

INFORMATION NUMÉRIQUE

Enjeux et Pratiques

Licence 3 / UE502 – Accès expert à l'information – 2025-2026

Découverte : les web services et LODEX

valerie.bonvallot@inist.fr

Sommaire : web services et LODEX

1. TDM et web services : quelques définitions
2. Navigation dans le site ISTEX TDM
3. Cas pratique : lancer des web services à partir de Lodex
4. Exercices sous LODEX

Découverte : web services LODEX

1. TDM et web services

1. TDM et web services

TDM : Text and data mining - Définitions fouille de textes et de données

« Toute technique *d'analyse automatisée* visant à analyser des textes et des données sous une forme numérique afin d'en *dégager des informations*, ce qui comprend, à titre non exhaustif, des *constantes*, des *tendances* et des *corrélations* »

(Ordonnance du 24 novembre 2021,
directive sur le droit d'auteur
[ouvrirlascience.fr](https://www.ouvrirlascience.fr))

Ensemble des méthodes et des traitements informatiques qui consistent à *analyser le sens des textes* en langage naturel pour en donner une *représentation utilisable* par les humains et les ordinateurs.

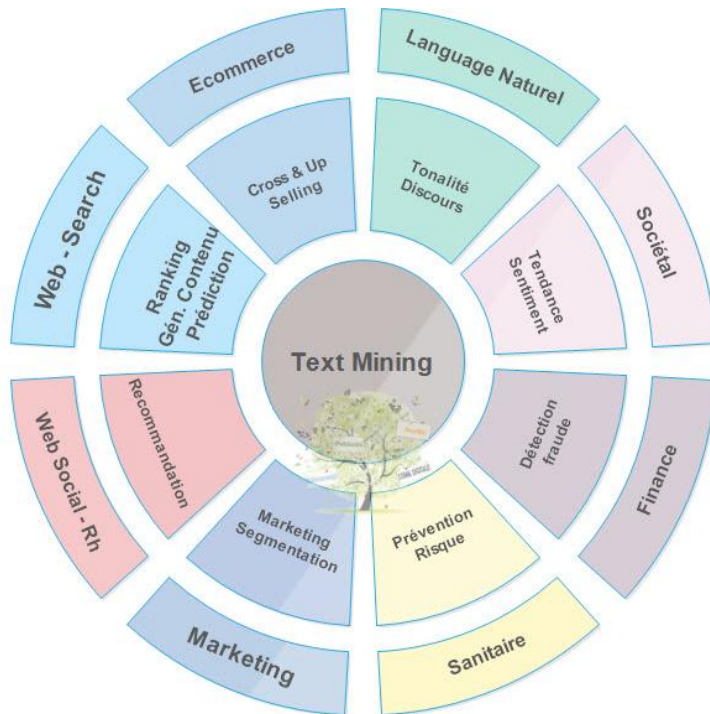
Données → Connaissances

C'est une spécialisation de la fouille de données (data mining) qui fait appel aux *méthodes de l'Intelligence Artificielle*, du *Traitement Automatique des Langues* et des *Statistiques*

