de 3 mots-clés sont générés.

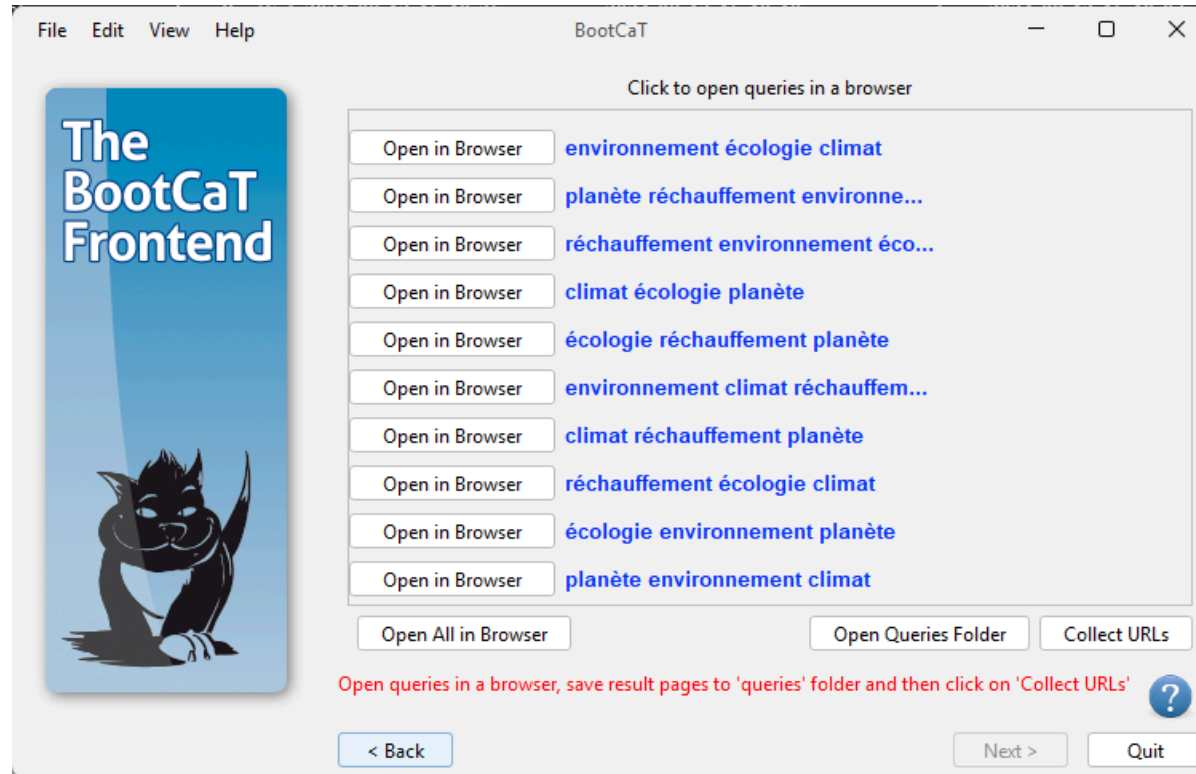
On peut également modifier manuellement les uplets.



Possibilité de filtrer le contenu à extraire.



Voici les requêtes récupérées.



