

# Introduction au TDM

ANF Corpus ISTEX

06 novembre 2025

Léo GAILLARD

Ingénieur en fouille de données et en fouille de textes

[leo.gaillard@inist.fr](mailto:leo.gaillard@inist.fr)

Qu'est ce que le  
TDM ?

L'évolution du  
TDM

Les enjeux du  
TDM

Quelques outils  
de TDM

L'INIST-CNRS et le  
TDM





Qu'est ce que le  
TDM ?

Photo de Codioful (Formerly Gradienta)  
sur Unsplash

# La fouille de textes

4

## DEFINITION

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

Données □ Connaissances

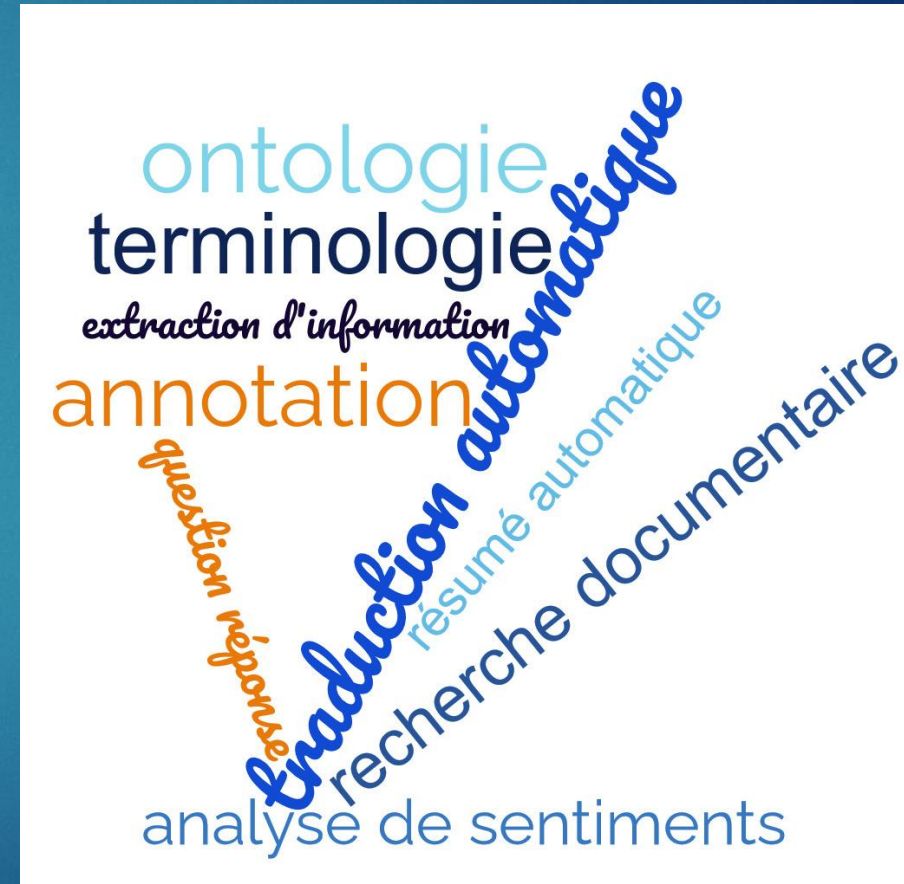
C'est une spécialisation de la fouille de données (*data mining*) qui fait appel aux méthodes de l'**Intelligence Artificielle**<sup>1</sup>, du **Traitement Automatique des Langues** et des **Statistiques**.

<sup>1</sup> L'**apprentissage profond** ou **apprentissage en profondeur**<sup>1</sup> (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

# La fouille de textes : des technologies qui nous accompagnent déjà largement au quotidien...

5

- Filtrage de spam
- Recommandations
- Assistant personnel
- Service client, agent conversationnel
- Intelligence économique
- Intelligence stratégique
- Sécurité
- Assistance au diagnostic médical
- Génération textuelle
- Recherche scientifique
  - Détection de textes générés par IA
  - Validation de références bibliographiques
  - Résumé automatique
  - Indexation automatique
    - Mots-clés
    - Entités nommées
    - ...



