

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

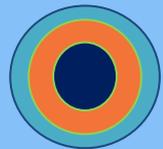
1 TDM : Théorie

2 TDM à l'Inist

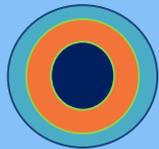




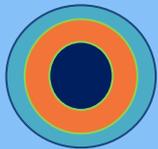
Théorie



Définitions



Contexte



Techniques



Théorie : Définitions

- Contexte - Techniques de fouille de textes

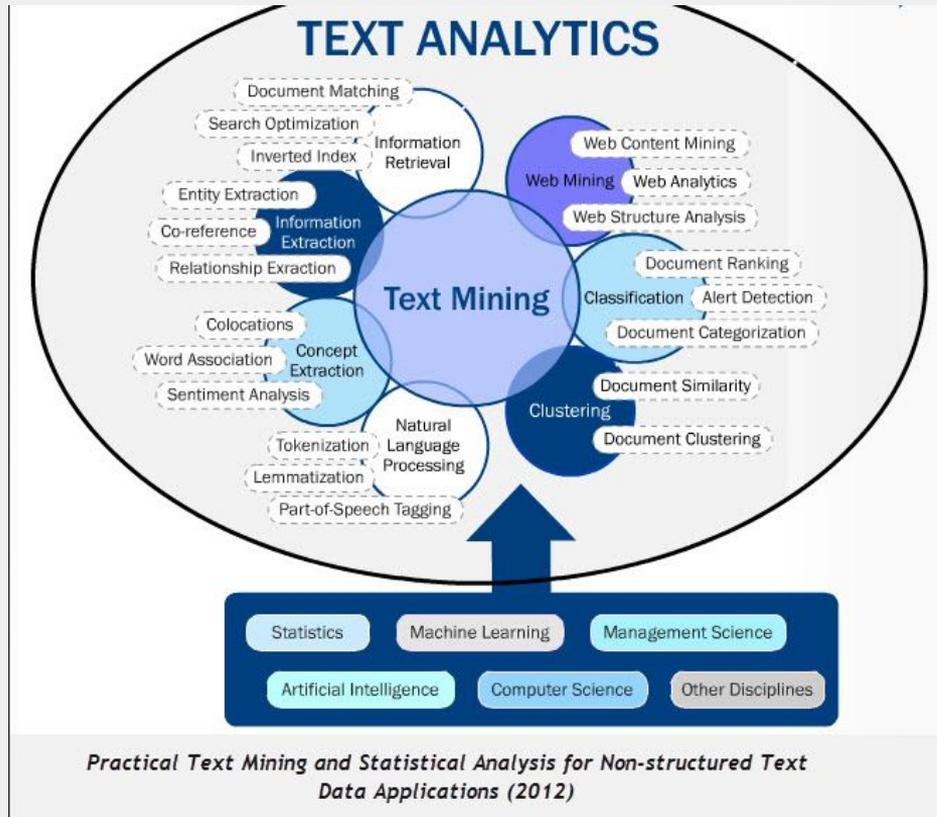
Ensemble des méthodes et des traitements informatiques qui consistent à **analyser le sens des textes** en langage naturel pour en donner une **représentation utilisable** par les humains et les ordinateurs.

Données (non structurées)
⇒ **Connaissances**

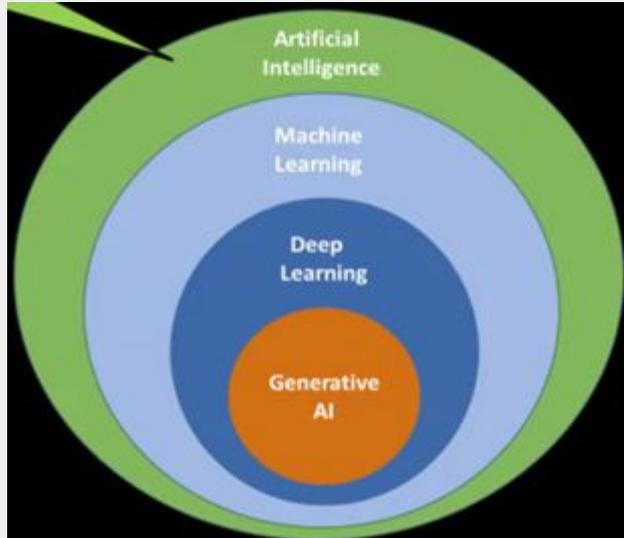
C'est une spécialisation de la fouille de données (data mining) qui fait appel aux **méthodes** de l'**Intelligence Artificielle**, du **Traitement Automatique des Langues** et des **Statistiques**

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



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)