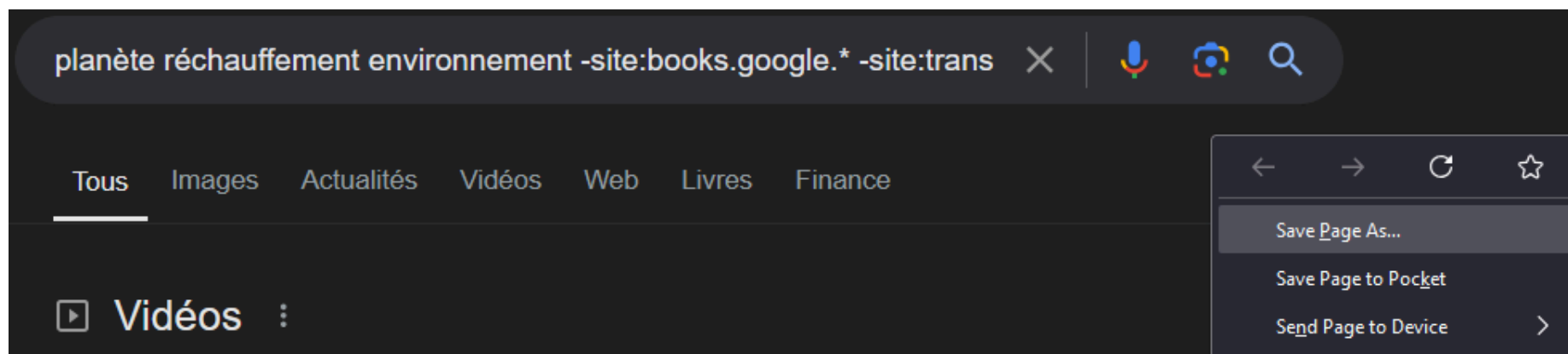
Sauvegarde des pages de requêtes

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

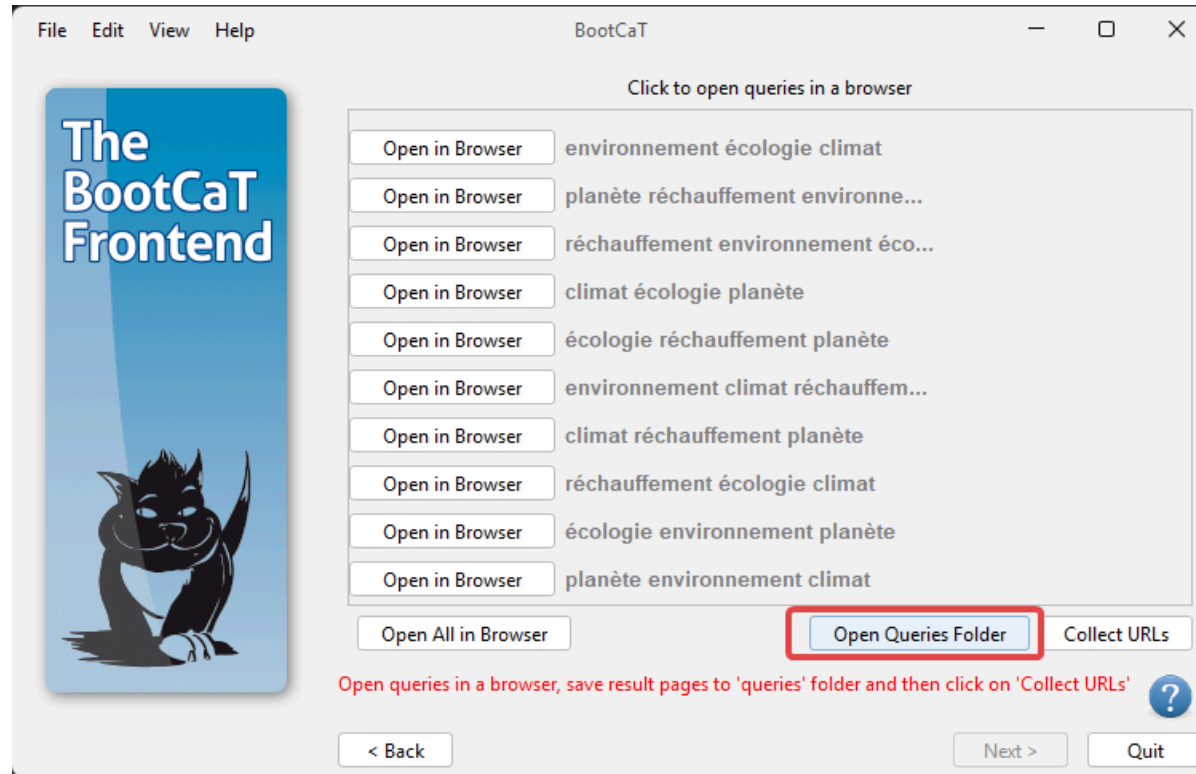
On peut cliquer sur « **Open All in Browser** » pour ouvrir toutes les pages de requêtes Google en même temps : cela ouvrira plusieurs onglets dans votre navigateur.

