Text & Data Mining
Fouille de textes et Istex

Valérie Bonvallot
valerie.bonvallot@inist.fr

Construire et analyser un corpus avec l'infrastructure Istex
Urfist Paris - 6 mai 2025

Adaptation du support de Fabienne Kettani

Déroulement de l’intervention

1. TDM : Théorie

2. TDM à l’Inist

Théorie

- Définitions
- Contexte
- Techniques

Théorie : Définitions - Contexte - Techniques de fouille de textes

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 (non structurées)
⇒ Connaissances

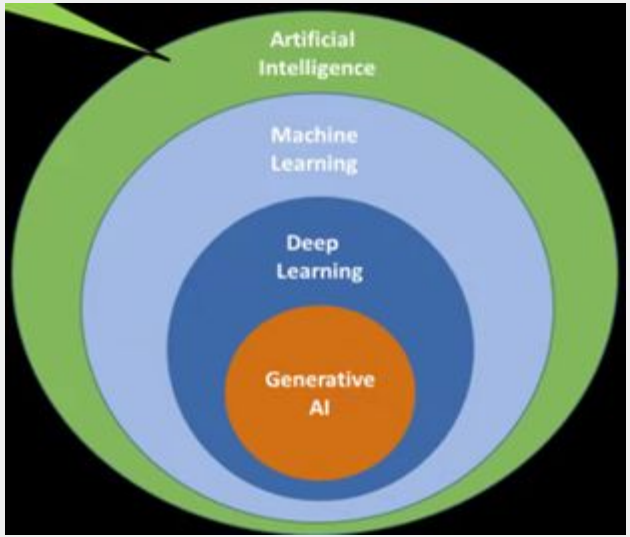
C'est une spécialisation de la fouille de données (data mining) qui fait appel aux **méthodes** de l'**Intelligence Artificielle**, du **Traitement Automatique des Langues** et des **Statistiques**

La directive sur le droit d'auteur définit le text and data mining ou la fouille de textes et de données, comme « 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** » **(nov. 2021)**

Théorie : Définitions - Contexte - Techniques de fouille de textes

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications (2012)

Théorie : Définitions - Contexte - Techniques de fouille de textes



IA : un programme qui va permettre de réaliser une tâche humaine

Apprentissage automatique

(Large Language Models : stat, prédiction)

Apprentissage profond* et réseaux de neurones

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

Théorie : Définitions - Contexte - Techniques de fouille de textes

IA : un programme qui va permettre de réaliser une tâche humaine

Apprentissage automatique

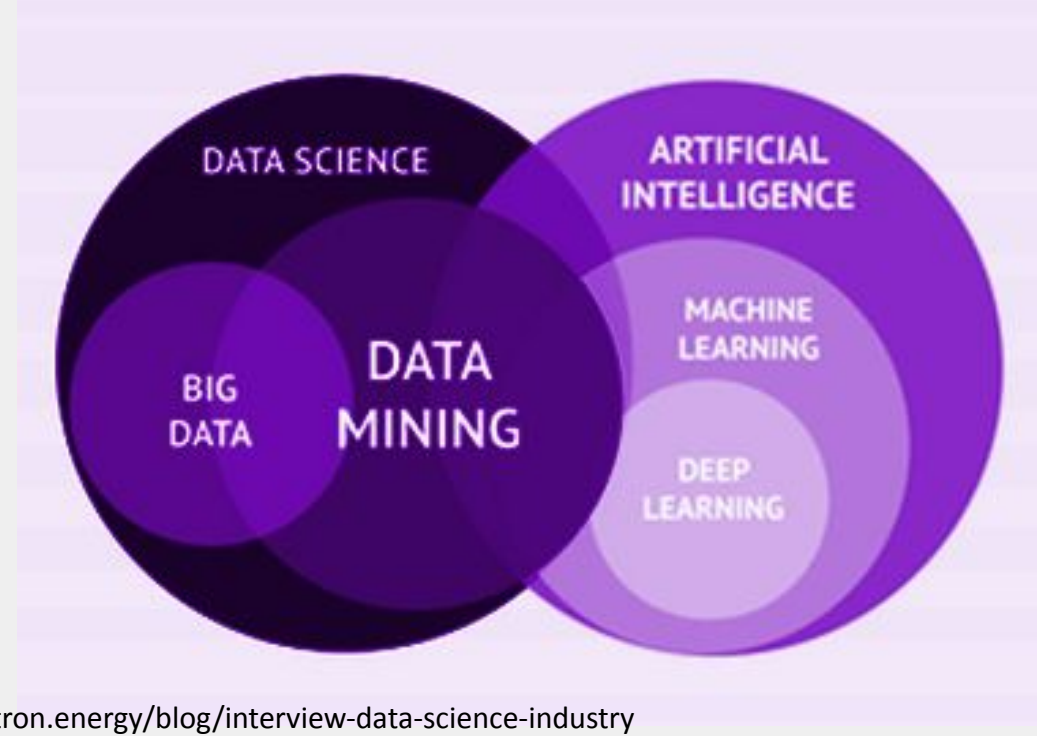
(Large Language Models : stat, prédiction)

Apprentissage profond et réseaux de neurones*

IA générative (agent conversationnel : chatbot)

*Un réseau de neurones, c'est comme un système avec plusieurs couches, où chaque couche apprend un petit morceau du problème, et ensemble, elles trouvent la bonne réponse

