

INFORMATION NUMÉRIQUE

Enjeux et Pratiques

Licence 3 / UE502 – Accès expert à l'information – 2024-2025

Découverte : les web services et LODEX

valerie.bonvallot@inist.fr
lucile.bourguignon@inist.fr
philippe.houdry@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.

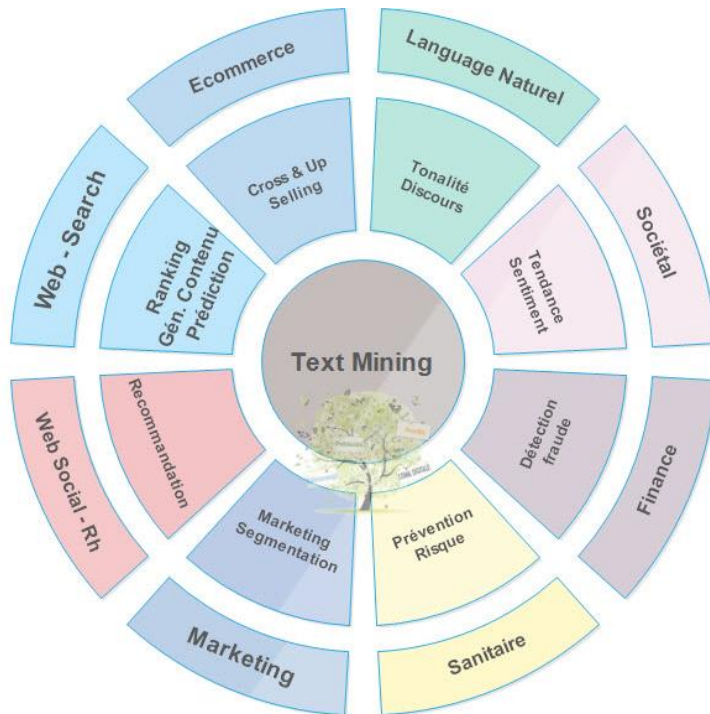
Données → Connaissances

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

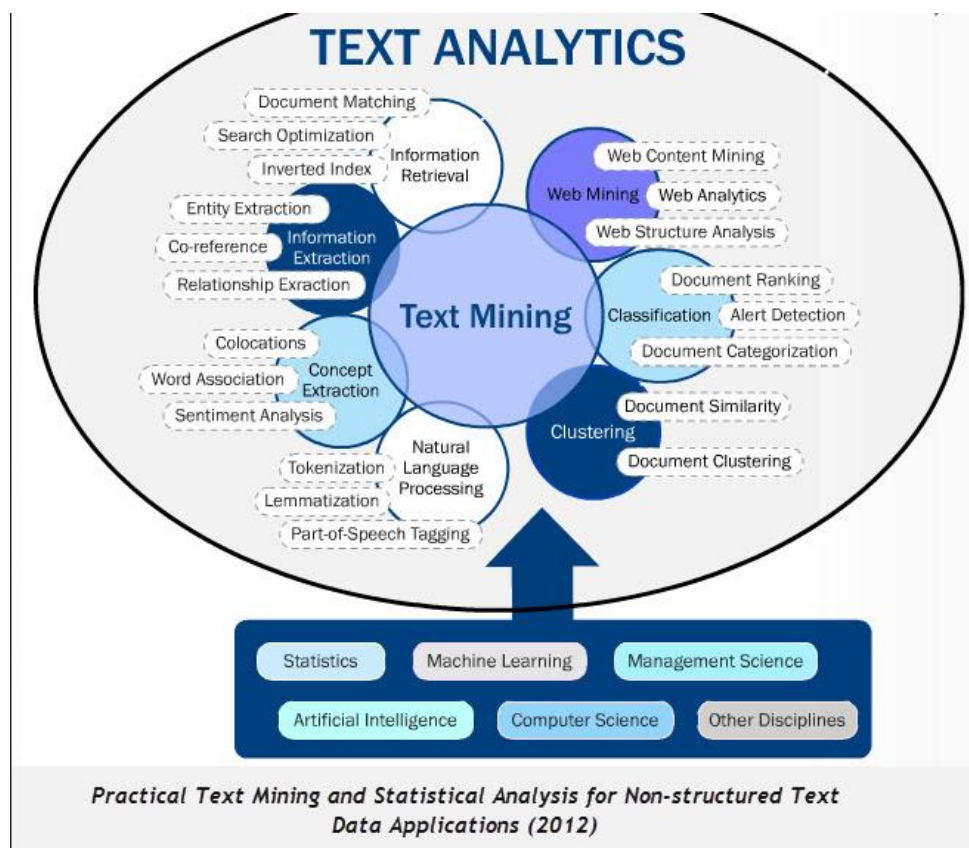


- Faire de la bibliométrie
- Extraire de l'information pertinente
- Répondre à une question (recherche d'information)
- Analyser de gros volumes de textes
- Identification de thèmes
- Détecter des sentiments dans les textes
- Construire des résumés automatiques
- Désambiguïser des lieux, des personnes...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