Sauvegarde des pages de requêtes

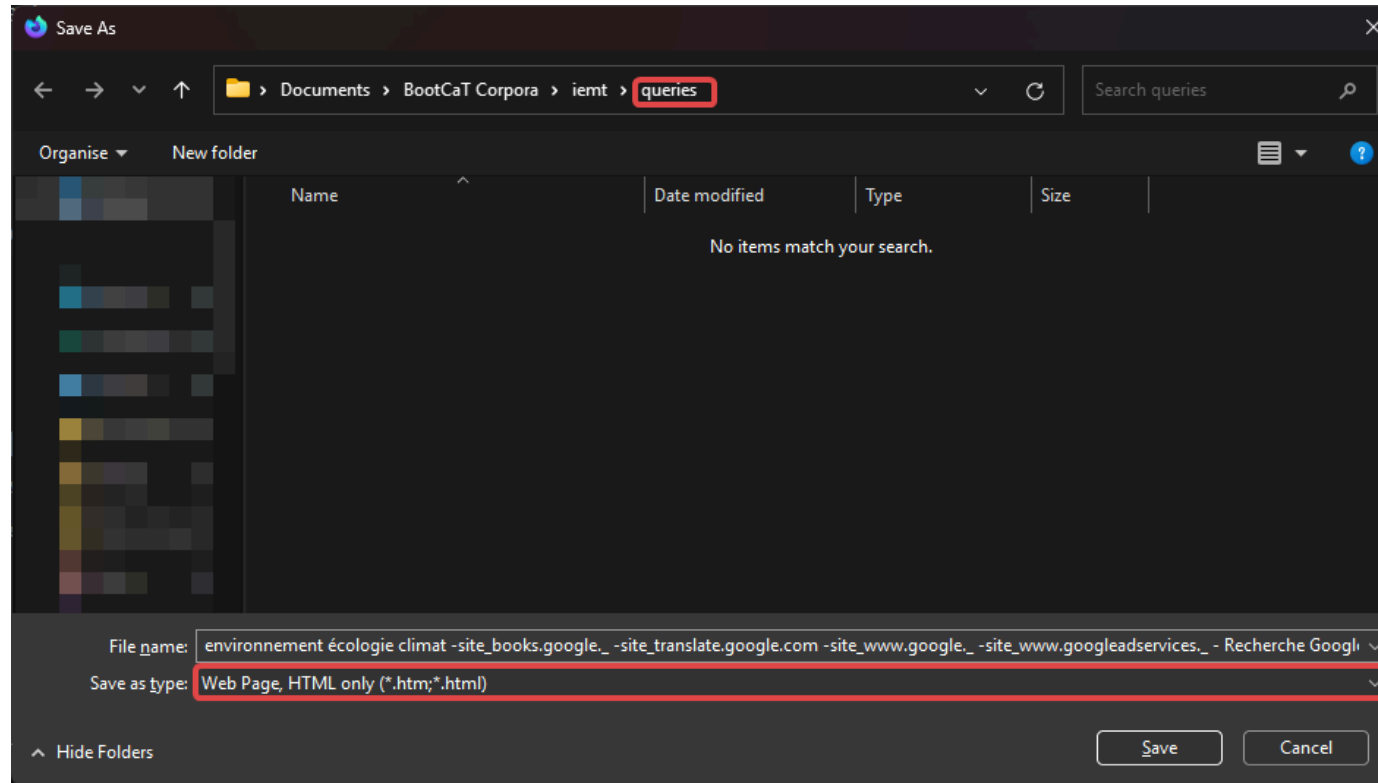
Par la suite, vous pouvez appuyer sur CTRL + S pour sauvegarder chaque page, ou bien faire un clic droit → **Enregistrer sous**.