Théorie : Définitions - Contexte - Techniques de fouille de textes



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

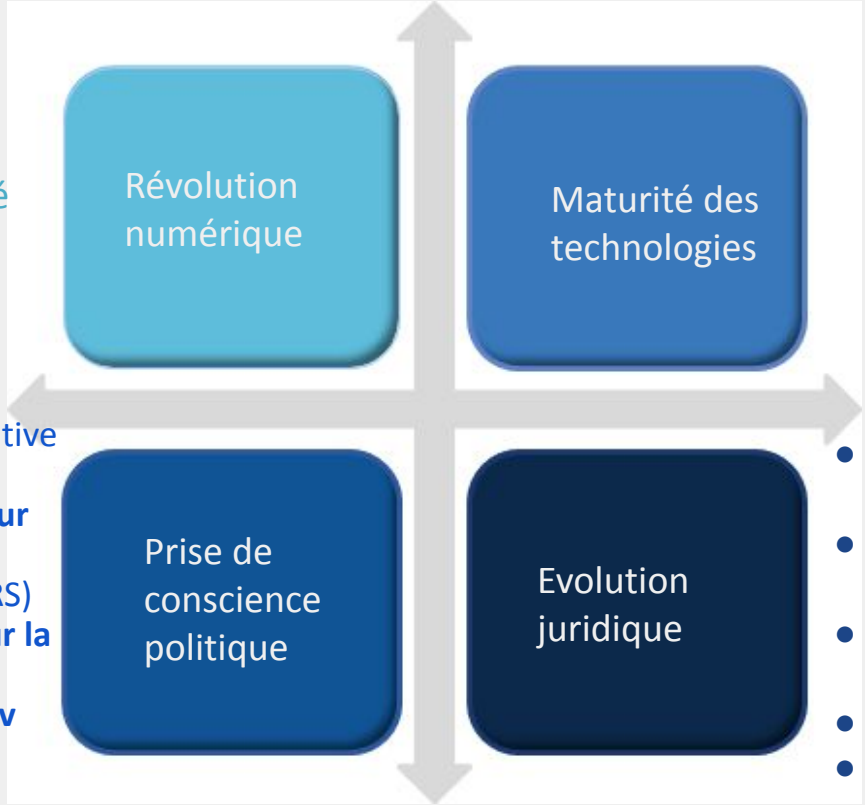
Théorie : Définitions - Contexte - Techniques de fouille de textes

- Extraire de l'information pertinente
- Répondre à une question (recherche d'information)
- Analyser de gros volumes de textes
- Détecter des sentiments dans les textes
- Générer des résumés automatiques
- Traduire automatiquement

- Assister le diagnostic médical
- Désambiguïser des lieux, des personnes ...
- Faire des systèmes de recommandations
- Détecter des « fake news »
- Identifier des thèmes : trier des mails, des textes
- Faire de la bibliométrie
- ...

Théorie : Définitions - **Contexte** - Techniques de fouille de textes

- **Big data**
- **5V** : volume, vélocité, variété, valeur, véracité
- **180** zettaoctets (10^{21})



- **30** ans d'expérience
- Puissance de **calcul**
- **Algorithme** : des statistiques à l'apprentissage profond

- 2001 Budapest Open Initiative
- 2003 Déclaration de Berlin
- 2018 **1er Plan national pour la science ouverte**
- 2019 Feuille de route (CNRS)
- 2021 **2è Plan national pour la science ouverte**
- 2022 **Recherche Data Gouv**
- 2023 Comité de l'IAG

- 2016 **Loi pour une République numérique**
- 2019 Directive européenne "Copyright"
- 2021 **Transposition en droit français**
- 2022 Décret sur **l'exception** TDM
- 2024 Règlement européen sur IA

Théorie : Définitions - Contexte - Techniques de fouille de textes

Complexité des langues ⇒ besoin d'interpréter et comprendre

Paris	capitale de la France, ville US
ne... pas...	négation
Orange	polysémie : couleur, fruit, société, ville
Cocker	hyperonymie (chien)
Boire un verre	métonymie

S'appuyer sur le traitement de la langue

Multilinguisme

Alphabet : latin, cyrillique, grec, arabe, …

Le **découpage** des mots, des phrases, des paragraphes

La **graphie** des mots, leur genre et leur(s) catégorie(s) syntaxique(s)

La **syntaxe** ou comment sont construites les phrases

La **sémantique** des mots ou comment les désambiguïser

Théorie : Définitions - Contexte - Techniques de fouille de textes

Stanford CoreNLP : http://corenlp.run/

Des techniques de TAL (traitement automatique des langues)

"Comment transformez vous un document et son contenu en chiffres ?"

Tokenisation

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

POS tagging
(Part Of Speech)

ADV	VERBE	PRON	DET	NOM	CCONJ	DET	NOM	ADV	NOM	PONCT
Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

Lemmatisation
(forme canonique)

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

transformer — chiffres

Stemming
(racinisation)

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

transform — docu — chiffr

Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques pour traiter les données : vectorisation

"Comment transformez vous un document et son contenu en chiffres ?"

Embedding*

réseaux de neurones

Comment → [0.01, 0.8, -0.1 , … , 0.2 , -1.4]

transformez → [-0.8, 0.2, … , -1.4]

Opération (+ ou - simple) sur l'ensemble des vecteurs.

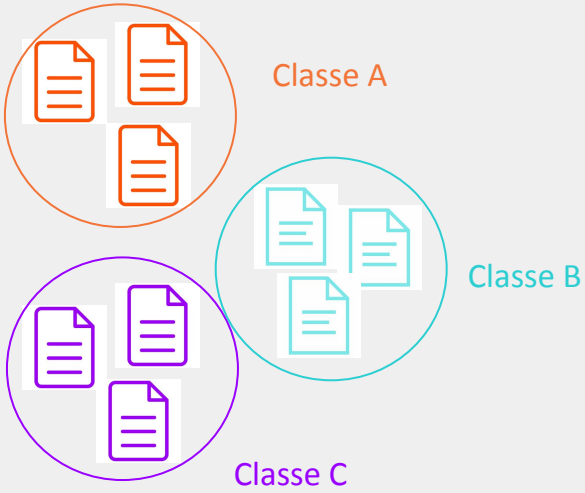
"Comment transformez vous un document et son contenu en chiffres ?" → [-0.79, 1, … , -2.8]

***plongement (plonger un mot dans un espace numérique) - représentation vectorielle - encodage vectoriel**

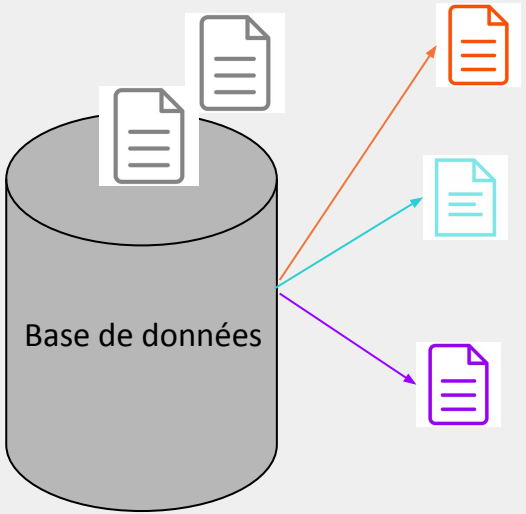
Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : classification supervisée