1. TDM et web services

TDM : Text and data mining - Utilisation fouille de textes et de données

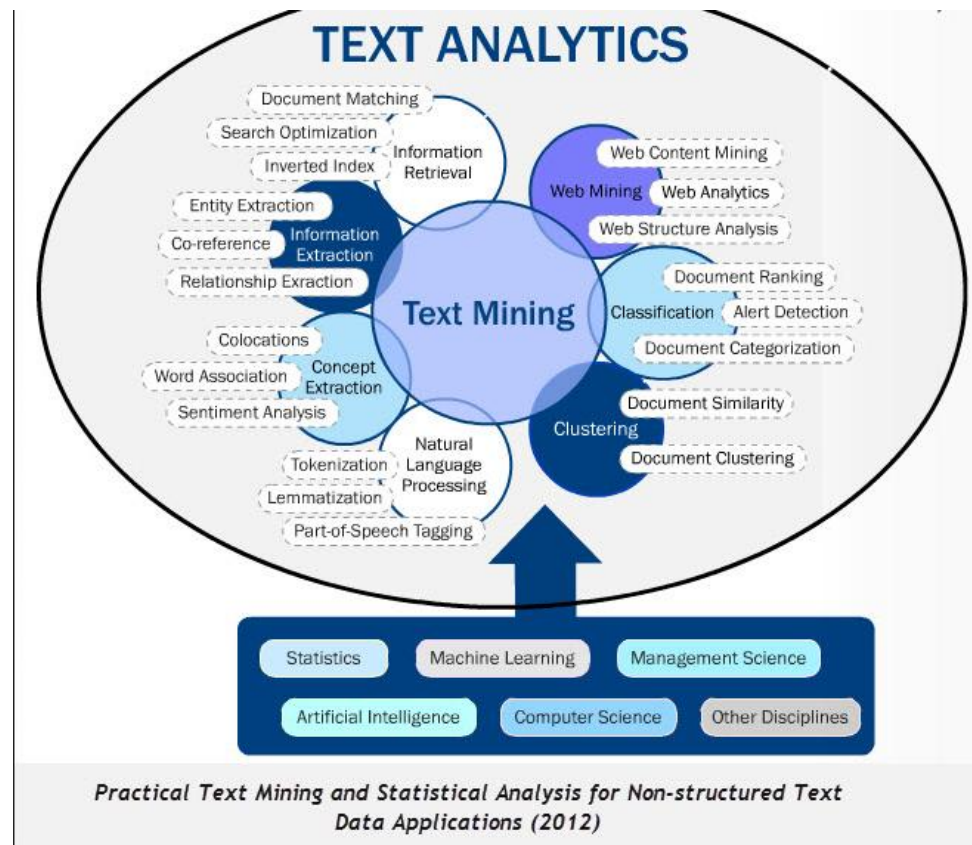
ontologie
terminologie
extraction d'information
annotation
production automatique
résumé automatique
recherche documentaire
analyse de sentiments



- Faire de la bibliométrie
- Extraire de l'information pertinente
- Répondre à une question (recherche d'information)
- Analyser de gros volumes de textes
- Identification de thèmes
- Détecter des sentiments dans les textes
- Construire des résumés automatiques
- Désambiguïser des lieux, des personnes...
- Faire des systèmes de recommandations
- Détecter des « fake news »
- Trier des mails, des textes

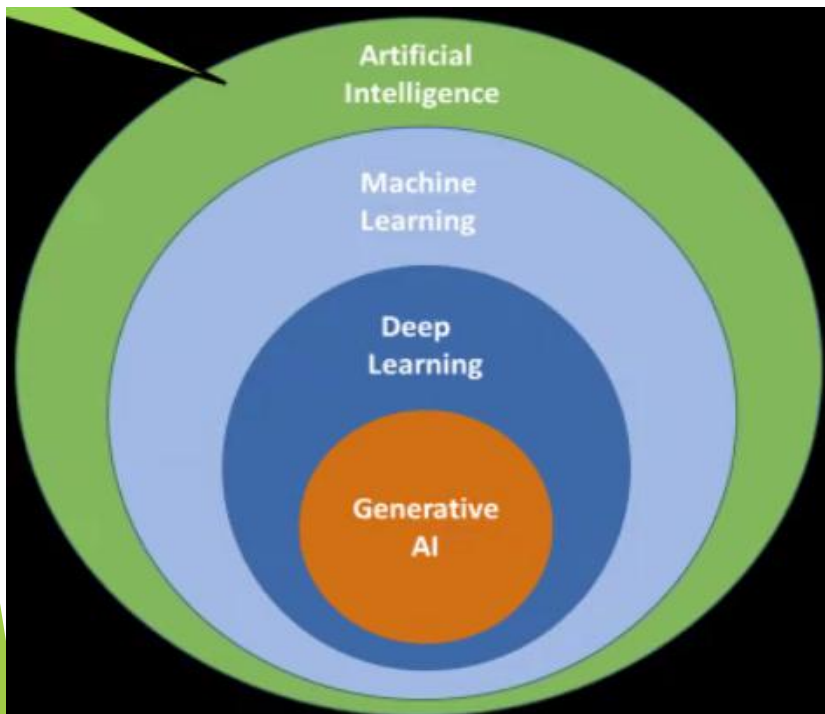
1. TDM et web services

TDM : Text and data mining - Technique
fouille de textes et de données



1. TDM et web services

TDM : Text and data mining et Intelligence artificielle
fouille de textes et de données