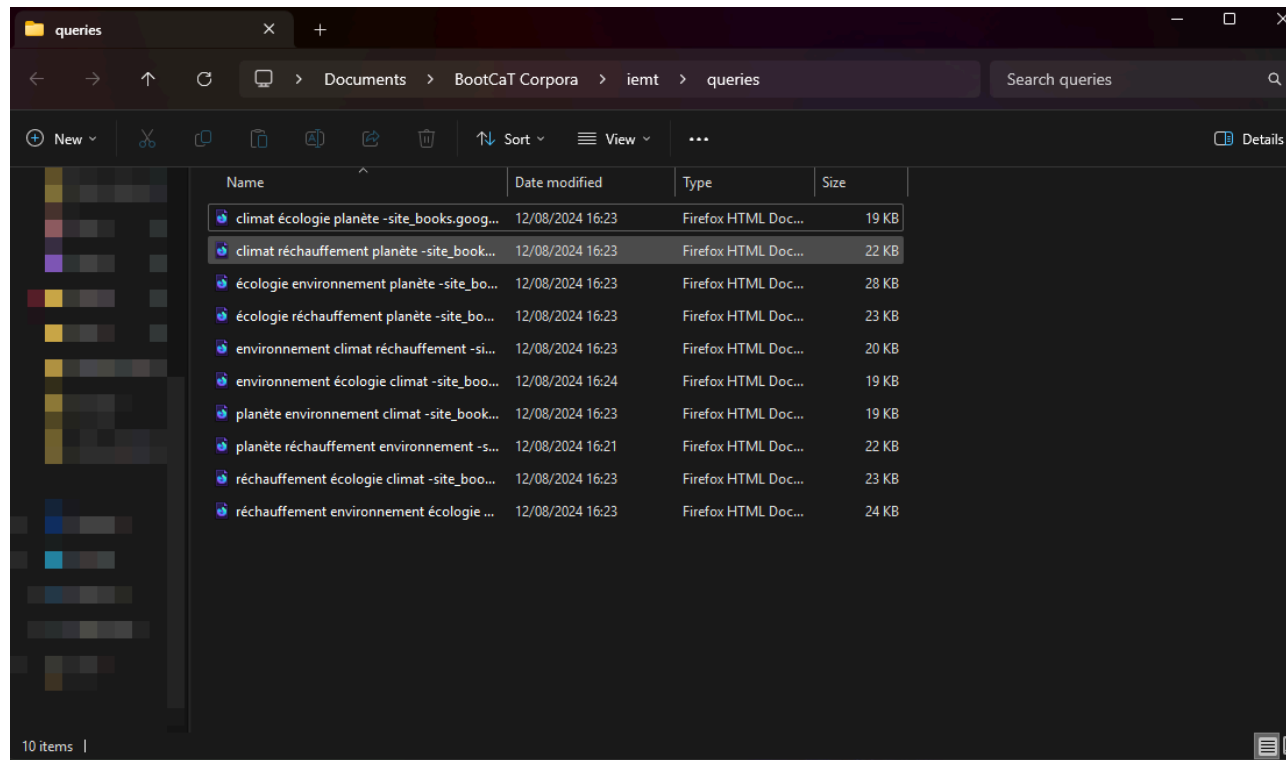
Les pages doivent être sauvegardées dans le dossier « queries » :