Classer les documents avec apprentissage sur données étiquetées



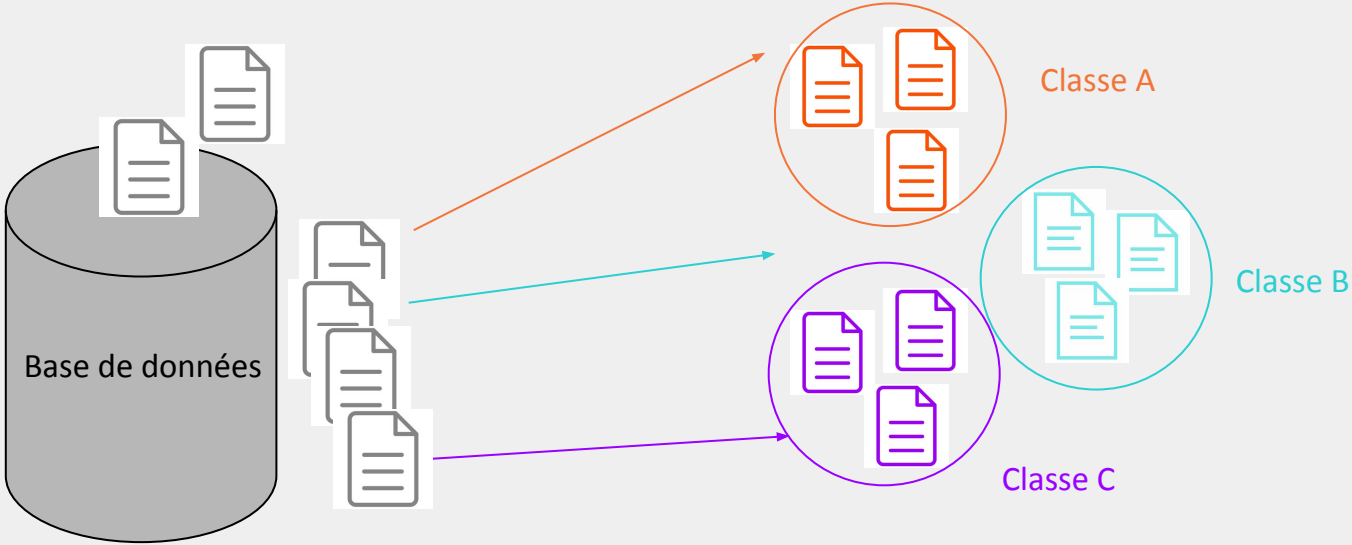
Apprentissage



Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : classification NON supervisée (clustering-regroupement)

Classer les documents sans apprentissage



Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : Reconnaissance d'entités nommées

Repérage de personnes, lieux géographiques, institution …

Méthodes : comparaison avec des nomenclatures

apprentissage automatique à partir d'exemples

Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : Indexation libre et/ou contrôlée

Repérage de termes caractérisant le document et permettant de le retrouver ensuite au sein d'un corpus

TDM à l’Inist

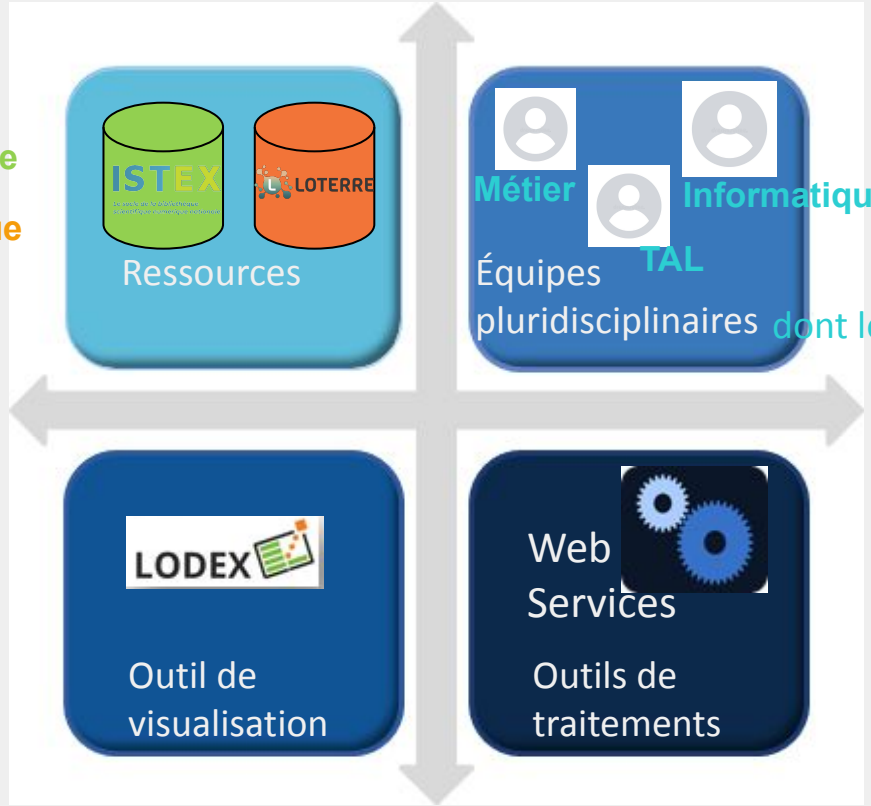
- Environnement propice
- Services et outils (démonstrations)

TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Base bibliographique

Base terminologique



TDM à l'Inist : Environnement - Services et outils
(Démonstrations)

Des web services

Programmes accessibles sur Internet pour que 2 machines communiquent
Un type spécifique API

1 web service = 1 tâche, 1 traitement

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

Paramétrage minimal

Données issues de différentes sources

Programmes en **open source** sous github : https://github.com/Inist-CNRS/web-services