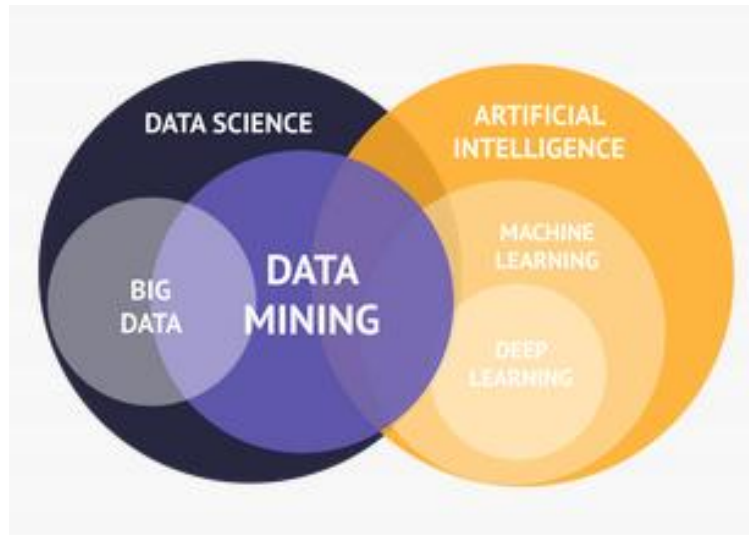
- IA : un programme qui va permettre de réaliser une tâche humaine
 - Apprentissage automatique (Large Language Models : stat, prédiction)
 - Apprentissage profond* et réseaux de neurones (transformeurs)
 - IA générative

Mohand Boughanem

¹*L'apprentissage profond ou apprentissage en profondeur (en anglais : *deep learning*, *deep structured learning*, *hierarchical learning*) est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur, du traitement automatisé du langage

1. TDM et web services

TDM : Text and data mining et Intelligence artificielle
fouille de textes et de données



<https://www.metron.energy/blog/interview-data-science-industry/>

1. TDM et web services

TDM : Text and data mining - Contexte
fouille de textes et de données



Big data

3V : Volume, Vitesse et Variété
4V : + Valeur
5V : + Véracité

TAL IA

Puissance de calcul
Algorithme

Science ouverte

Dispositions légales

1. TDM et web services

Web services

interface et protocole d'échange en ligne de données

1 WS = 1 tâche, un traitement spécifique

Peu de compétences informatiques (transparence du langage, pas d'installation)

Paramétrage minimal

Données issues de différentes sources

Découverte : web services LODEX

2. Navigation dans ISTEX TDM

2. Navigation dans ISTEEX TDM

Un site en ligne : [ISTEX TDM](https://services.istex.fr)