<https://www.mauricelarger.com/analyse-de-texte-et-seo/2015>

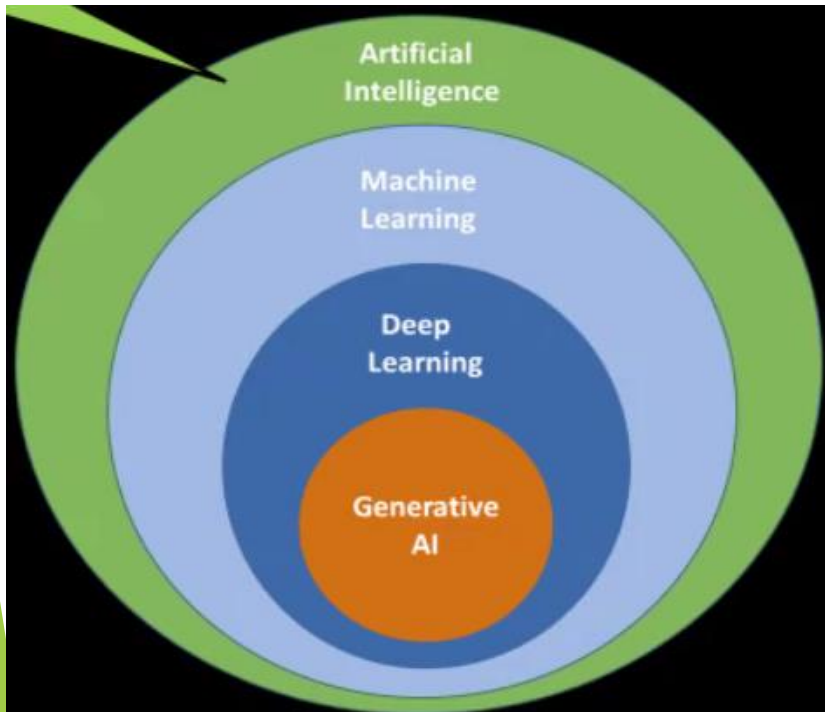
1. TDM et web services

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



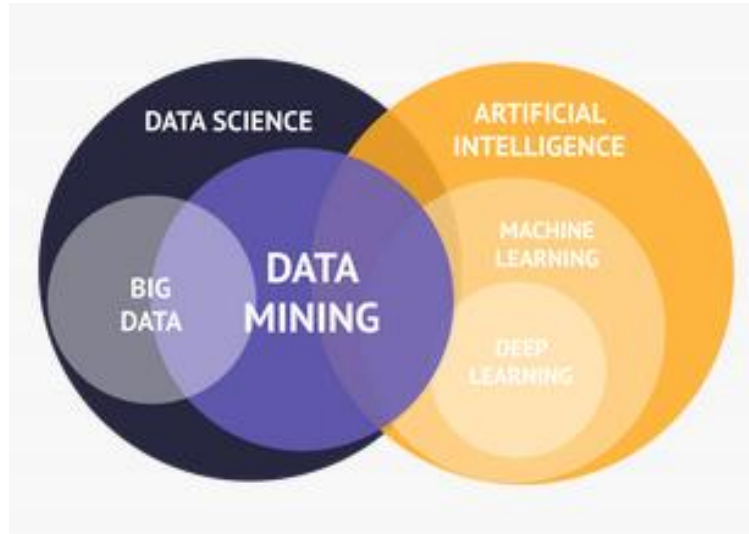
- IA : un programme qui va permettre de réaliser une tâche humaine
 - Apprentissage automatique (Large Language Models : stat, prédiction)
 - Apprentissage profond* et réseaux de neurones (transformeurs)
 - IA générative