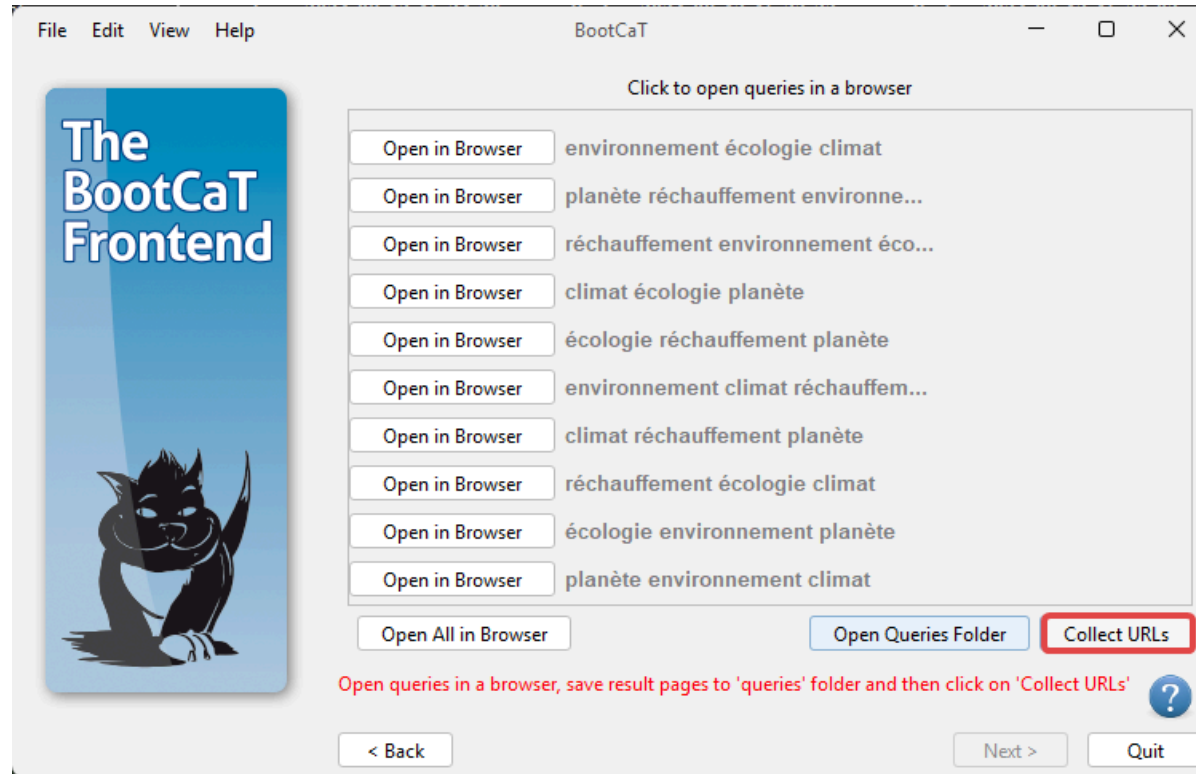
Attention à bien sauvegarder les fichiers au format HTML !