<https://services.istex.fr>

The screenshot displays the ISTEEX TDM website interface. At the top, a dark blue header contains the text "ISTEX TDM" and "Les services Istex pour la fouille de textes". Below this, a light blue section titled "Rechercher un web service" features a search bar with the placeholder text "Tapez ici votre recherche, p.ex. : Classification" and a "RECHERCHER" button. To the right of the search bar, a vertical blue sidebar contains the text "Trouvez un web service correspondant à vos besoins", followed by a paragraph explaining the service, and a counter showing "49" web services available. Below the search bar, a grid of six service cards is displayed, each with a gear icon, a title, and a description. The cards are: 1. "hiddenTextDetect" (Texte intégral) - DÉTECTION DE TEXTE CACHÉ DANS UN PDF; 2. "OALDocTypeClass" (Éléments catalographiques) - CLASSIFICATION DE DOCUMENTS OPENALEX PAR TYPE DE DOCUMENT; 3. "Grobid" (Texte intégral) - EXTRACTION ET STRUCTURATION DE PUBLICATION SCIENTIFIQUE AU FORMAT PDF; 4. "dataGraph" (Résumés - Texte intégral) - GRAPHE DE MOTS CLÉS; 5. "TranslITaL" (Adresses et affiliations - Auteurs - Éléments catalographiques - Citations - Résumés - Texte intégral) - TRANSLITTÉRATION EN CARACTÈRES LATINS; 6. "softwareTag" (Résumés - Texte intégral) - EXTRACTION DE NOMS DE LOGICIELS. Each card has a right-pointing arrow. At the bottom of the grid, a button labeled "VOIR TOUS LES SERVICES" is visible.

ISTEX TDM
Les services Istex pour la fouille de textes

Rechercher un web service

Tapez ici votre recherche, p.ex. : Classification **RECHERCHER**

Trouvez un web service correspondant à vos besoins

Nous développons et mettons à votre disposition des web services de TDM (Text and Data Mining) faciles à mettre en œuvre, couplés à un outil de création de tableaux de bord dynamiques.

Actuellement **49** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

hiddenTextDetect
DÉTECTION DE TEXTE CACHÉ DANS UN PDF

OALDocTypeClass
CLASSIFICATION DE DOCUMENTS OPENALEX PAR TYPE DE DOCUMENT

Grobid
EXTRACTION ET STRUCTURATION DE PUBLICATION SCIENTIFIQUE AU FORMAT PDF

dataGraph
GRAPHE DE MOTS CLÉS

TranslITaL
TRANSLITTÉRATION EN CARACTÈRES LATINS

softwareTag
EXTRACTION DE NOMS DE LOGICIELS

VOIR TOUS LES SERVICES

2. Navigation dans ISTEEX TDM

Un site en ligne : [ISTEX TDM](https://services.istex.fr)

<https://services.istex.fr>

ISTEX TDM
Les services Istex pour la fouille de textes

Rechercher un web service

Tapez ici votre recherche, p.ex. : Classification

RECHERCHER

Texte intégral
hiddenTextDetect
Détection de texte caché dans un PDF

Éléments catalographiques
OALDocTypeClass
Classification de documents OpenAlex par type de document

Texte intégral
Grobid
Extraction et structuration de publication scientifique au format PDF

Résumés - Texte intégral
dataGraph
Graphe de mots clés

Adresses et affiliations - Auteurs - Éléments catalographiques - Citations - Résumés - Texte intégral
TranslITAL
Translittération en caractères latins

Résumés - Texte intégral
softwareTag
Extraction de noms de logiciels

Trouvez un web service correspondant à vos besoins

Nous développons et mettons à votre disposition des web services de TDM (Text and Data Mining) faciles à mettre en œuvre, couplés à un outil de création de tableaux de bord dynamiques.

Actuellement **49** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

VOIR TOUS LES SERVICES

OBJET TRAITÉ

☐ Adresses et affiliations (10)
☐ Auteurs (3)
☐ Éléments catalographiques (6)
☐ Citations (3)
☐ Résumés (24)
☐ Texte intégral (28)

LANGUES (3)

TRAITEMENT (7)

☐ Classification (10)
☐ Extraction d'entités nommées (11)
☐ Homogénéisation (9)
☐ Indexation (7)
☐ Traitement automatique du langage (4)
☐ Prétraitement (6)
☐ Validation (5)

TYPE DE DONNÉES (2)

PRÉSENCE SUR TDM FACTORY (2)

☐ Oui (21)
☐ Non (28)

hiddenTextDetect
Détection de texte caché dans un PDF

Ce web service analyse un fichier PDF pour détecter la présence de texte caché ou invisible — c'est-à-dire du texte présent dans le document mais qui n'est pas visible à l'œil nu lors de la lecture humaine.

