

Corpus *Réfugié·es*

“Je débute une thèse sur les réfugié·es à travers le monde au 21^e siècle. Je souhaite construire une bibliographie sur le sujet, distinguer la part de documents traitant des réfugié·es climatiques et des réfugié·es politiques et enfin identifier leur origine géographique.”

Description générale

Pour répondre à cette problématique, il faut constituer un corpus de documents traitant des réfugié·es grâce au réservoir Istex. Ce corpus va ensuite être exploré et enrichi dans Lodex.

Objectif : repérer les études sur les réfugié·es et débiter leur analyse.

Outil de TDM utilisé : [Lodex](#) et les [web services associés](#).

Contraintes imposées par l'outil : Istex propose un format de sortie adaptée à Lodex. Pour utiliser les web services qui m'intéressent, il faut vérifier que les documents sont en anglais et contiennent des résumés.

Exercice 1 – Construire une requête Istex

☐ **Étape 1 :** Se rendre sur [Istex Search](#).

☐ **Étape 2 :** Rechercher les formes anglaises et françaises *refugee*, *réfugié*, *asylum seeker* et *demandeur d'asile*. Pour une aide sur la syntaxe des requêtes consulter les [Astuces de recherche](#) ? (à droite dans Istex Search).

☐ **Étape 3 :** Limiter le bruit et le silence.

⇒ Pour supprimer le silence : rechercher les variantes (singulier / pluriel et féminin / masculin) de chacun des termes de la requête.

⇒ Pour supprimer le bruit, on peut préciser les champs les plus à même de renvoyer des résultats pertinents (soit le *titre* et les *mots-clés d'auteur·ices* dont les

dénominations sont *title* et *subject.value*). La requête prend alors la forme *champ:()*. La liste des champs est accessible dans [la recherche assistée](#).

⇒ Pour limiter le bruit, on doit s'assurer de ne sélectionner que des articles de recherche (filtre : *Type de contenu* ; nom technique : *genre*).

□ **Étape 4 :** Répondre aux contraintes scientifiques et techniques.

⇒ Pour répondre à la problématique, il faut limiter les dates de publication au 21^e siècle.

⇒ Pour répondre aux contraintes imposées par l'outil, il faut s'assurer que les documents sont en **anglais** et possèdent un **résumé**.

Quelle est la taille de votre corpus ?

Exercice 2 – Chargement des données et création du site Lodex

□ Étape 1 : Télécharger le corpus.

- ⇒ Extraire le corpus *Réfugié-es* en utilisant l'équation corrigée et en choisissant le format adapté pour un import dans Lodex.

□ Étape 2 : Importer le corpus dans Lodex.

- ⇒ Se rendre sur votre instance Lodex : se connecter avec votre nom d'utilisateur et votre mot de passe.
- ⇒ Aller dans l'interface administrateur en cliquant sur *Voir plus > Admin*.
- ⇒ Importer le corpus en glissant le fichier .zip téléchargé **sans décompression préalable**.
- ⇒ Choisir le loader¹ *ZIP Istex Search*.
- ⇒ Cliquer sur *Importer les données*.

□ Étape 3 : Importer le modèle.

- ⇒ Importer le modèle fourni, cliquer sur le menu en haut à droite *Modèle > Importer un modèle*.
- ⇒ Configurer l'affichage : cliquer sur le menu en haut à droite *Configuration*. Choisir le thème *ISTEX – Thème ISTEX (restreint)*². Cliquer sur *Sauvegarder*.
- ⇒ Publier votre site en cliquant sur *Publier* en haut à droite.
- ⇒ Cliquer sur l'icône en forme d'œil pour voir le résultat.
- ⇒ Explorer les différents graphiques à partir de l'onglet *Graphiques* en bas à gauche. Consulter quelques ressources grâce à l'onglet *Recherche*.

□ Étape 4 : Modifier un champ existant dans Lodex

- ⇒ Depuis l'administration de l'instance se rendre sur l'onglet *Affichage*. Dans le menu de gauche, cliquer sur *Page d'accueil*.
- ⇒ Cliquer sur le champ *Date de création*.
- ⇒ Modifier le contenu du champ *Valeur arbitraire* avec la date du jour.
- ⇒ Cliquer sur le bouton *Sauvegarder*.
- ⇒ Sur la page d'accueil, vérifier que la date a bien été mise à jour.

¹ Un loader est un script d'adaptation du fichier à Lodex. Il dépend du format de fichier fourni en entrée.

² Le texte présent dans le bandeau ainsi que les liens sont modifiables depuis la configuration.

Exercice 3 – Premiers pas vers le TDM

□ Étape 1 : Extraire les mots-clés des résumés via le web service [Teeft](#).

- ⇒ Aller dans *Données > Enrichissements* et cliquer sur *+ Ajouter*.
- ⇒ Donner le nom *Mots-clés (WS)*, aller chercher l'url du web service **Teeft eng** (onglet *Indexation*) dans le catalogue en cliquant sur le bouton vert à droite du champ *URL du web service*.
- ⇒ Choisir *Résumé* dans la *Colonne de la source*, cliquer sur *Sauvegarder*. Cliquer enfin sur *Lancer*.

□ Étape 2 : Créer un diagramme en barres.

Dans Lodex, avant de faire un graphique, il est nécessaire de déclarer la colonne comme une ressource.