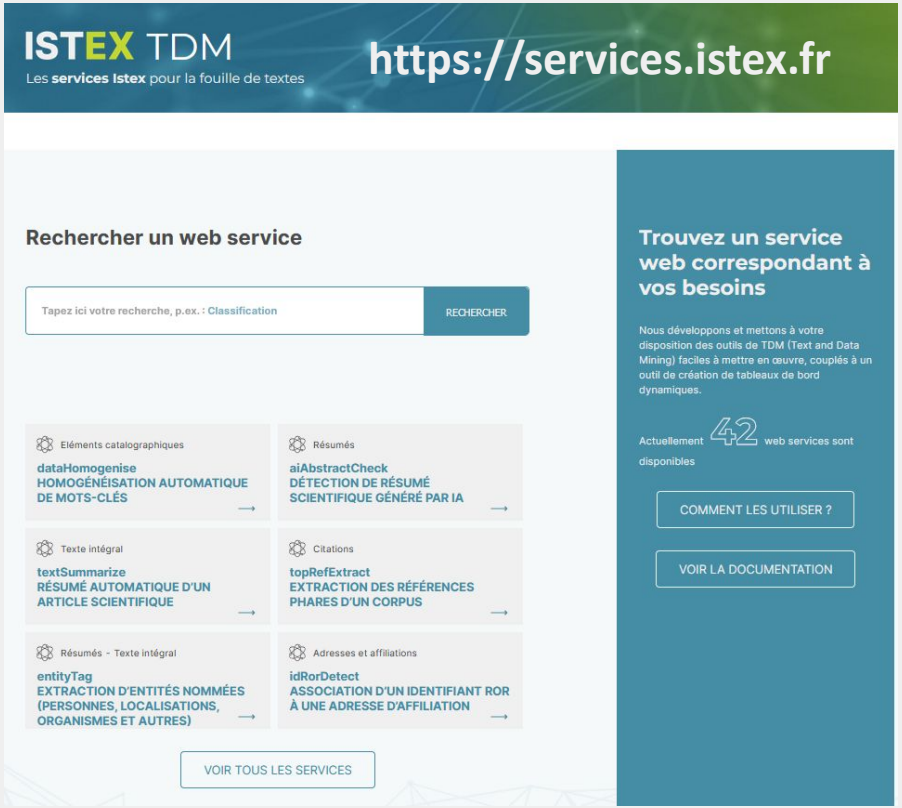
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Un catalogue en ligne

Recensement et description
Modalités d'utilisation
- url du web service
- lien vers swagger
- lien vers github

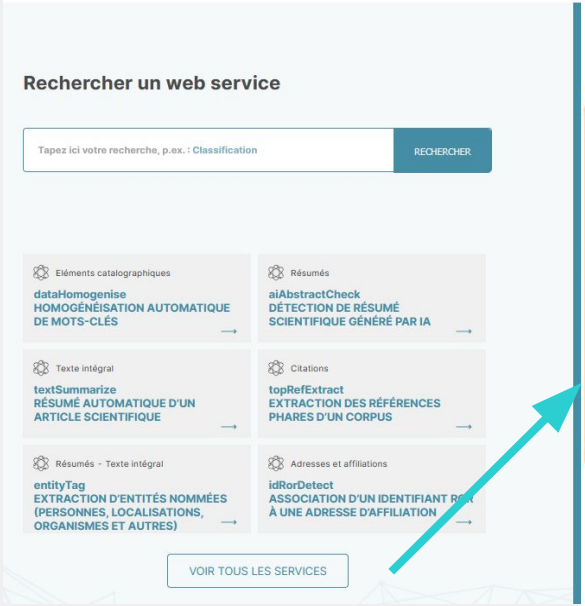
Cas d'usage - illustration

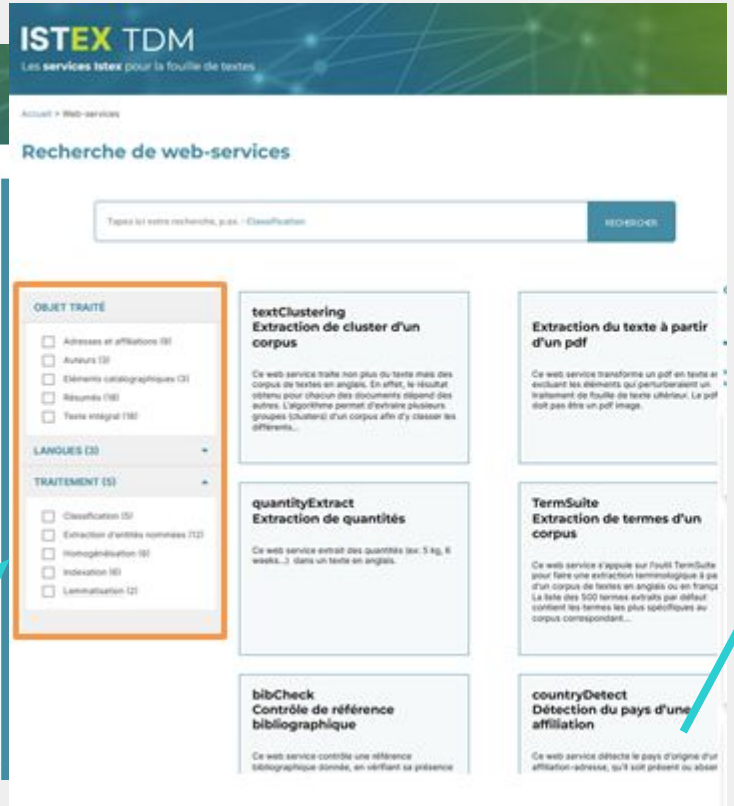


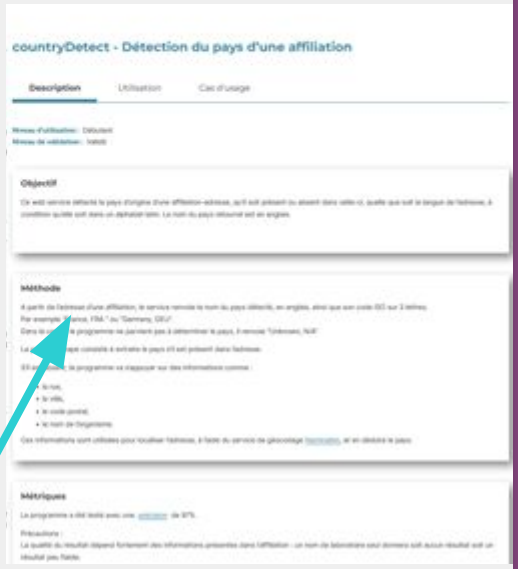
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Un catalogue en ligne







TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Un catalogue en ligne

Onglet "Description"

- Niveau d'utilisation / Niveau de validation
- Objectif
- Méthode, modèles et ressources
- Métrique et précaution d'utilisation
- Variantes
- Références
- Ces web services qui peuvent vous intéresser

Onglet "Utilisation"

- URL pour Lodex
- Exemple du traitement (entrée ⇒ sortie)
- Pour aller plus loin
- Lien vers OpenApi (pour tester)
- Lien vers le code source (github)

Onglet "Cas d'usage"