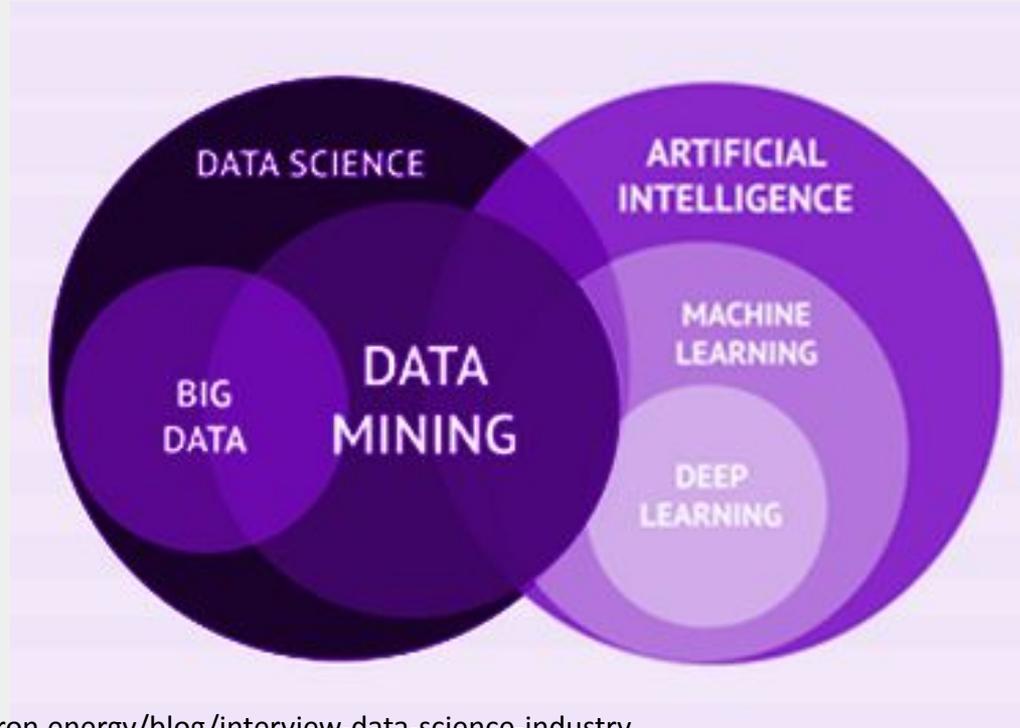
Apprentissage profond et réseaux de neurones*

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

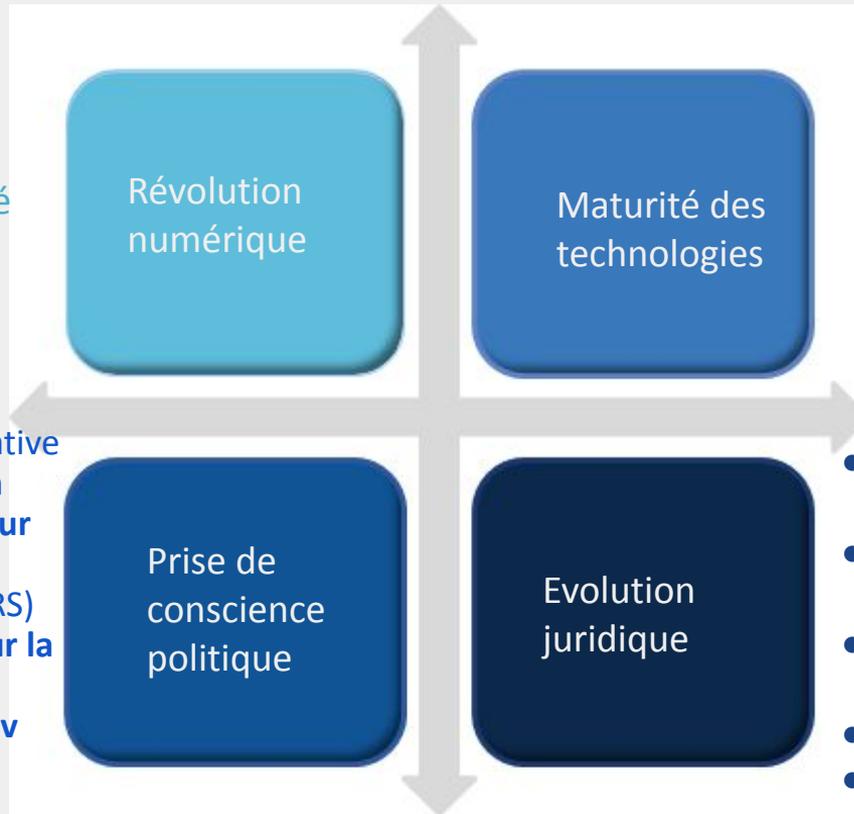
- Extraire de l'information pertinente
- Répondre à une question (recherche d'information)
- Analyser de gros volumes de textes
- Détecter des sentiments dans les textes
- Générer des résumés automatiques
- Traduire automatiquement



- 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})



- 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 Govv**
- 2023 Comité de l'IAG

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

- 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 | hyponymie (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 désambigüiser

Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TAL (traitement automatique des langues)

Stanford CoreNLP : <http://corenlp.run/>

“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**

Embedding*

“Comment transformez vous un document et son contenu en chiffres ?”