- ⇒ Aller dans *Affichage > Ressource principale*, cliquer sur *+ Nouveau champ*. Pour paramétrer ce nouveau champ : dans *Étiquette* nommer le champ *Mots-clés*, sélectionner *Colonne(s) existante(s)* et aller chercher la colonne nommée *Mots-clés (WS)*.
- ⇒ Le web service Teeft renvoie, pour chaque mot-clé, des informations incluant sa fréquence et sa spécificité. Dans notre cas, nous souhaitons uniquement récupérer les termes eux-mêmes : cliquer sur *Ajouter une opération*, sélectionner GET, puis indiquer "term" dans le champ *path*. Cliquer sur *Sauvegarder*.
- ⇒ Pour créer le graphique : aller dans *Affichage > graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Mots-clés les plus représentés dans les résumés* dans le champ *Étiquette*. Choisir la routine *distinct-by* puis choisir le champ sur lequel la routine va s'appliquer grâce au menu déroulant (soit le champ *Mots-clés*).
- ⇒ Enfin, dans *Affichage*, choisir le format *Diagramme en barres* en filtrant les résultats : dans *Paramètres des Données* mettre *valeur minimum à afficher* à 5 et choisir le tri *Descendant*. Cliquer sur *Sauvegarder*.

□ Étape 3 : Créer une nouvelle facette *Mots-clés*.

- ⇒ Depuis l'administration de l'instance se rendre sur l'onglet *Affichage*. Dans le menu de gauche cliquer sur *Recherche et facette*.
- ⇒ Dans la section *Facettes* sélectionner *Mots-clés*.
- ⇒ Sur la page d'accueil, depuis l'onglet recherche, vérifier que la facette a bien été créée.

Combien de documents comportent le terme *asylum seekers* et *environmental refugees* ?

Exercice 4 – Cartographie

□ Étape 1 : Récupérer les noms de lieux mentionnés.

- ⇒ Les données contiennent des enrichissements déjà fournis par le réservoir Istex. C'est le cas des noms de lieux renseignés dans la colonne *Entités nommées (Unitex)* avec les autres entités nommées.
- ⇒ Pour créer une colonne contenant uniquement ces noms de lieux, aller dans *Données > Enrichissements* et cliquer sur + *Ajouter*. Donner le nom *Entités nommées de géographie*, cliquer sur *Mode avancé* et coller le code ci-après, cliquer sur *Sauvegarder*. Cliquer enfin sur *Lancer*.

```
[assign]
path = value
value = get("value.Entités nommées (Unitex).placeName").compact()
```

Remarque : le mode avancé offre une meilleure flexibilité pour transformer les données. Il s'agit de code en Lodash, une librairie Javascript (cf. [tutoriel](#)). Les transformations les plus usuelles sont disponibles via [ce lien](#).

□ Étape 2 : Uniformiser les noms de lieux pour créer une cartographie.

- ⇒ Pour construire la carte, nous avons besoin des coordonnées du pays, retournées par le web service [Associer un terme au vocabulaire Pays et Subdivisions](#). Aller dans *Données > Enrichissements > + Ajouter*, puis donner le nom *Pays et subdivisions (WS)*, en utilisant l'URL de ce web service <https://loterre-resolvers.services.istex.fr/v1/identify?loterreID=9SD> sur la colonne nouvellement créée *Entités nommées de géographie*. Cliquer sur *Sauvegarder*. Cliquer enfin sur *Lancer*.

□ Étape 3 : Créer une carte en fonction des pays mentionnés dans les articles.

- ⇒ Dans *Affichage > Ressource principale*, créer une ressource *Coordonnées des pays cités* à partir de l'enrichissement *Pays et subdivisions (WS)*. Après avoir sélectionné la colonne, cliquer sur *Ajouter une opération* pour appliquer une transformation GET, renseigner *latitude;longitude* dans le *path*. Dans *Affichage* décocher *Visible*. Cliquer sur *Sauvegarder*.
- ⇒ Pour créer le graphique : aller dans *Affichage > graphiques* cliquer sur + *Nouveau champ*, nommer le graphique en renseignant *Cartographie des pays cités* dans le champ *Étiquette*. Choisir la routine *distinct-by* puis choisir le champ sur lequel la routine va s'appliquer ressource *Coordonnées des pays cités*. Enfin, dans *Affichage*, choisir le format *Carte OpenStreetMap*. Cliquer sur *Sauvegarder*.

□ **Étape 4 :** Créer une nouvelle facette *Pays cités*.

- ⇒ Dans *Affichage > Ressource principale*, créer une ressource *Pays cités* à partir de l'enrichissement *Pays et subdivisions (WS)*. Après avoir sélectionné la colonne, cliquer sur *Ajouter une opération* pour appliquer une transformation GET, renseigner *prefLabel@fr* dans le *path*. Dans *Affichage* choisir *Texte-Liste de valeurs*. Cliquer sur *Sauvegarder*.
- ⇒ Dans le menu de gauche cliquer sur *Recherche et facette*.
- ⇒ Dans la section *Facettes* sélectionner *Pays cités*.
- ⇒ Sur la page d'accueil, depuis l'onglet recherche, vérifier que la facette a bien été créée.

Exercice 5 – Premier précalcul

□ **Étape 1 :** Extraction des [thématiques du corpus](#)³.

- ⇒ Dans *Données > Précalculs* cliquer sur + *Ajouter*, lancer le précalcul *ldaSegment* sur la colonne *Résumé*.
- ⇒ **Les [précalculs](#) peuvent directement être utilisés pour faire des graphiques.** Aller dans *Affichage > graphiques* cliquer sur + *Nouveau champ*, puis créer le graphique *Thématiques extraites du corpus* à partir du précalcul *lda-segment*. Choisir la routine *segments-precomputed-nofilter*. Enfin, dans *Affichage*, choisir le format *diagramme à barres groupées*.

³ Le LDA ou *Latent Dirichlet Allocation* est un algorithme de topic modeling. Son objectif est de découvrir automatiquement les thèmes présents dans les documents.