- Définition des besoins
- Illustrations

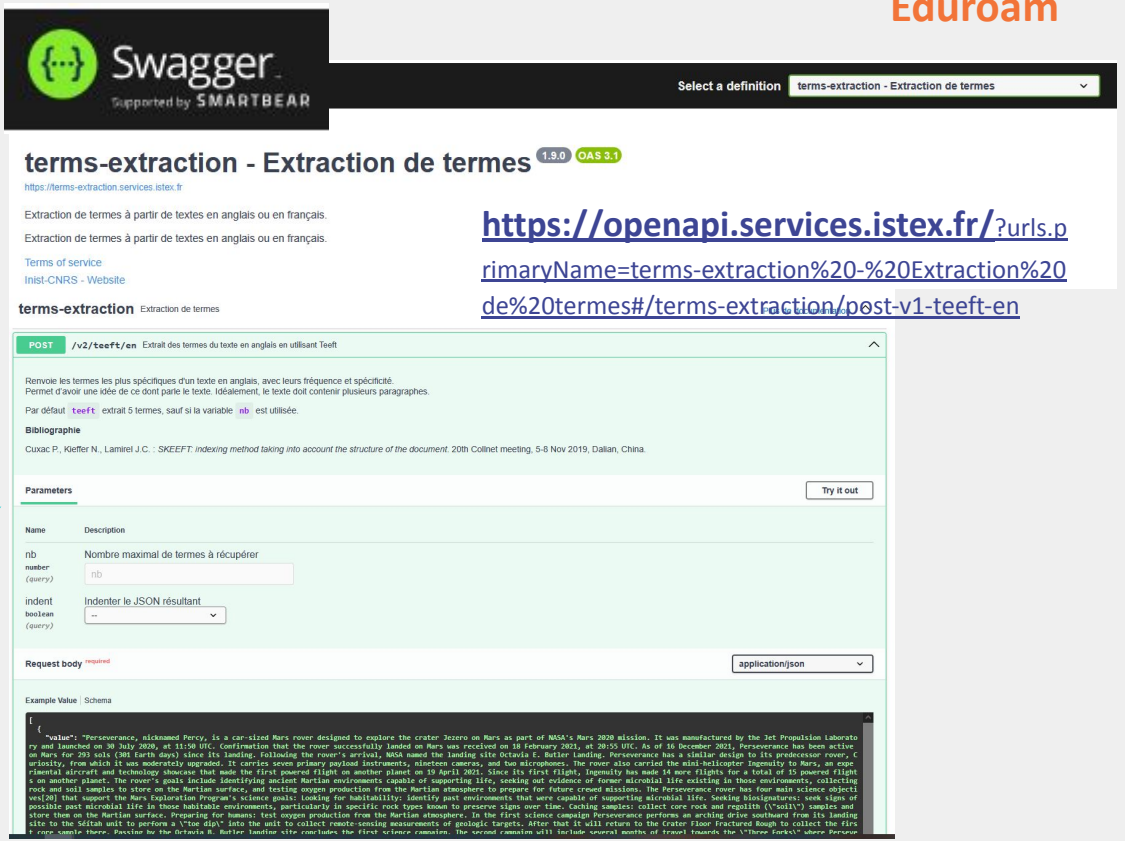
TDM à l'Inist : Environnement - Services et outils

Droits Istex

Eduroam

(Démonstrations)

Des tests en ligne



https://openapi.services.istex.fr/?urls.primaryName=terms-extraction%20-%20Extraction%20de%20termes#/terms-extraction/post-v1-teeft-en

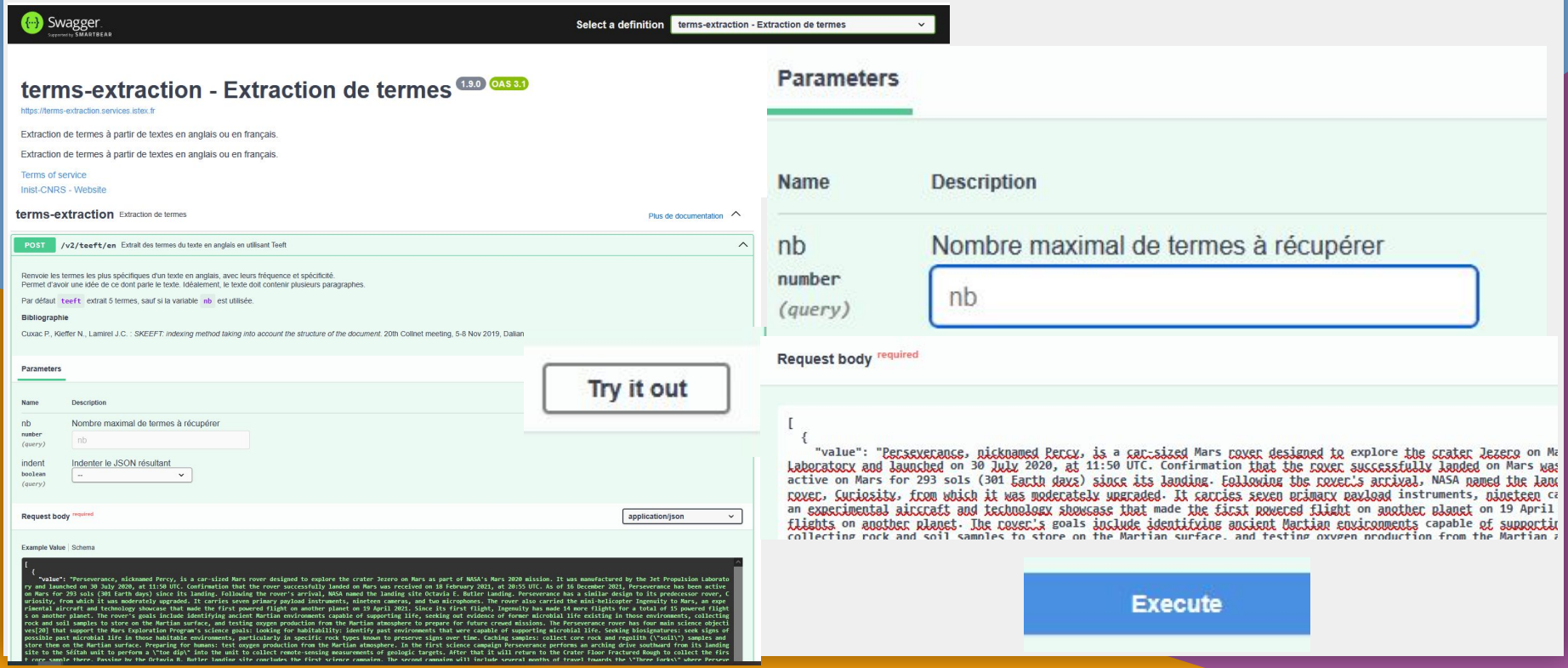
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Des tests en ligne