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.

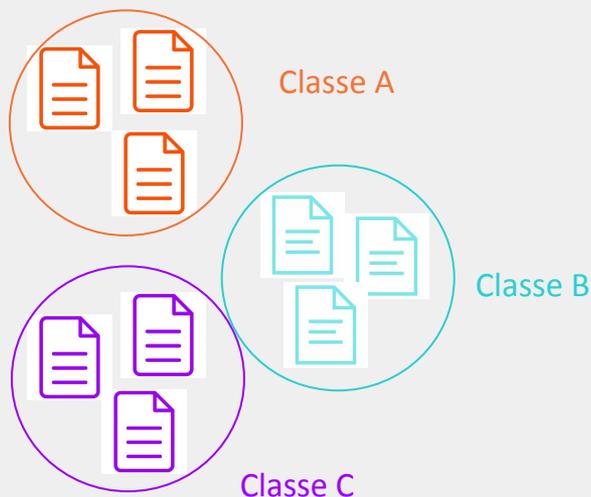
“Comment transformez vous un document et son contenu en chiffres ?” → [-0.79, 1, ... , -2.8]

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

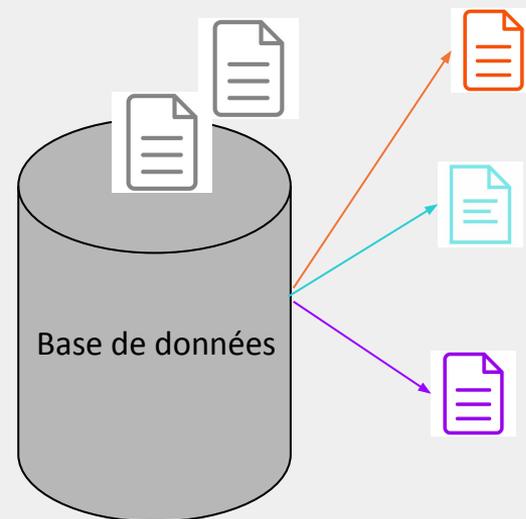
Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : **classification supervisée**

Classer les documents avec **apprentissage sur données étiquetées**

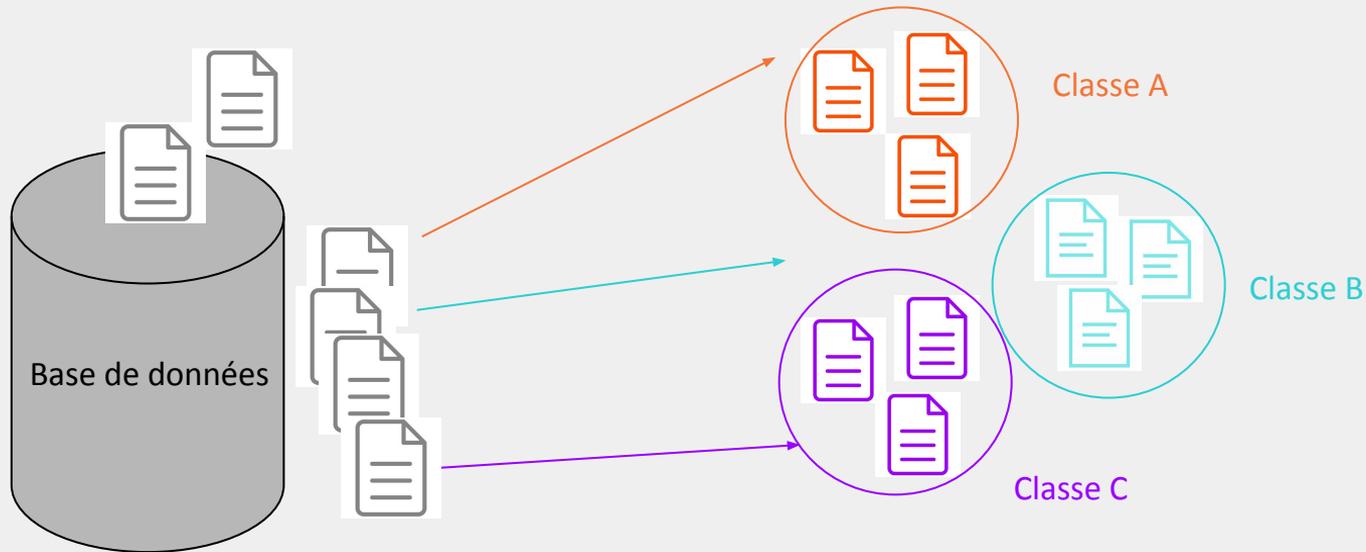


Apprentissage



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 ...