Mohand Boughanem

¹L'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

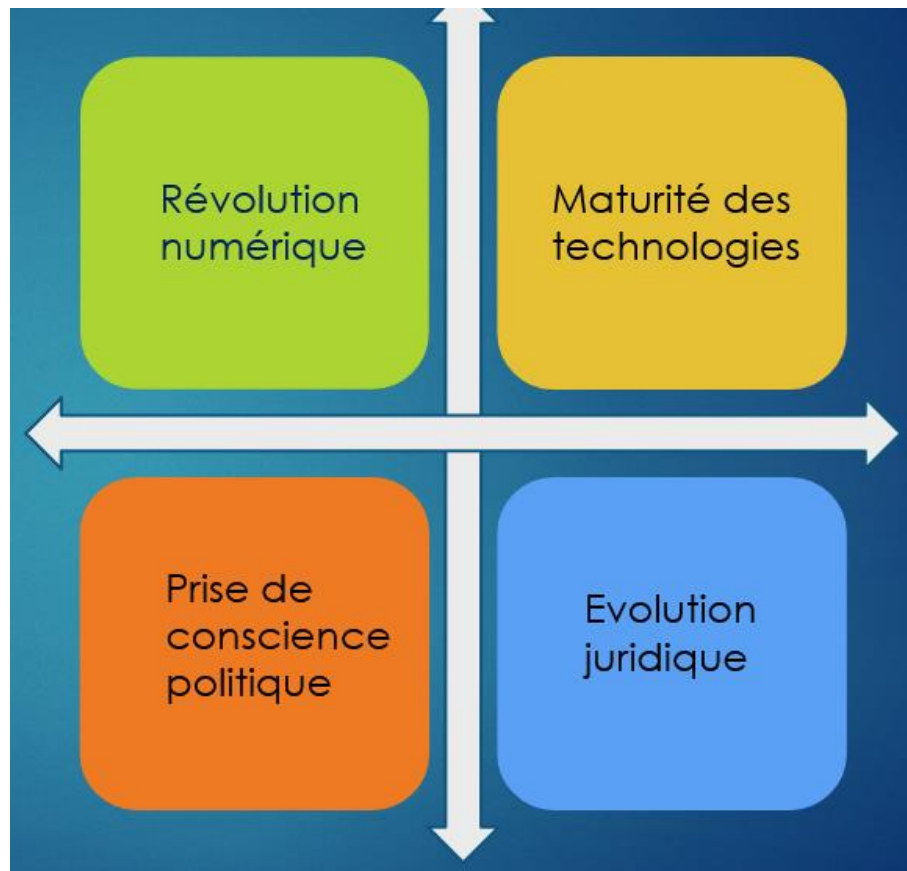
TDM : Text and data mining et Intelligence artificielle
fouille de textes et de données



<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
Algorithmes

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 ISTEX 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**

Éléments catalographiques dataHomogenise HOMOGENÉISATION AUTOMATIQUE DE MOTS-CLÉS →	Résumés aiAbstract-check DÉTECTION DE RÉSUMÉ SCIENTIFIQUE GÉNÉRÉ PAR IA →
Texte intégral textSummarize RÉSUMÉ AUTOMATIQUE D'UN ARTICLE SCIENTIFIQUE →	Citations topRefExtract EXTRACTION DES RÉFÉRENCES PHARES D'UN CORPUS →
Résumés - Texte intégral entityTag EXTRACTION D'ENTITÉS NOMMÉES (PERSONNES, LOCALISATIONS, ORGANISMES ET AUTRES) →	Adresses et affiliations idRorDetect ASSOCIATION D'UN IDENTIFIANT ROR À UNE ADRESSE D'AFFILIATION →

Trouvez un service web correspondant à vos besoins

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

Actuellement **42** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

VOIR TOUS LES SERVICES

2. Navigation dans ISTEX 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

Éléments catalographiques

dataHomogenise
HOMOGÉNÉISATION AUTOMATIQUE
DE MOTS-CLÉS



Texte intégral

textSummarize
RÉSUMÉ AUTOMATIQUE D'UN
ARTICLE SCIENTIFIQUE



Résumés - Texte intégral

entityTag
EXTRACTION D'ENTITÉS NOMMÉES
(PERSONNES, LOCALISATIONS,
ORGANISMES ET AUTRES)



Résumés

aiAbstract-check
 DÉTECTION DE RÉSUMÉ
SCIENTIFIQUE GÉNÉRÉ PAR IA



Citations

topRefExtract
EXTRACTION DES RÉFÉRENCES
PHARES D'UN CORPUS



Adresses et affiliations

idRorDetect
ASSOCIATION D'UN IDENTIFIANT ROR
À UNE ADRESSE D'AFFILIATION



VOIR TOUS LES SERVICES

Trouvez un service
web correspondant à
vos besoins

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

Actuellement **42** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

Tapez ici votre recherche, p.ex. : Classification

RECHERCHER

Description Utilisation Cas d'usage

Niveau d'utilisation : Avancé
Niveau de validation : Expérimental

Objectif

Ce web service traite non plus du texte mais de corpus de textes en anglais. En effet, le résultat obtenu pour chacun des documents dépend des autres.