Droits Istex

Eduroam



TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Des tests en ligne

Request URL

```
https://terms-extraction.services.istex.fr/v2/teeft/en
```

Server response

Code	Details
200	Response body


```
    {
      "term": "perseverance",
      "frequency": 6,
      "specificity": 1
    },
    {
      "term": "martian surface",
      "frequency": 2,
      "specificity": 0.7063
    },
    {
      "term": "martian atmosphere",
      "frequency": 2,
      "specificity": 0.7063
    },
    {
      "term": "crater jezero",
      "frequency": 1,
      "specificity": 0.3532
    },
    {
      "term": "jet propulsion laboratory",
      "frequency": 1,
      "specificity": 0.3532
    }
  ]
 }
]
```

TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Des lignes de commandes pour les plus avancés

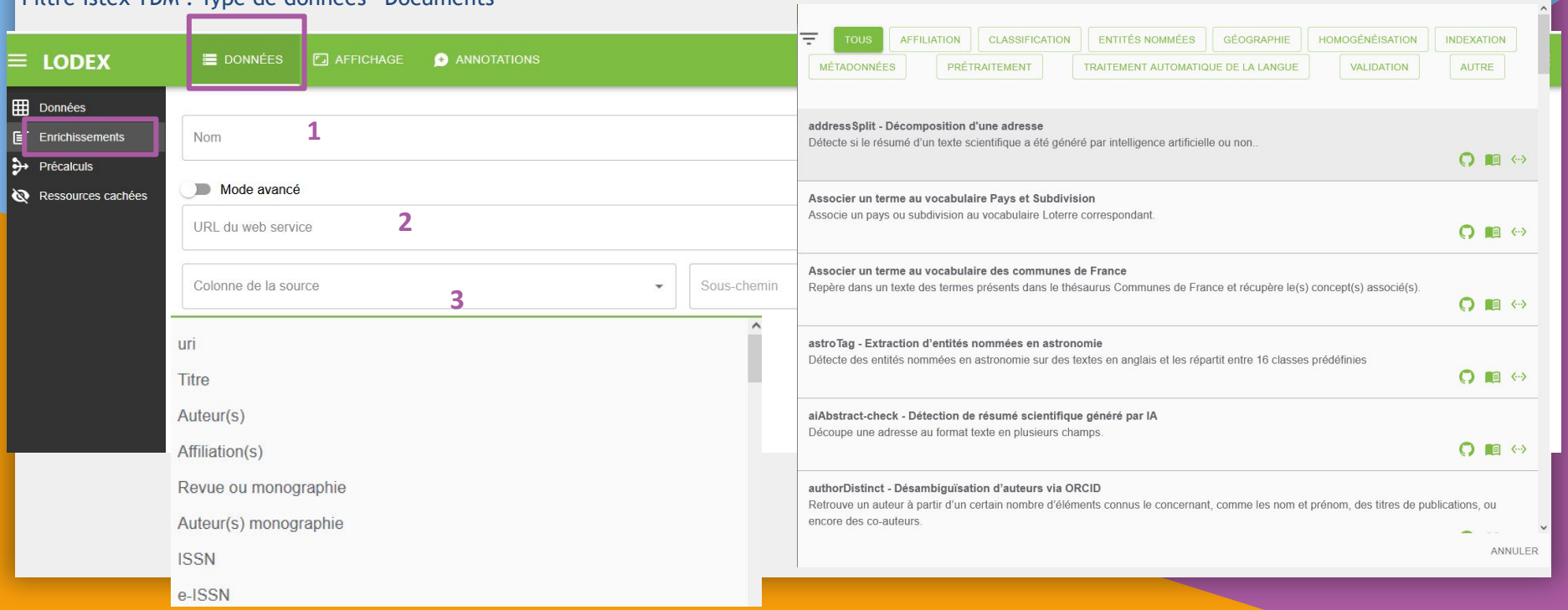
```
[~]$ curl -X 'POST' \
'https://terms-extraction.services.istex.fr/v2/teeft/en' \
-H 'accept: application/json' \
-H 'Content-Type: application/json' \
-d @data.json \

[{"id":"https://en.wikipedia.org/wiki/Perseverance_(rover)","value":[{"term":"persev
erance","frequency":6,"specificity":1},{"term":"martian surface","frequency":2,"spec
ificity":0.7063},{"term":"martian atmosphere","frequency":2,"specificity":0.7063},{"
term":"crater jezero","frequency":1,"specificity":0.3532},{"term":"jet propulsion la
boratory","frequency":1,"specificity":0.3532}]}]%
```

TDM à l'Inist : Environnement - Services et outils (Démonstrations)

Enrichissements exécutables et paramétrables depuis Lodex

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

Filtre Istex TDM : Type de données "Documents"



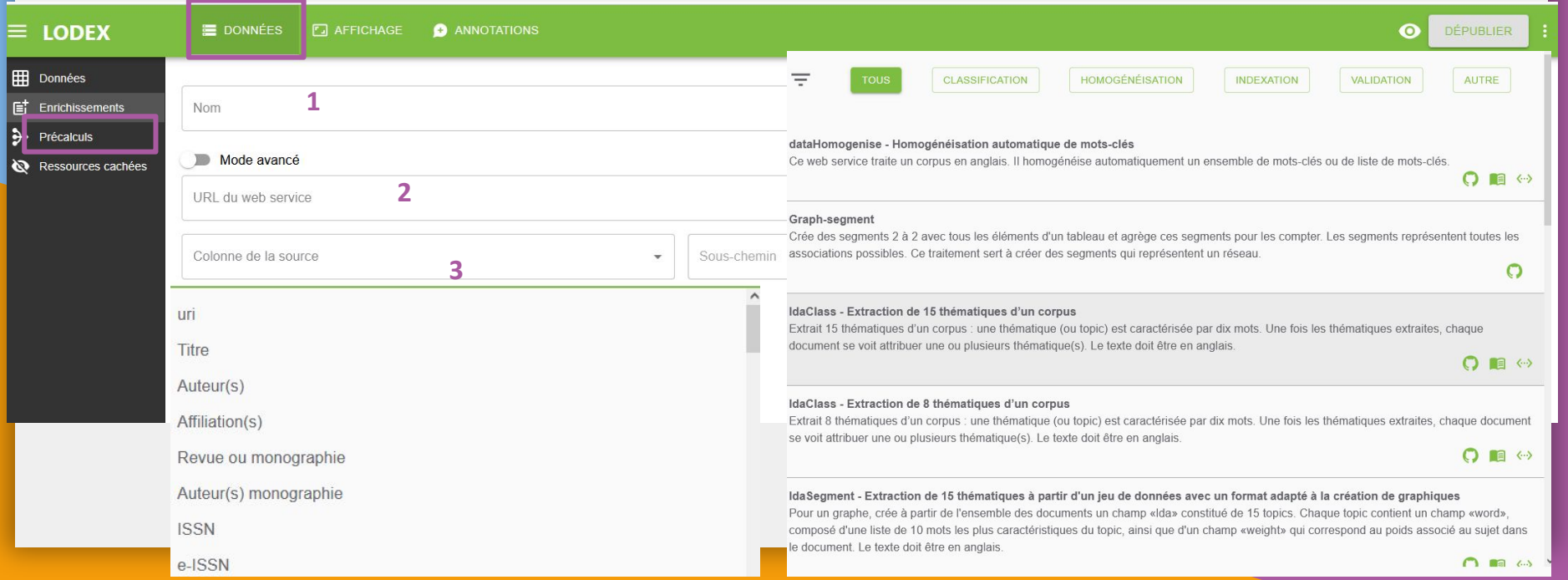
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Précalculs exécutables et paramétrables depuis Lodex