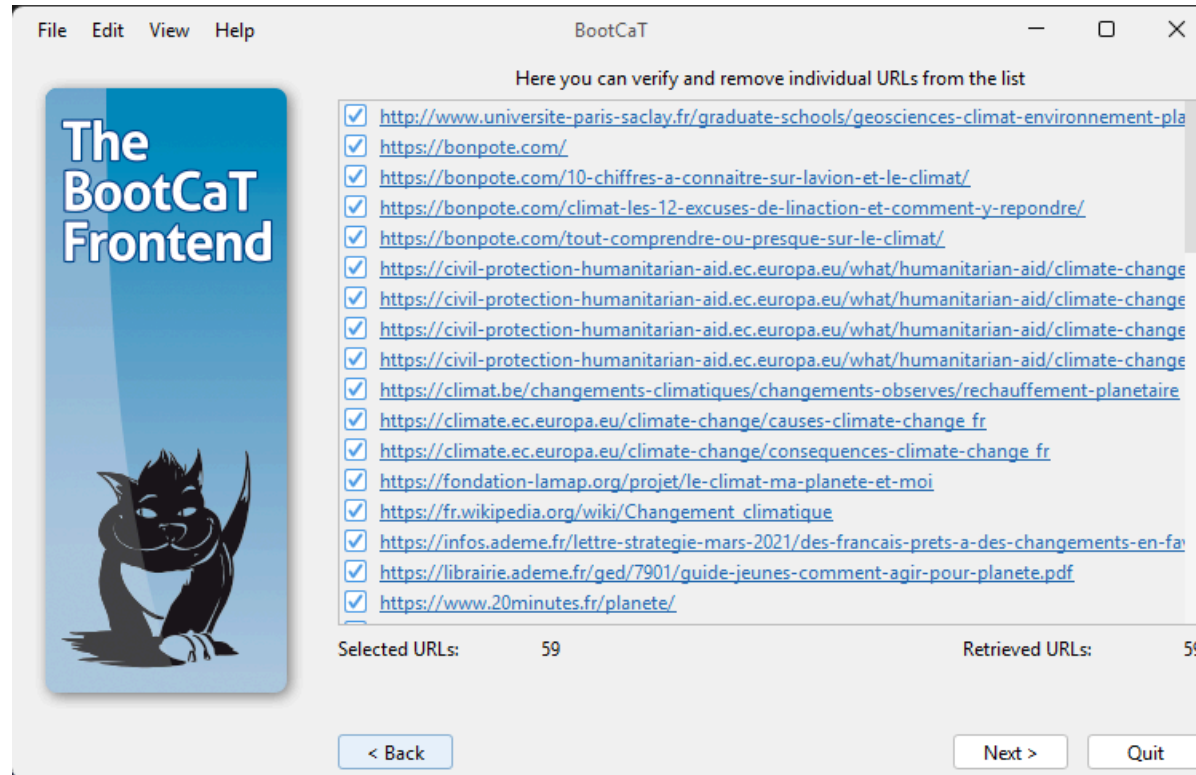
Voici ce que l'on obtient :



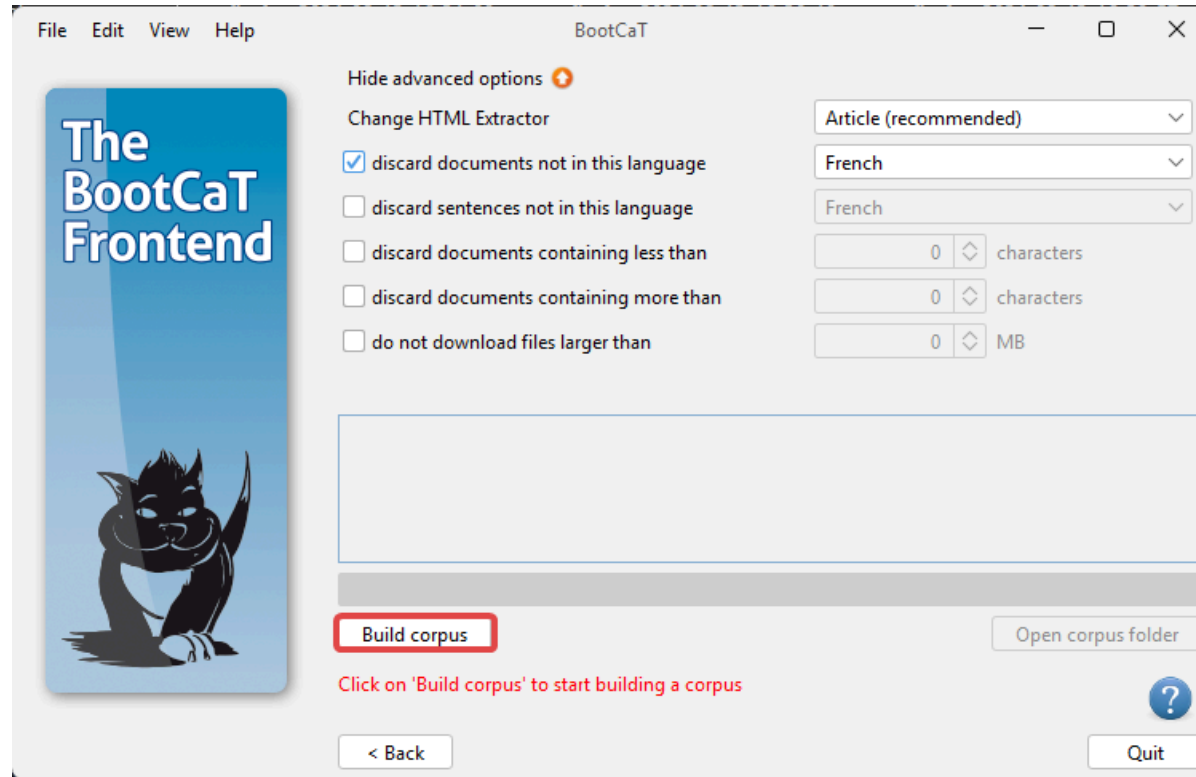
Ensuite, on récupère les adresses URL :