L'algorithme extrait plusieurs groupes (clusters) d'un corpus afin d'y classer les différents textes en fonction de leur similarité. Un document est présent dans un seul groupe.

Chaque cluster est caractérisé par 20 termes.

Méthode

Dans un premier temps, un embedding est utilisé pour vectoriser les documents. Une fois représentés sous forme de vecteurs, il est possible de calculer leur ressemblance. Pour ce faire, nous réduisons la dimension des vecteurs en utilisant l'algorithme [UMAP](#) puis nous comparons les proximités entre ces vecteurs en utilisant la [distance cosinus](#). Enfin, on les regroupe en cluster en utilisant l'algorithme des [k-means](#).

- Le nombre de clusters est déterminé de manière automatique (en utilisant la méthode de la silhouette). Si des documents ne permettent pas d'être traités, ils seront considérés comme du bruit (dans ce cas précis, le label de leur cluster sera 0 (zéro) ; les documents appartenant au cluster 0 ne sont pas regroupés ensemble).
- L'entrée doit être un texte court (type titre ou un abstract). Fonctionne également sur un tableau de mots-clés pertinents extraits d'un texte (pouvant être obtenus avec [text](#) par exemple).

obtenu pour chacun des documents dépend des autres. L'algorithme repère la liste des identifiants des documents considérés comme du bruit dans...

textClustering
Extraction de clusters d'un corpus

Ce web service traite non plus du texte mais de corpus de textes en anglais. En effet, le résultat obtenu pour chacun des documents dépend des autres. L'algorithme extrait plusieurs groupes (clusters) d'un corpus afin d'y classer les différents textes...

OBJET TRAITÉ

- ☐ Adresses et affiliations (10)
- ☐ Auteurs (2)
- ☐ Éléments catalographiques (3)
- ☐ Citations (1)
- ☐ Résumés (20)
- ☐ Texte intégral (21)

LANGUES (3)

- ☐ Anglais (32)
- ☐ Français (25)
- ☐ Autre (14)

TRAITEMENT (7)

- ☐ Classification (8)
- ☐ Extraction d'entités nommées (9)
- ☐ Homogénéisation (8)
- ☐ Indexation (6)
- ☐ Traitement automatique du langage (3)
- ☐ Prétraitement (3)
- ☐ Validation (2)

TYPE DE DONNÉES (2)

- ☐ Corpus (5)
- ☐ Document (30)

topRefExtract
Extraction des références
phares d'un corpus

Ce web service identifie les N publications les plus citées dans un corpus donné, par défaut 10.

idRorDetect
Associer un identifiant ROR à
une adresse d'affiliation

Ce web service prend en entrée une affiliation pour interroger l'API ROR (Research Organization Registry) et renvoie les informations suivantes :
id_ror : Lien vers la fiche ROR de l'organisme
score : score de similarité name : nom de l'organisme...

sciencematrixClass
Classification en domaines
scientifiques Science-Matrix

Ce web service classe des documents en anglais dans les trois niveaux de la classification Science-Matrix.

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."

2. Navigation dans ISTEX TDM

Détection de code RNSR

rnsrRuleDetect Attribution d'identifiant(s) RNSR à une adresse (Alignements)

Le web service attribue, à l'aide de règles, un ou plusieurs identifiants RNSR à partir d'une adresse d'affiliation d'auteur et d'une année de publication. Quand aucun code RNSR n'est trouvé, le service renvoie un tableau vide.

rnsrLearnDetect Attribution d'identifiant(s) RNSR à une adresse (Apprentissage)

Ce web service attribue un ou plusieurs identifiant(s) RNSR à partir d'une adresse d'affiliation d'auteur en langue française.

Avant

"CNRS UMR AMAP MONTPELLIER FRA"

Après

<> RNSR://200317641S

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

The screenshot displays the LODEX web interface. On the left is a dark sidebar with a menu containing 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main area has a green header with 'LODEX', 'DONNÉES', 'AFFICHAGE', and a 'DÉPUBLIER' button. Below the header, the 'Enrichissements' section is active. It includes a 'Nom' field with a blue '1' next to it, a 'Mode avancé' toggle switch, an 'URL du web service' field with a blue '2' and a green icon button, and two dropdown menus labeled 'Colonne de la source' (with a blue '3' below it) and 'Sous-chemin'. At the bottom are 'ANNULER' and 'SAUVEGARDER' buttons, with a blue '4' next to the latter. On the right, a grey box titled 'Aperçu de la valeur*' shows 'Aucune donnée' and a footnote: '* valeur de la colonne source ou de la règle avancée'.