Grobid
Extraction et structuration de publication scientifique au format PDF

Ce service extrait le texte d'une publication scientifique au format PDF et le structure au format XML-TEI avec l'API de Grobid.

TranslITAL
Translittération en caractères latins

TranslITAL est un ensemble de web services qui permettent la translittération de caractères non latins en caractères latins. Il est le fruit d'un projet de recherche conjoint entre la Bibliothèque universitaire des langues et civilisations (BULAC) et l'Institut national des...

OALDocTypeClass
Classification de documents OpenAlex par type de document

Ce web service prend en entrée un identifiant OpenAlex et vérifie si le document associé est un document de recherche ou non. Ce service a été développé par Nick Haupka (Université d'Ottlingen, Allemagne) et financé par : the Federal...

dataGraph
Graphe de mots clés

Ce web service génère un graphe de mots à partir d'un corpus de résumés en anglais ensemble de mots-clés déjà extraits.

softwareTag
Extraction de noms de logiciels

Ce web service détecte des noms de logiciels sur des textes en anglais.

softwareTag - Extraction de noms de logiciels

Accueil > Web services > Extraction de noms de logiciels

Malgré les contrôles qualité que nous effectuons avant de mettre nos web services en production, les outils d'IA peuvent commettre des erreurs. Nous vous recommandons de vérifier les informations importantes.

Description

Utilisation

Niveau d'utilisation : Débutant
Niveau de validation : Expérimental

Objectif

Ce web service détecte des noms de logiciels sur des textes en anglais.

Méthode

Le modèle utilisé par ce web service est un réseau de neurones entraîné en utilisant Flair sur le jeu de données [softwareTag](#). Le code pour entraîner et évaluer le modèle est disponible à [ce lien](#).

Métriques

Précision : Le web service fonctionne uniquement sur du texte en anglais.

Variante

Vérifier si le logiciel extrait est open source

2. Navigation dans ISTEX TDM

Indexation

Teeft Extraction de termes d'un texte via Teeft

Le service web Teeft extrait, par défaut, les 5 termes les plus spécifiques d'un texte en anglais ou en français. Il permet ainsi d'avoir une idée de ce dont il est question dans le texte.

Avant

"The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 20 January 2020, and later a pandemic on 11 March 2020. As of 2 April 2021, more than 129 million cases have been confirmed, with more than 2.82 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history."

Après

"severe acute respiratory syndrome coronavirus2",
"international concern",
"ongoing global pandemic",
"coronavirus disease",
"covid-19",
"december",
"wuhan",
"coronavirus pandemic",
"deadly pandemic",
"covid-19 pandemic"

2. Navigation dans ISTEX TDM

Détection de la langue

langDetect
Détection de la langue d'un
texte

Le web service détecte la langue d'un document
texte.

Avant

"User experience design (UXD, UED, or XD) is the process of supporting user behavior[1] through usability, usefulness, and desirability provided in the interaction with a product.[2] User experience design encompasses traditional human-computer interaction (HCI) design and extends it by addressing all aspects of a product or service as perceived by users. Experience design (XD) is the practice of designing products, processes, services, events, omnichannel journeys, and environments with a focus placed on the quality of the user experience and culturally relevant solutions."

Après

<> "en"

2. Navigation dans ISTEX TDM

Classification supervisée

pascalFrancisClass Classification en domaines scientifiques Pascal-Francis

Le web service classe automatiquement des documents scientifiques en anglais dans le plan de classement Pascal (Sciences, Techniques et Médecine) ou Francis (Sciences Humaines et Sociales). Après traitement, chaque document possède un domaine scientifique homogène, dans la mesure où les...

Avant

"Rhesus Monkey Model Self Injury effect
Relocation Stress Behavior Neuroendocrine
Functionbackground self injurious behavior
SIB disorder many individual clinical
nonclinical population state stress arousal
longitudinal datum relationship increase
(...)
significant stressor rhesus macaque stressor
increase self behavior sleep disturbance
monkey SIB finding life stress SIB human
disorder HPA axis result potential role CBG
long term neuroendocrine response major
stressor"

Après



"003": "Sciences humaines et sociales",
"770": "Psychologie. Psychanalyse. Psychiatrie.",
"770D": "Psychopathologie. Psychiatrie."