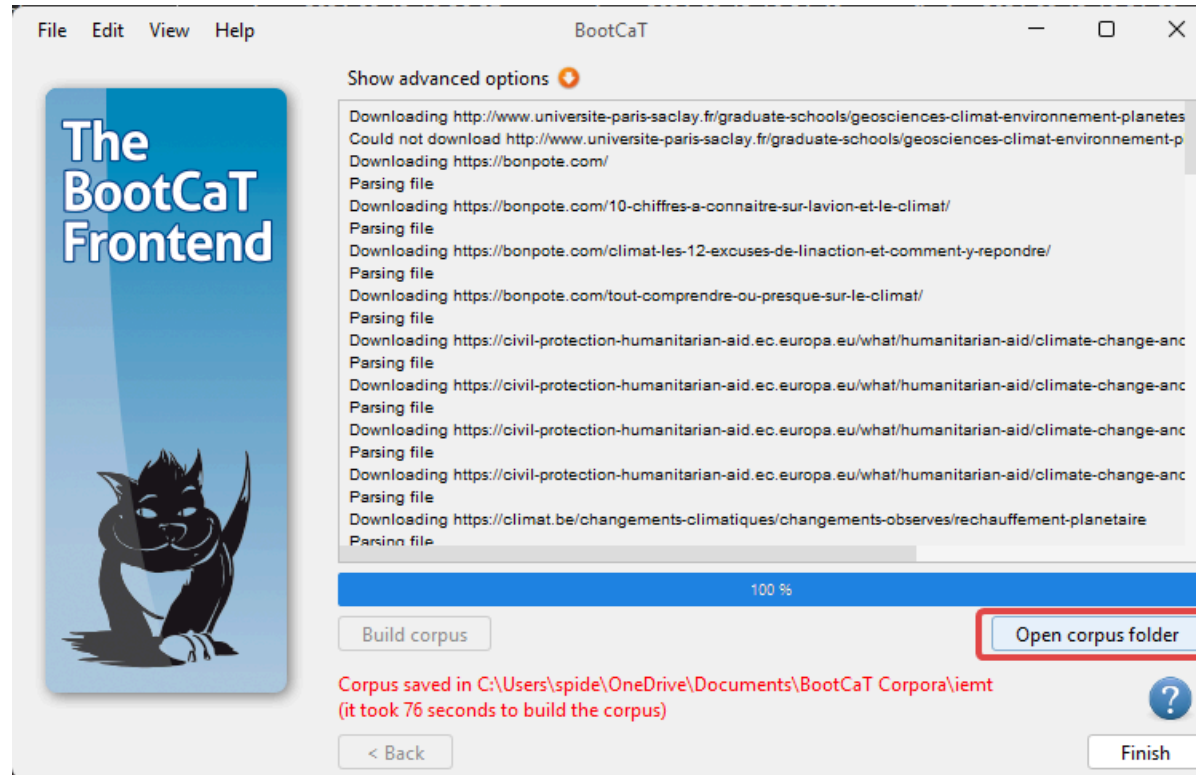
On peut filtrer les adresses URL :

