

Corpus Fascism

“Comment les sciences humaines et sociales ont-elles étudié le fascisme depuis 1945, et quelles évolutions thématiques peut-on observer dans la littérature scientifique internationale ?”

Description générale

Pour répondre à cette problématique, il faut constituer un corpus de publications scientifiques traitant du fascisme. Ce corpus va ensuite être exploré et enrichi dans Lodex.

Objectif : repérer les études sur le fascisme et débuter 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 *fascism* et *fascist*. 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) 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*, le *résumé* et les *mots-clés d'auteur·ices* dont les dénominations sont *title*, *abstract* et *subject.value*). La requête prend alors la forme *champ:()*. La liste des champs est accessible dans [la recherche assistée](#).

Étape 4 : Répondre aux contraintes scientifiques et techniques.

- ⇒ Pour répondre à la problématique, il faut s'assurer que les documents appartiennent aux sciences sociales (*categories.scopus:"1 - Social Sciences"*) et ont été publiés après 1945.

- 
- ⇒ 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 ?

Étape 5 : Télécharger le corpus.

- ⇒ Extraire le corpus en utilisant [l'équation corrigée](#) et en choisissant le format adapté pour un import dans Lodex.

Exercice 2 – Crédit du site Lodex et ajout de données OpenAlex

Étape 1 : Importer le corpus dans Lodex.

- ⇒ Se rendre sur votre instance Lodex : se connecter avec votre nom d'utilisateur et votre mot de passe (accessibles depuis [le site des supports](#)).
- ⇒ 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 2 : Importer le modèle.

- ⇒ Importer [le modèle fourni](#), cliquer sur le menu en haut à droite *Modèle > Importer un modèle*.
- ⇒ 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*.

Que constatez-vous en visualisant le graphique des années de publication ?

Étape 3 : Ajouter des données provenant d'une autre source (OpenAlex). Il est possible d'importer des données de sources variées, il faut au préalable vérifier leur compatibilité (ex. la colonne *Titre* doit avoir le même nom, pas *title* ou *Titre de l'Article*).

- ⇒ Aller dans l'interface administrateur en cliquant sur *Voir plus > Admin*.
- ⇒ Dans *Données*, cliquer sur *Ajouter*.
- ⇒ Glisser le fichier *.jsonl* contenant [les données OpenAlex](#).

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

- 
- ⇒ Choisir le loader JSON - format *text JSON Lines*.
 - ⇒ Cliquer sur *Importer les données*.
 - ⇒ Consulter de nouveau le graphique des années de publication.

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 (arbitraire) *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*.
- ⇒ Dans *Étiquette*, nommer le champ *Mots-clés (WS)*, 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. Nous souhaitons uniquement récupérer les termes.

- ⇒ 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 (WS)*).
- ⇒ 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 *Confirmer* puis *Sauvegarder*.

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

- ⇒ Depuis l'administration de l'instance, se rendre sur l'onglet *Affichage*. Dans le menu de gauche cliquer sur *Recherche et facettes*.
- ⇒ Dans la section *Facettes*, cocher *Mots-clés (WS)*.

- 
- ⇒ Sur la page d'accueil, depuis l'onglet recherche, vérifier que la facette a été créée.
 - En quelle année apparaît le plus le mot-clé Ukraine ?**

Exercice 4 – Premier précalcul

Étape 1 : Extraction des [thématisques du corpus](#)².

- ⇒ Dans *Données > Précalculs* cliquer sur + Ajouter, donner le nom *lda*.
- ⇒ Aller chercher l'url du web service **ldaSegment** 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 sur Lancer.

Étape 2 : Créer un diagramme à barres groupées

Contrairement aux enrichissements, les [précalculs](#) peuvent directement être utilisés pour faire des graphiques.

- ⇒ Aller dans *Affichage > Graphiques*, cliquer sur + Nouveau champ, puis créer le graphique *Thématisques extraites du corpus* à partir du précalcul *lda-segment*. Choisir la routine *segments-precomputed-nofilter*.
- ⇒ Dans *Affichage*, choisir le format *Groupement de diagrammes à barres*.
- ⇒ Cliquer sur *Confirmer* puis *Sauvegarder*.

Quels topics voyez-vous émerger ?

Exercice 5 – Pour aller plus loin : identification des entités nommées

Étape 1 : Extraire les entités nommées pour déterminer les noms des personnes les plus cités.

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

Étape 2 : Créer un graphe à bulle.

² Le LDA ou *Latent Dirichlet Allocation* est un algorithme de topic modeling. Son objectif est de découvrir automatiquement les thèmes présents dans les documents.

- 
- ⇒ Aller dans *Affichage > Ressource principale*, cliquer sur + *Nouveau champ*.
 - ⇒ Dans *Étiquette*, nommer le champ *Personnes citées (WS)*, sélectionner *Colonne(s) existante(s)* et aller chercher la colonne nommée *Entités nommées (WS)*.

Le web service *entityTag* extrait, pour chaque document, des noms de personnes (PER), de localisations (LOC) ou d'organismes (ORG). Nous souhaitons récupérer les PER.

- ⇒ Cliquer sur *Ajouter une opération*, sélectionner GET, puis indiquer PER dans le champ *path*. Sélectionner ensuite l'opération UNIQ pour dédoublonner les valeurs, puis l'opération ARRAY qui permet transformer les données en tableau.
- ⇒ Cliquer sur *Sauvegarder*.
- ⇒ Dans *Affichage > Graphiques*, cliquer sur + *Nouveau champ*, nommer le graphique en renseignant *Personnes citées (WS)* 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 *Personnes citées (WS)*).
- ⇒ Dans *Affichage*, choisir le format *Graphe à bulles*. Dans *Paramètres des Données* choisir le tri *Descendant*.
- ⇒ Cliquer sur *Confirmer* puis *Sauvegarder*.

Quels noms de personnes ressortent le plus ? Quelle(s) anomalie(s) constatez-vous ?