Découverte : web services LODEx

3. Cas pratique

3. Cas pratique

Lodex

Linked Open Data EXperiment

Transformer

Ses données (CSV, TXT, JSON) en site web

Attribuer

Des identifiants pérennes (ARK)

Exporter

Ses données dans plusieurs formats



Explorer

Ses données à travers des graphiques dynamiques

Enrichir

Ses données avec des web-services

Aligner

Ses données avec des référentiels

3. Cas pratique

Enrichissements et précalculs

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

Données

Affichage

Dépublier

Nom

1

Mode avancé

URL du web service

2

Colonne de la source

3

Sous-chemin

4

ANNULER

SAUVEGARDER

Aperçu de la valeur*

Aucune donnée

* valeur de la colonne source ou de la règle avancée

ISTEX TDM

3. Cas pratique

Enrichissements

traitements document par document au sein de Lodex (web service **synchrone**).

Filtre Istex TDM : Type de données “Documents”

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

Données

Enrichissements

Précalculs

Ressources cachées

1

Repérage des espèces

LANCER

Aperçu de la valeur*

Résumé

Statut : Non démarré

Mode avancé

URL du web service

2

https://irc3-species.services.istex.fr/v1/irc3sp

Colonne de la source

Résumé

Sous-chemin

SUPPRIMER

ISTEX TDM

Les services Istex pour la fouille de textes

speciesTag - Extraction de noms d'espèces

Description

Utilisation

Cas d'usage

URL DU WEB SERVICE À RENSEIGNER DANS LODEX ENRICHISSEMENT EST :

https://irc3-species.services.istex.fr/v1/irc3sp

TOUS

AFFILIATION

CLASSIFICATION

ENTITÉS NOMMÉES

ENTITÉS NOMMÉES

GÉOGRAPHIE

HOMOGENÉISATION

HOMOGENÉISATION

INDEXATION

MÉTADONNÉES

PRÉTRAITEMENT

TRAITEMENT AUTOMATIQUE DE LA LANGUE

VALIDATION

AUTRE

addressSplit - Décomposition d'une adresse

Découpe une adresse au format texte en plusieurs champs.

Associer un terme au vocabulaire Pays et Subdivision

Associe un pays ou subdivision au vocabulaire Loterie correspondant.

Associer un terme au vocabulaire des communes de France

Repère dans un texte des termes présents dans le thésaurus Communes de France et récupère le(s) concept(s) associé(s).

astroTag - Extraction d'entités nommées en astronomie

Détecte des entités nommées en astronomie sur des textes en anglais et les répartit entre 16 classes prédéfinies

authorDistinct - Désambiguïsation d'auteurs via ORCID

Retrouve un auteur à partir d'un certain nombre d'éléments connus le concernant, comme les nom et prénom, des titres de publications, ou encore des co-auteurs.

authorDistinct - Désambiguïsation d'auteurs via ORCID

Retrouve un auteur à partir d'un certain nombre d'éléments connus le concernant, comme les nom et prénom, des titres de publications, ou

20

ANNULER

3. Cas pratique

Enrichissements

traitements document par document au sein de Lodex (web service **synchrone**).

Filtre Istex TDM : Type de données "Documents"

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

Données

AFFICHAGE

DÉPUBLIER

Nom

Repérage des espèces

1

5

▶ LANCER

Statut :

Non démarré

VOIR LES LOGS

Mode avancé

URL du web service

https://irc3-species.services.istex.fr/v1/irc3sp

2

📄

Colonne de la source

Résumé

3

▼

Sous-chemin

▼

SUPPRIMER

ANNULER

SAUVEGARDER

📷 Aperçu de la valeur*

Résumé

"During the International Tapir Symposium 16–21 Oct 2011, the conservation of Baird's tapir (Tapir...