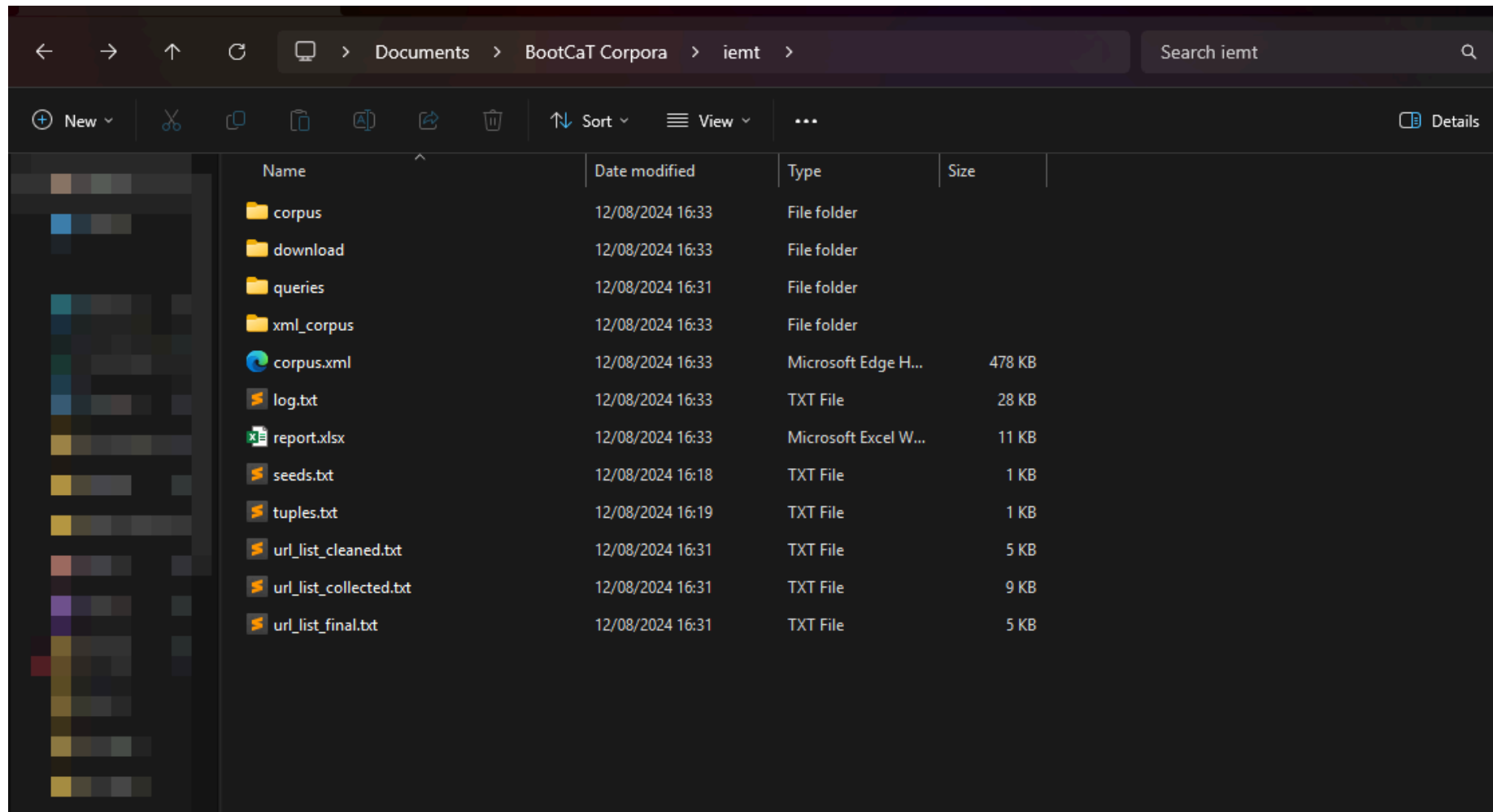


Il n'y a plus qu'à compiler le corpus !



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





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