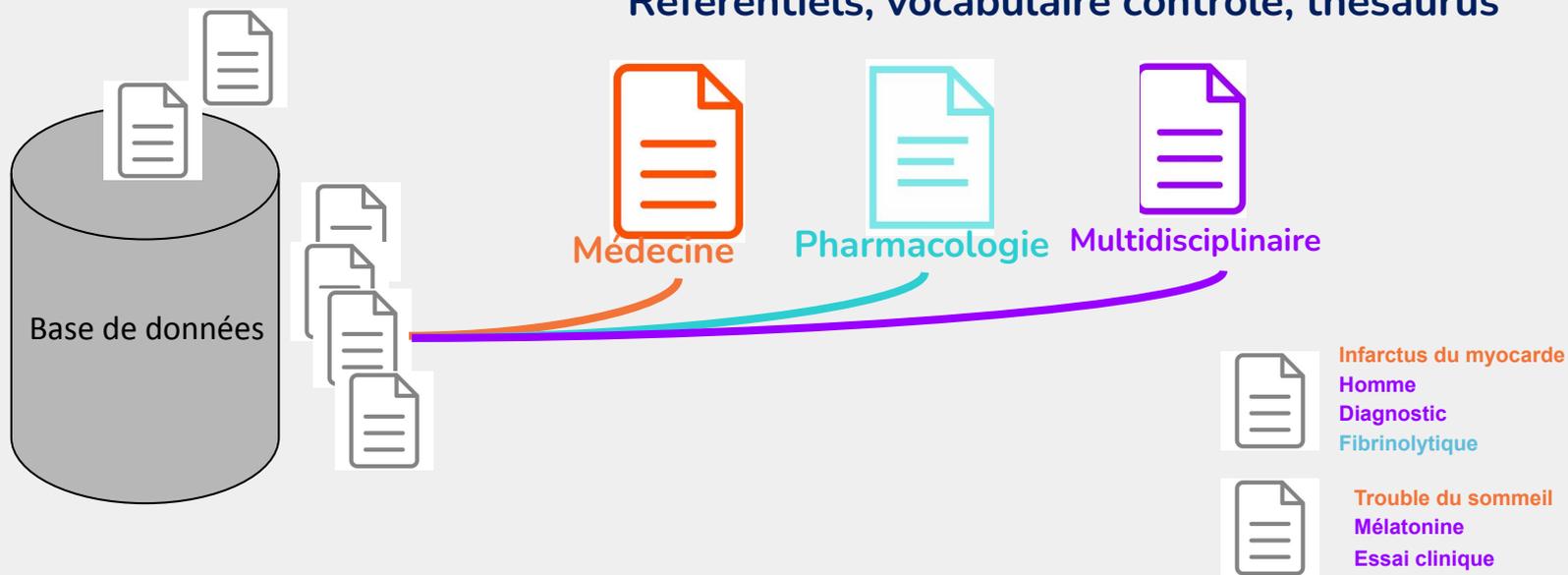
Méthodes : comparaison avec des nomenclatures
apprentissage automatique à partir d'exemples

Théorie : Définitions - Contexte - Techniques de fouille de textes

Des techniques de TDM : **Indexation libre et/ou contrôlée**

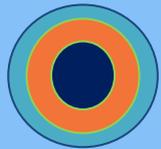
Repérage de termes **caractérisant le document** et permettant de le retrouver ensuite au sein d'un corpus

Référentiels, vocabulaire contrôlé, thésaurus

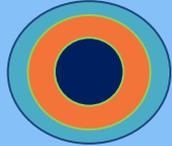




TDM à l'Inist



Environnement propice



Services et outils (démonstrations)