ISTEX TDM

Enrichissements

Filtre Istex TDM : Type de données "Documents"

TOUS

AFFILIATION

CLASSIFICATION

ENTITÉS NOMMÉES

ENTITÉS NOMMÉES

GÉOGRAPHIE

HOMOGÉNÉISATION

HOMOGÉNÉISATION

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 Loterre 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 le 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 le nom et prénom, des titres de publications, ou

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

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..."

Web services et LODEX – 2024-2025 – Valérie Bonvallot – CNRS-INIST

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"

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

Données

Enrichissements

Précalculs

Ressources cachées

Nom *

noiseDetect

Statut :

Non démarré

URL du web service *

https://text-clustering.services.istex.fr/v1/noise-lodex

Colonne(s) source(s) *

Résumé Colonne(s) source(s) *

SUPPRIMER

ISTEX TDM

Les services istex pour la fouille de textes.

Description

Utilisation

Cas d'usage

URL DU WEB SERVICE À RENSEIGNER DANS LODEX PRÉCALCUL EST :

https://text-clustering.services.istex.fr/v1/noise-lodex

IdaClass - Extraction de 15 thématiques d'un corpus

Extrait 15 thématiques d'un corpus : une thématique (ou topic) est caractérisée par dix mots. Une fois les thématiques extraites, chaque document se voit attribuer une ou plusieurs thématique(s). Le texte doit être en anglais.

IdaClass - Extraction de 8 thématiques d'un corpus

Extrait 8 thématiques d'un corpus : une thématique (ou topic) est caractérisée par dix mots. Une fois les thématiques extraites, chaque document se voit attribuer une ou plusieurs thématique(s). Le texte doit être en anglais.

IdaSegment - Extraction de 15 thématiques à partir d'un jeu de données avec un format adapté à la création de graphiques

Pour un graphe, crée à partir de l'ensemble des documents un champ «Ida» constitué de 15 topics. Chaque topic contient un champ «word», composé d'une liste de 10 mots les plus caractéristiques du topic, ainsi que d'un champ «weight» qui correspond au poids associé au sujet dans le document. Le texte doit être en anglais.

IdaSegment - Extraction de 8 thématiques à partir d'un jeu de données avec un format adapté à la création de graphiques

Pour un graphe, crée à partir de l'ensemble des documents un champ «Ida» constitué de 8 topics. Chaque topic contient un champ «word», composé d'une liste de 10 mots les plus caractéristiques du topic, ainsi que d'un champ «weight» qui correspond au poids associé au sujet dans le document. Le texte doit être en anglais.

noiseDetect - Détection de bruit d'un corpus

ANNULER

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"

LODEX

DONNÉES AFFICHAGE

DÉPUBLIER

Données

Enrichissements

Précalculs

Ressources cachées

Nom * 1

noiseDetect

LANCER 5

Statut : Non démarré

VOIR LES LOGS

URL du web service *

https://text-clustering.services.istex.fr/v1/noise-lodex 2

Colonne(s) source(s) *

Résumé Colonne(s) source(s) *

SUPPRIMER

RETOUR SAUVEGARDER 4

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...

Découverte : web services LODEX

4. Exercices sous Lodex

4. Exercices sous Lodex

A partir du corpus blob

- Indexation – 8 termes
- Classification supervisée avec apprentissage (Science Metrix)

Démarche pour chaque enrichissement

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

4. Exercices sous Lodex

Questions

- Combien de documents sont indexés avec le terme Teeft physarum seule ou dans une expression ?
- Combien de documents sont indexés avec le terme Teeft plasmodium ?
- Combien de documents ont été classifiés ?
- Combien de documents appartiennent à la classe nanosciences & nanotechnologie ?
- Combien de web services peuvent être utilisés sur du texte écrit en français ?
- Qu'obtient-on lorsque l'on essaie de détecter la langue de cette expression : "Time flies" ? (en testant le service "détection de langues" via openAPI)
- En utilisant le bon web service, donner un mot-clé pertinent de : "Le journal CNRS est un site d'information scientifique."

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://ia-factory.services.istex.fr/>
- TMTools Explorer : <https://data.istex.fr/instance/tm-tools-explorer>

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 !

LODEX