"We collected data on habitat use and locomotion of the François' langur (Trachypitecus francoisi...

"Aim: Climate change assessments are largely based on correlative species distribution models...

 Aperçu de la valeur*

Résumé

"During the International Tapir Symposium 16–21 Oct 2011, the conservation of Baird's tapir (Tapir...

"We collected data on habitat use and locomotion of the François' langur (*Trachypithecus francoisi*...

"Aim: Climate change assessments are largely based on correlative species distribution models...

Précalculs

Le résultat obtenu pour chacun des documents dépend des

autres (web service **asynchrone**)

Filtre Istex TDM : Type de données “Corpus”

22

3. Cas pratique

Précalculs

traitements de l'ensemble des documents en dehors de Lodex.
Le résultat obtenu pour chacun des documents dépend des autres (web service **asynchrone**)

Filtre Istex TDM : Type de données "Corpus"

The screenshot displays the LODEX web interface for pre-calculations. The sidebar on the left highlights the 'Précalculs' option. The main content area has a green header bar with 'LODEX', 'DONNÉES', and 'AFFICHAGE' tabs. The 'Précalculs' form contains the following elements:

- Nom ***: A text input field containing 'noiseDetect' (labeled 1).
- Statut**: A dropdown menu showing 'Non démarré'.
- URL du web service ***: A text input field containing 'https://text-clustering.services.istex.fr/v1/noise-lodex' (labeled 2).
- Colonne(s) source(s) ***: A dropdown menu showing 'Résumé' (labeled 3).
- Buttons**: 'LANCER' (labeled 5), 'VOIR LES LOGS', 'SUPPRIMER', 'RETOUR', and 'SAUVEGARDER' (labeled 4).

On the right side, there is a preview box titled 'Aperçu de la valeur*' showing a sample text snippet:

Résumé

"During the International Tapir Symposium 16-21 Oct 2011, the conservation of Baird's tapir (Tapir...)

"We collected data on habitat use and locomotion of the François' langur (Trachypithecus francoisi...)

"Aim: Climate change assessments are largely based on correlative species distribution models..."

Découverte : web services LODEX

4. Exercices sous Lodex

4. Exercices sous Lodex

A partir du corpus

- Extraire 8 termes à partir du résumé
- A partir du DOI vérifier le statut des références citées
- A partir du DOI repérer les références les plus citées dans le corpus

Démarche pour chaque enrichissement

- Donnez un nom à la colonne résultat
- Sélectionnez le web service dans le catalogue ou cherchez l'url sur ISTEEX TDM
- Sélectionnez la colonne sur laquelle le traitement va être lancé
- Affichez le résultat dans le détail de la notice
- Créez la facette correspondante
- Créez un graphe

Découverte : web services LODEX

5. Documentation

Adresses & Co



Se connecter :

- ↴ ISTEX TDM : <https://services.istex.fr/>
- ↴ SWAGGER : <https://openapi.services.istex.fr/>
- ↴ IA Factory : <https://tdm-factory.services.istex.fr/>

S'authentifier :

- ↴ Vérifier ses droits d'accès : <https://api.istex.fr/auth>
- ↴ Vérifier son accès par fédération d'identité :
<https://api.istex.fr/auth?auth=fede>

Documentation & Tutoriels



Se documenter :

- ↳ Documentation Lodex : <https://www.lodex.fr/docs/documentation/>



Se former :

- ↳ Tutos Lodex : <https://callisto-formation.fr/course/view.php?id=194>

Informations & Contact



Se tenir informé :

- ↴ Article d'actualité : <https://www.istex.fr/category/actualites/>



Chercher de l'aide / Contribuer à l'amélioration :

- ↴ Contact :
 - Via le formulaire : <https://www.istex.fr/contact/>
 - Via la liste : contact@listes.istex.fr
- ↴ Liste de discussion Lodex : <https://groupes.renater.fr/sympa/info/lodex>

Découverte : web services et LODEX

Merci pour votre attention !