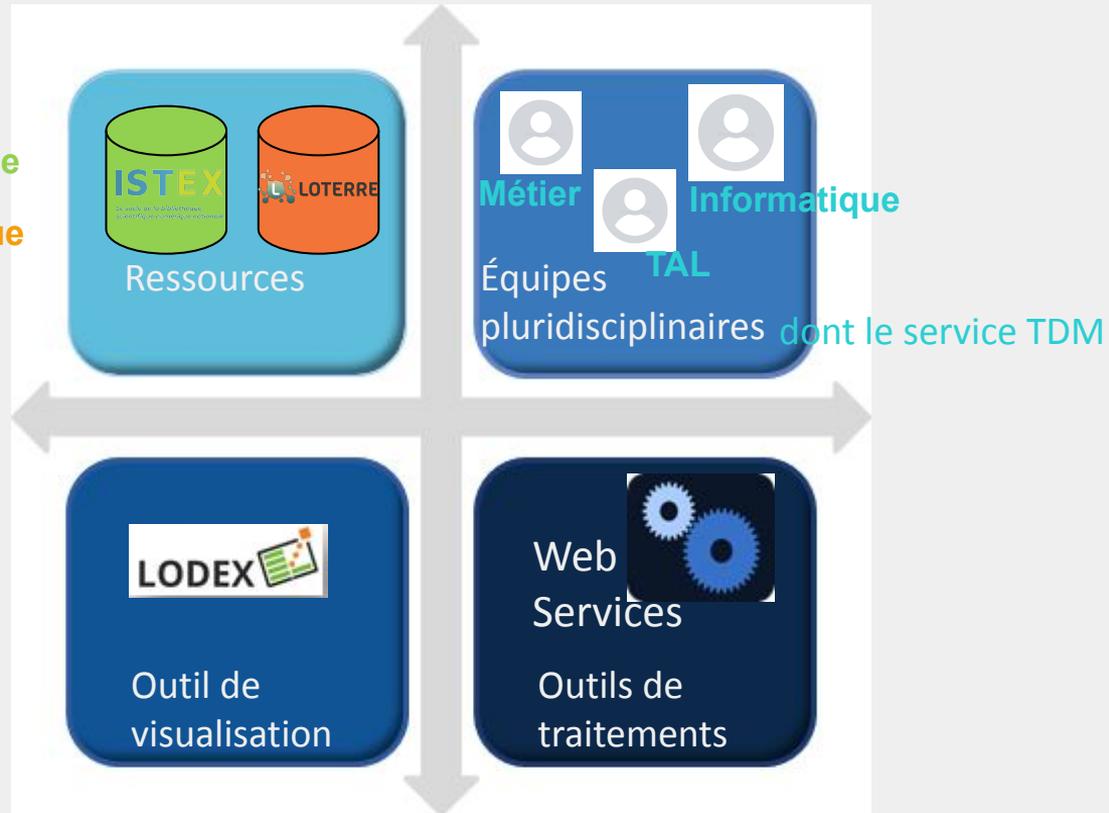


TDM à l'Inist : Environnement - Services et outils

(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

The screenshot shows the ISTEX TDM website interface. At the top, the logo 'ISTEX TDM' is displayed with the tagline 'Les services Istex pour la fouille de textes'. The URL 'https://services.istex.fr' is prominently featured. Below the header, there is a search bar with the placeholder text 'Tapez ici votre recherche, p.ex. : Classification' and a 'RECHERCHER' button. The main content area is titled 'Rechercher un web service' and displays a grid of service cards. Each card includes an icon, a title, and a brief description of the service. The services listed are: 'dataHomogenise' (HOMOGÉNÉISATION AUTOMATIQUE DE MOTS-CLÉS), 'aiAbstractCheck' (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)), and 'Adresses et affiliations' (idRorDetect ASSOCIATION D'UN IDENTIFIANT ROR À UNE ADRESSE D'AFFILIATION). A 'VOIR TOUS LES SERVICES' button is located at the bottom of the grid. On the right side, a blue sidebar contains the text 'Trouvez un service web correspondant à vos besoins', a paragraph about the tools, and a counter showing '42' services available. It also includes two buttons: 'COMMENT LES UTILISER ?' and 'VOIR LA DOCUMENTATION'.

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

Un catalogue en ligne

ISTEX TDM

Les services Istex pour la fouille de textes

Accueil > Web-services

Recherche de web-services

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

OBJET TRAITÉ

- Adresses et affiliations (3)
- Auteurs (3)
- Éléments catalographiques (3)
- Résumés (3)
- Texte intégral (3)

LANGUES (3)

TRAITEMENT (5)

- Classification (3)
- Extraction d'entités nommées (2)
- Homogénéisation (3)
- Indexation (3)
- Lémmatisation (2)

textClustering
Extraction de cluster d'un corpus

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

Extraction du texte à partir d'un pdf

Ce web service transforme un pdf en texte et extrait les éléments qui constitueront un traitement de fouille de texte ultérieur. Le pdf doit pas être un pdf image.

TermSuite
Extraction de termes d'un corpus

Ce web service s'appuie sur l'outil TermSuite pour faire une extraction terminologique à partir d'un corpus de textes en anglais ou en français. La liste des 500 termes extraits par défaut contient les termes les plus spécifiques du corpus correspondant...

countryDetect
Détection du pays d'une affiliation

Ce web service détecte le pays d'origine d'un affiliation-adresse, qu'il soit présent ou absent.

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

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

VOIR TOUS LES SERVICES

countryDetect - Détection du pays d'une affiliation

Description Utilisation Cas d'usage

Revue d'affiliation: Détectée
Revue de référence: Non

Objectif

Ce web service détecte le pays principal d'une affiliation-adresse, qu'il soit présent ou absent dans celle-ci, quelle que soit la langue de l'adresse. Il confirme quelle soit dans un pays donné. Le nom du pays retourné est en anglais.