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"



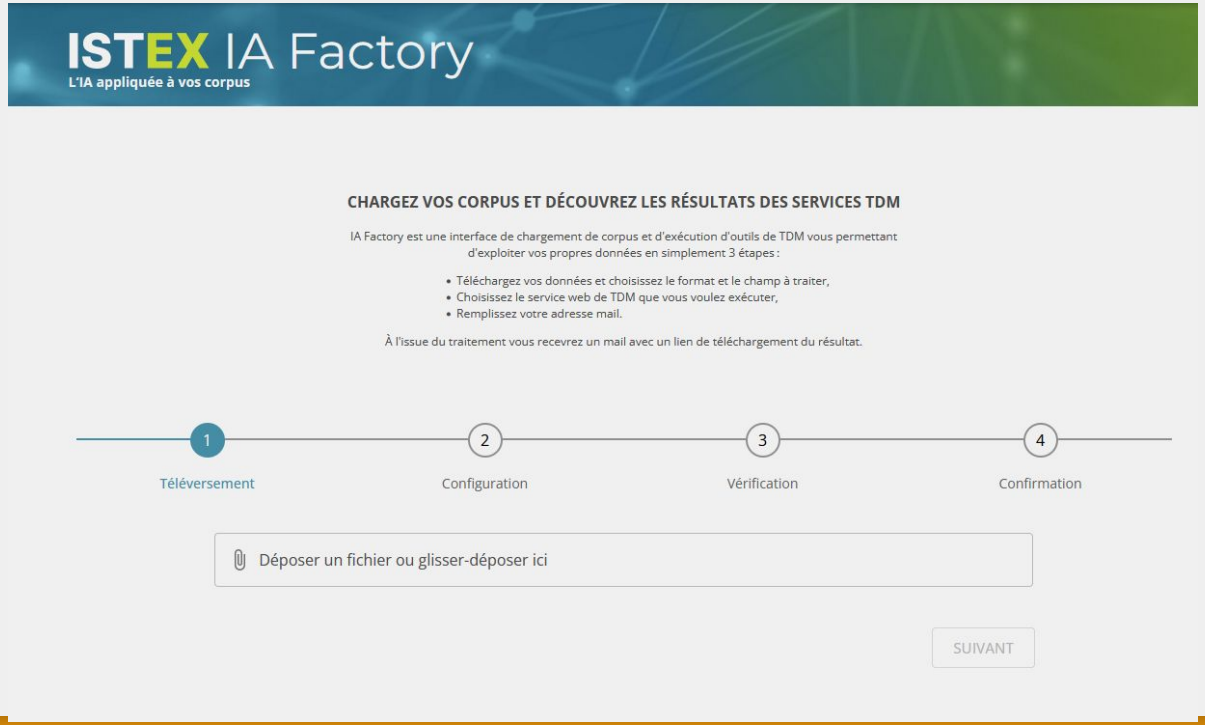
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Droits Istex

Eduroam

Web services exécutables depuis IA Factory



- TermSuite
- Teeft
- corpusSimilarity (en cours)
- lda
- noiseDetect
- textClustering
- scienceMetrixClass
- topRefExtract
- bibCheck

TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Exemples de traitements

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"
```

TDM à l'Inist : Environnement - Services et outils (Démonstrations)

Exemples de traitements

langDetect
Détection de la langue d'un texte

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

TDM à l'Inist : Environnement - Services et outils (Démonstrations)

Exemples de traitements

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."
```

TDM à l'Inist : Environnement - Services et outils (Démonstrations)

Exemples de traitements

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.

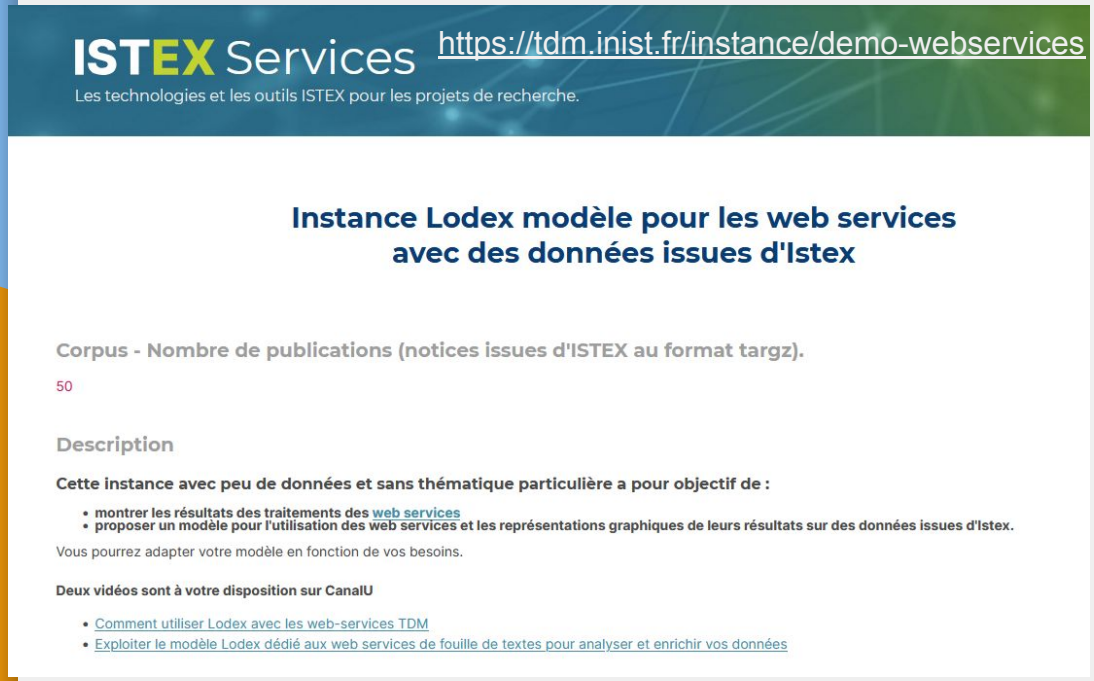
rnsrRuleDetect
Attribution d'identifiant(s) RNSR à une adresse (Alignements)

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.