Méthode

À partir de l'adresse d'une affiliation, le service renvoie le nom du pays détecté, en anglais, ainsi que son code ISO sur 3 lettres.
Par exemple: France, FR ou "Germany, DE".
Dans le cas où le programme ne parvient pas à déterminer le pays, il renvoie "Unknown, Null".
Le programme est capable de traiter le pays qu'il est présent dans l'adresse.
Si vous avez le programme ou souhaitez plus d'informations contactez:
• Inist
• l'Inist
• le web-service
• le nom de l'organisme.
Ces informations sont utiles pour valider l'adresse, à l'aide du service de géolocalisation [Geonames](#), et en détectant le pays.

Métriques

Le programme a été testé avec une [base de données](#) de 30%.

Évaluation:
La qualité du résultat dépend fortement des informations présentes dans l'affiliation - un nom de laboratoire ou domaine qui sont aussi utilisés est un résultat plus fiable.

TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Un catalogue en ligne

ISTEX TDM
Les services Istex pour la fouille de textes

<https://services.istex.fr>

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

(Démonstrations) **Des tests en ligne**

Droits Istex
Eduroam

ISTEX TDM

Les services Istex pour la fouille de textes

Accueil > Web-services > Extraction de termes d'un texte via Teeft

Teeft - Extraction de termes d'un texte via Teeft

Description **Utilisation** Cas d'usage

URL DU WEB SERVICE À RENSEIGNER DANS LODEX ENRICHISSEMENT EST :

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

VARIANTES ENRICHISSEMENT

Langue française

<https://terms-extraction.services.istex.fr/v2/teeft/fr>

Nombre de termes français (ex : 10)

<https://terms-extraction.services.istex.fr/v2/teeft/fr?nb=10>

DÉMONSTRATION



CODE SOURCE



The screenshot shows the Swagger UI for the 'terms-extraction' service. At the top, it says 'Swagger Supported by SMARTBEAR'. A dropdown menu shows 'Select a definition terms-extraction - Extraction de termes'. The main heading is 'terms-extraction - Extraction de termes' with version '1.9.0 OAS 3.1'. Below this, there are links to the service and terms of use. The 'terms-extraction' endpoint is highlighted, showing a 'POST /v2/teeft/en' method. The description states it returns terms with frequency and specificity. The 'Parameters' section shows 'nb' (Nombre maximal de termes à récupérer) with a value of '10' and 'indent' (Indenter le JSON résultant) set to 'true'. The 'Request body' is set to 'application/json'. An 'Example Value' section shows a JSON response with a detailed description of the Perseverance rover.

TDM à l'Inist : Environnement - Services et outils

(Démonstrations)

Des tests en ligne

Droits Istex
Eduroam



Select a definition terms-extraction - Extraction de termes

terms-extraction - Extraction de termes 1.9.0 OAS 3.1

<https://terms-extraction.services.istex.fr>

Extraction de termes à partir de textes en anglais ou en français.

Extraction de termes à partir de textes en anglais ou en français.

Terms of service

Inist-CNRS - Website

terms-extraction Extraction de termes

Plus de documentation

POST /v2/teefr/en Extrait des termes du texte en anglais en utilisant Teef

Remvoie les termes les plus spécifiques d'un texte en anglais, avec leurs fréquence et spécificité.
Permet d'avoir une idée de ce dont parle le texte. Idéalement, le texte doit contenir plusieurs paragraphes.

Par défaut `teefr` extrait 5 termes, sauf si la variable `nb` est utilisée.

Bibliographie

Cuacac P., Klefner N., Lamirel J.C. SKEEFF: indexing method taking into account the structure of the document. 20th Colinet meeting, 5-8 Nov 2019, Dallas

Parameters

| Name | Description |
|---------|--------------------------------------|
| nb | Nombre maximal de termes à récupérer |
| number | <input type="text" value="nb"/> |
| indent | Indenter le JSON résultant |
| boolean | <input type="checkbox"/> |

Request body required

Example Value Schema

```
{
  "value": "Perseverance, nicknamed Percy, is a car-sized Mars rover designed to explore the crater Jezero on Mars as part of NASA's Mars 2020 mission. It was manufactured by the Jet Propulsion Laboratory and launched on 30 July 2020, at 11:50 UTC. Confirmation that the rover successfully landed on Mars was received on 18 February 2021, at 20:55 UTC. As of 16 December 2021, Perseverance has been active on Mars for 293 sols (301 Earth days) since its landing. Following the rover's arrival, which named the landing site Octavia E. Butler Landing, Perseverance has a similar design to its predecessor rover, Curiosity, from which it was moderately upgraded. It carries seven primary payload instruments, nineteen cameras, and two microphones. The rover also carried the mini-helicopter Ingenuity to Mars, an experimental aircraft and technology showcase that made the first powered flight on another planet on 19 April 2021, since its first flight, Ingenuity has made 14 more flights for a total of 15 powered flights on another planet. The rover's goals include identifying ancient Martian environments capable of supporting life, seeking out evidence of former microbial life existing in those environments, collecting rock and soil samples to store on the Martian surface, and testing oxygen production from the Martian atmosphere to prepare for future crewed missions. The perseverance rover has four main science objectives that support the Mars Exploration Program's science goals: looking for habitability; identify past environments that were capable of supporting microbial life; seeking biosignatures; seek signs of possible past microbial life in those habitable environments, particularly in specific rock types known to preserve signs over time; caching samples; collect core rock and regolith (Soil) samples and store them on the Martian surface. Preparing for humans: test oxygen production from the Martian atmosphere. In the first science campaign Perseverance performs an arching drive southward from its landing site to the fifth unit to perform a 'Vise dig' into the unit to collect trace-venting measurements of geologic targets. After that it will return to the crater floor Fractured Bench to collect the first 'core sample' there. Back to the Octavia E. Butler landing site concludes the first science campaign. The second campaign will include several months of travel around the 'Three Forks' where Perseverance will collect rock and soil samples to store on the Martian surface, and testing oxygen production from the Martian atmosphere."
}
```

Try it out

Parameters

| Name | Description |
|------|-------------|
|------|-------------|

| | |
|---------|--------------------------------------|
| nb | Nombre maximal de termes à récupérer |
| number | <input type="text" value="nb"/> |
| (query) | |

Request body required

```
{
  "value": "Perseverance, nicknamed Percy, is a car-sized Mars rover designed to explore the crater Jezero on Mars as part of NASA's Mars 2020 mission. It was manufactured by the Jet Propulsion Laboratory and launched on 30 July 2020, at 11:50 UTC. Confirmation that the rover successfully landed on Mars was received on 18 February 2021, at 20:55 UTC. As of 16 December 2021, Perseverance has been active on Mars for 293 sols (301 Earth days) since its landing. Following the rover's arrival, which named the landing site Octavia E. Butler Landing, Perseverance has a similar design to its predecessor rover, Curiosity, from which it was moderately upgraded. It carries seven primary payload instruments, nineteen cameras, and two microphones. The rover also carried the mini-helicopter Ingenuity to Mars, an experimental aircraft and technology showcase that made the first powered flight on another planet on 19 April 2021, since its first flight, Ingenuity has made 14 more flights for a total of 15 powered flights on another planet. The rover's goals include identifying ancient Martian environments capable of supporting life, seeking out evidence of former microbial life existing in those environments, collecting rock and soil samples to store on the Martian surface, and testing oxygen production from the Martian atmosphere to prepare for future crewed missions. The perseverance rover has four main science objectives that support the Mars Exploration Program's science goals: looking for habitability; identify past environments that were capable of supporting microbial life; seeking biosignatures; seek signs of possible past microbial life in those habitable environments, particularly in specific rock types known to preserve signs over time; caching samples; collect core rock and regolith (Soil) samples and store them on the Martian surface. Preparing for humans: test oxygen production from the Martian atmosphere. In the first science campaign Perseverance performs an arching drive southward from its landing site to the fifth unit to perform a 'Vise dig' into the unit to collect trace-venting measurements of geologic targets. After that it will return to the crater floor Fractured Bench to collect the first 'core sample' there. Back to the Octavia E. Butler landing site concludes the first science campaign. The second campaign will include several months of travel around the 'Three Forks' where Perseverance will collect rock and soil samples to store on the Martian surface, and testing oxygen production from the Martian atmosphere."
}
```

Execute

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":"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)

Enrichissements exécutables et paramétrables depuis Lodex

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

Filtre Istex TDM : Type de données "Documents"

The screenshot displays the LODEX interface with a sidebar on the left and a main configuration area on the right. The sidebar contains a menu with 'LODEX' at the top, followed by 'DONNÉES' (highlighted with a purple box), 'AFFICHAGE', and 'ANNOTATIONS'. Below this, there are options for 'Données', 'Enrichissements' (highlighted with a purple box), 'Précalculs', and 'Ressources cachées'. The main configuration area has a top navigation bar with buttons for 'TOUS', 'AFFILIATION', 'CLASSIFICATION', 'ENTITÉS NOMMÉES', 'GÉOGRAPHIE', 'HOMOGÉNÉISATION', and 'INDEXATION'. Below this are buttons for 'MÉTADONNÉES', 'PRÉTRAITEMENT', 'TRAITEMENT AUTOMATIQUE DE LA LANGUE', 'VALIDATION', and 'AUTRE'. The main area contains several service configuration cards, each with a title, description, and a 'Mode avancé' toggle. The cards are: 'addressSplit - Décomposition d'une adresse', 'Associer un terme au vocabulaire Pays et Subdivision', 'Associer un terme au vocabulaire des communes de France', 'astroTag - Extraction d'entités nommées en astronomie', 'aiAbstract-check - Détection de résumé scientifique généré par IA', and 'authorDistinct - Désambiguïsation d'auteurs via ORCID'. Each card has a 'Nom' field (labeled '1'), a 'URL du web service' field (labeled '2'), and a 'Colonne de la source' dropdown menu (labeled '3'). The 'authorDistinct' card also has a 'Sous-chemin' field. At the bottom right, there is an 'ANNULER' button.