Avant	Après
"CNRS UMR AMAP MONTPELLIER FRA"	RNSR://200317641S

TDM à l'Inist : Environnement - Services et outils (Démonstrations)

Instance Lodex avec les différents web services



2 vidéos sur CanalU

https://www.canal-u.tv/chaines/inist-cnrs

- Comment utiliser Lodex avec les web-services TDM
- Exploiter le modèle Lodex dédié aux web services de fouille de textes pour analyser et enrichir vos données

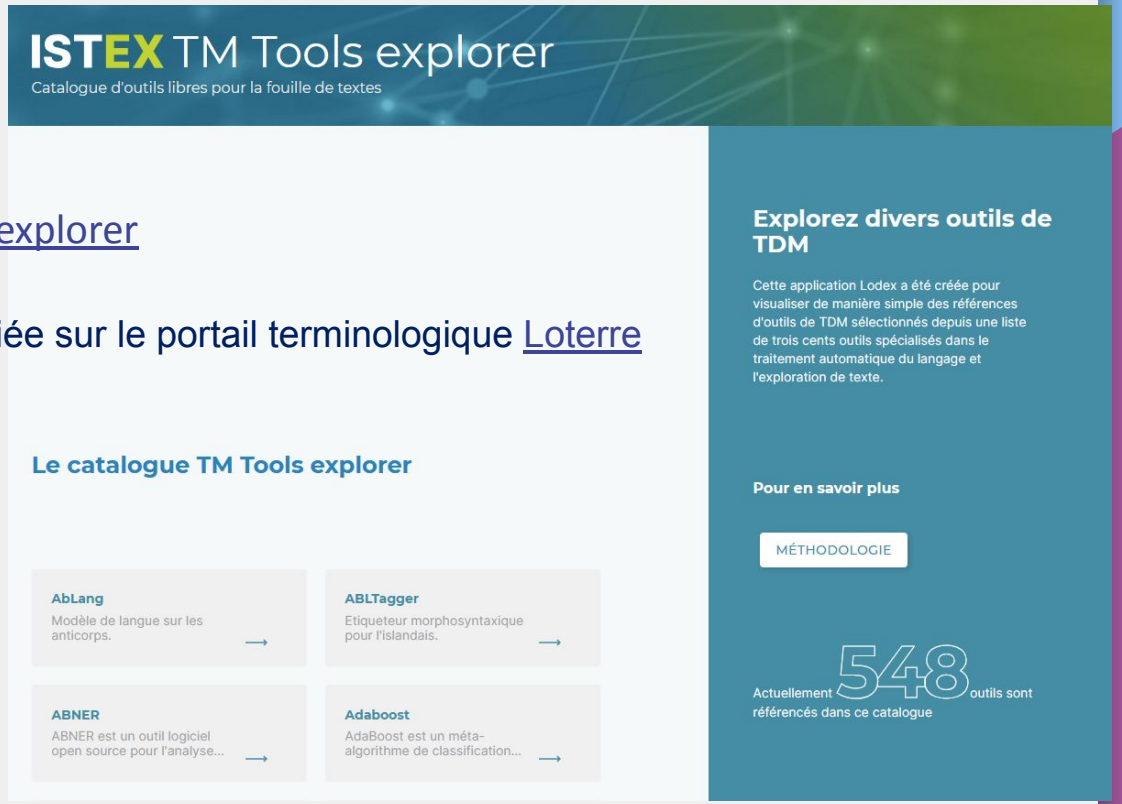
TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Un catalogue d'outils libres

https://data.istex.fr/instance/tm-tools-explorer

Repose sur un **thésaurus : ThesoTM** publiée sur le portail terminologique Loterre



Pour terminer

Conclusions

TDM

Données (non structurées) ⇒ Connaissances

IA + Statistiques + TAL

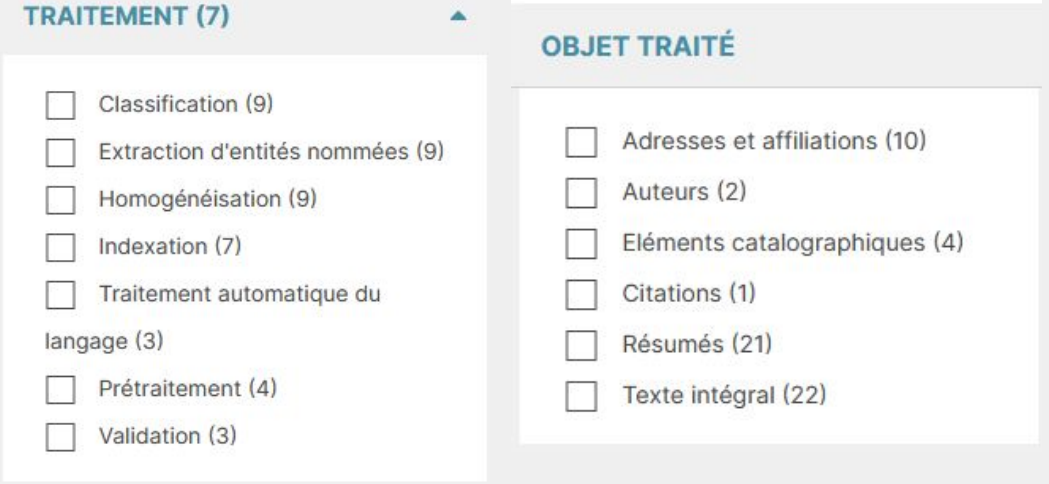
IA>Machine learning (LLM)>Deep learning (réseaux de neurones)

Outils et services

Web services pour enrichir les données

Catalogues

Lodex



Merci pour votre attention