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"

The screenshot displays the LODEX web interface. At the top, a green navigation bar contains the 'LODEX' logo, a 'DONNÉES' menu (highlighted with a purple box), and buttons for 'AFFICHAGE' and 'ANNOTATIONS'. On the right of the bar is a 'DÉPUBLIER' button. A left sidebar lists navigation options: 'Données', 'Enrichissements', 'Précalculs' (highlighted with a purple box), and 'Ressources cachées'. The main content area is divided into two columns. The left column contains a form with three fields: 'Nom' (labeled '1'), 'URL du web service' (labeled '2'), and 'Colonne de la source' (labeled '3') with a dropdown arrow. Below the form is a 'Mode avancé' toggle switch. The right column features a filter bar with buttons for 'TOUS', 'CLASSIFICATION', 'HOMOGENÉISATION', 'INDEXATION', 'VALIDATION', and 'AUTRE'. Below this are three service cards: 1. 'dataHomogenise - Homogénéisation automatique de mots-clés' with a description and a 'TOUS' button. 2. 'Graph-segment' with a description and a 'TOUS' button. 3. 'IdaClass - Extraction de 15 thématiques d'un corpus' with a description and a 'TOUS' button. Below that is another 'IdaClass - Extraction de 8 thématiques d'un corpus' card with a description and a 'TOUS' button. At the bottom is an 'IdaSegment - Extraction de 15 thématiques à partir d'un jeu de données avec un format adapté à la création de graphiques' card with a description and a 'TOUS' button. Each card includes a 'TOUS' button and a 'DÉPUBLIER' button.

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

Web services exécutables depuis IA Factory

Droits Istex
Eduroam

ISTEX IA Factory
L'IA appliquée à vos corpus

CHARGEZ VOS CORPUS ET DÉCOUVREZ LES RÉSULTATS DES SERVICES TDM

IA Factory est une interface de chargement de corpus et d'exécution d'outils de TDM vous permettant d'exploiter vos propres données en simplement 3 étapes :

- Téléchargez vos données et choisissez le format et le champ à traiter,
- Choisissez le service web de TDM que vous voulez exécuter,
- Remplissez votre adresse mail.

À l'issue du traitement vous recevrez un mail avec un lien de téléchargement du résultat.



 Déposer un fichier ou glisser-déposer ici

SUIVANT

- [TermSuite](#)
- [Teft](#)
- [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.

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"

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.



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

Instance Lodex avec les différents web services

ISTEX Services <https://tdm.inist.fr/instance/demo-webservices>

Les technologies et les outils ISTEX pour les projets de recherche.

Instance Lodex modèle pour les web services avec des données issues d'Istex

Corpus - Nombre de publications (notices issues d'ISTEX au format targz).

50

Description

Cette instance avec peu de données et sans thématique particulière a pour objectif de :

- montrer les résultats des traitements des [web services](#)
- proposer un modèle pour l'utilisation des [web services](#) et les représentations graphiques de leurs résultats sur des données issues d'Istex.

Vous pourrez adapter votre modèle en fonction de vos besoins.

Deux vidéos sont à votre disposition sur CanalU

- [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](#)

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

ISTEXTM Tools explorer

Catalogue d'outils libres pour la fouille de textes

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

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

Explorez divers outils de TDM

Cette application Lodox a été créée pour visualiser de manière simple des références d'outils de TDM sélectionnés depuis une liste de trois cents outils spécialisés dans le traitement automatique du langage et l'exploration de texte.

Pour en savoir plus

MÉTHODOLOGIE

Le catalogue TM Tools explorer

AbLang

Modèle de langue sur les anticorps.



ABLTagger

Etiqueteur morphosyntaxique pour l'islandais.



ABNER

ABNER est un outil logiciel open source pour l'analyse...



Adaboost

AdaBoost est un méta-algorithme de classification...



Actuellement **548** outils sont référencés dans ce catalogue



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

TRAITEMENT (7)

- Classification (9)
- Extraction d'entités nommées (9)
- Homogénéisation (9)
- Indexation (7)
- Traitement automatique du langage (3)
- Prétraitement (4)
- Validation (3)

OBJET TRAITÉ

- Adresses et affiliations (10)
- Auteurs (2)
- Éléments catalographiques (4)
- Citations (1)
- Résumés (21)
- Texte intégral (22)

Merci pour votre attention