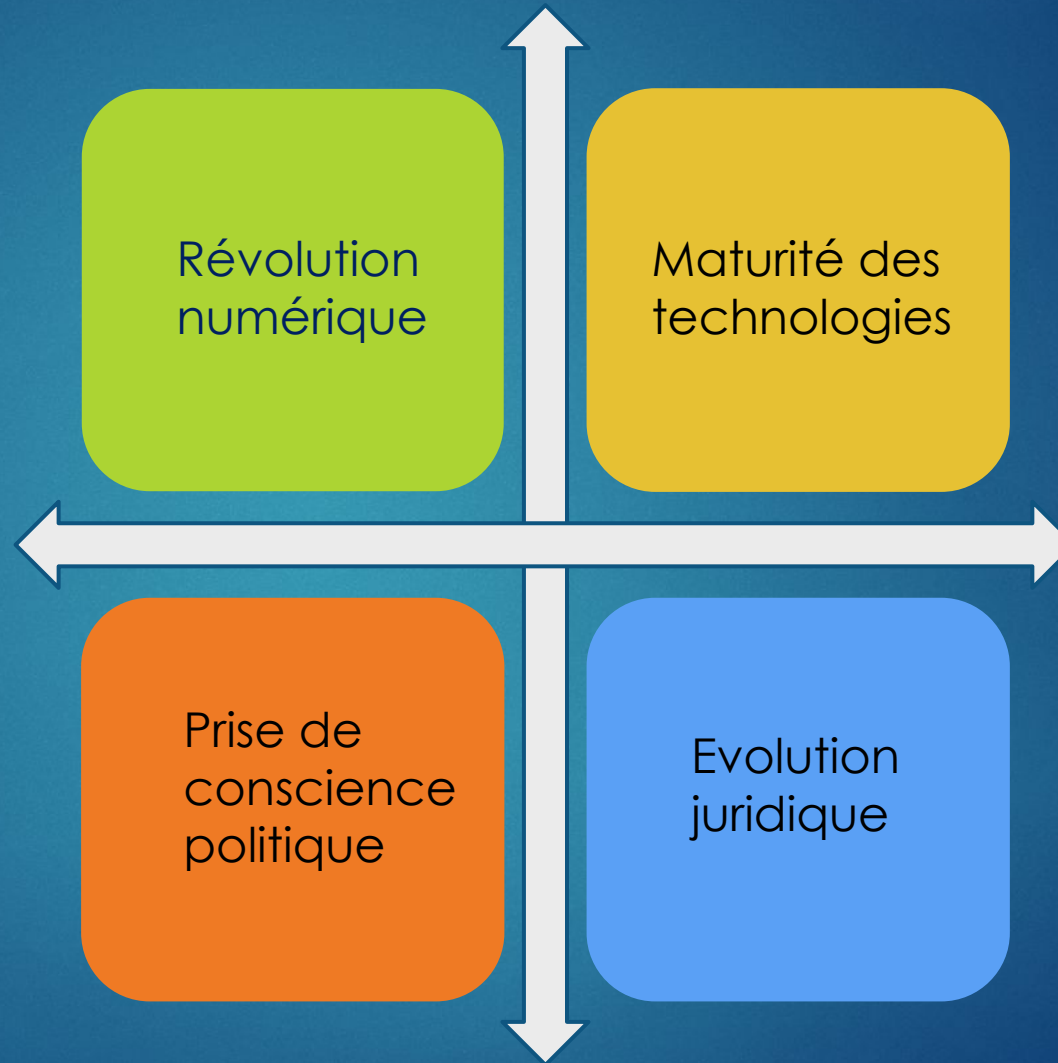




## L'évolution du TDM

# CONTEXTE

7





Nous ne sommes plus en capacité d'absorber la quantité d'information disponible...

**déluge** d'informations | **augmentation** des types de données produits.

Ere du **Big Data** ; les 3V : **V**olume, **V**élocité et **V**ariété

Le phénomène Big Data s'amplifie si vite que l'on n'arrive plus à suivre l'évolution des nouvelles unités de mesure : les **exaoctets** ( $10^{18}$  octets), les **zettaoctets** ( $10^{21}$ ), les **yottaoctets** ( $10^{24}$ )....

**> 180 zettaoctets en 2025**

## Publications scientifiques

*50% des articles ne sont **jamais lus**  
90% des articles ne sont **pas cités***

ANF TDM 2020/ R. Bossy & C. Nédellec

8

Révolution  
numérique



Image générée par [bing](#)



... mais nous disposons de technologies de plus en plus performantes

9

Maturité des technologies

30 ans d'expérience en **TAL et IA** (cf ChatGPT...), en partie majorée par l'**implication d'industriels** qui y trouvent un intérêt majeur (analyse de sentiments, de tendances, détection de buzz etc.)

Augmentation très importante de la **puissance de calcul et de stockage** en 40 ans

Évolution majeure des algorithmes : **statistiques > apprentissage profond > LLM (Large Languages Models)**

## Le TDM va s'inscrire dans la politique de science ouverte ...

On cherche à s'affranchir de la mainmise des éditeurs scientifiques sur les publications et les données de la science et à permettre une meilleure reproductibilité de la recherche.

10

Prise de conscience politique

- |       |                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2001  |    | <b>Budapest Open Initiative:</b> problématique du libre accès aux <u>publications scientifiques</u> et incitation à l'utilisation des archives ouvertes ou des revues en libre accès, prise de conscience des besoins en licences adaptées                                                                                                                                                                               |
| 2003  |    | <b>Déclaration de Berlin:</b> extension de l'ouverture aux <u>données de la recherche</u>                                                                                                                                                                                                                                                                                                                                |
| (...) |                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                          |
| 2018  |    | <b>Rapport Villani sur l'I.A</b> « Favoriser sans attendre les pratiques de fouille de texte et de données (TDM) » (page 35)<br><b>1<sup>er</sup> Plan national pour la Science ouverte</b> - Frédérique VIDAL- MESRI <b>5 M € /an</b><br>« La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans paiement.» |
| 2019  |    | <b>Le Grand Débat:</b> le TDM devient une « réalité publique » <a href="https://iscpif.fr/chavalarias/?p=1495">https://iscpif.fr/chavalarias/?p=1495</a><br><b>Feuille de route pour la science ouverte du CNRS</b><br><b>Engagement des universités :</b> politiques et interlocuteurs désignés pour la science ouverte                                                                                                 |
| 2021  |  | <b>2e Plan national pour la science ouverte (2021-2024): 15 M € /an</b><br>« Transformer les pratiques pour faire de la science ouverte le principe par défaut » 100% de publications en accès ouvert en 2030                                                                                                                                                                                                            |
| 2022  |  | <b>Plateforme Recherche Data Gouv</b>                                                                                                                                                                                                                                                                                                                                                                                    |
| 2025  |  | <b>Développement de LLM européens, ouverts, multilingues et spécialisés 80 M € / 3 ans</b>                                                                                                                                                                                                                                                                                                                               |



## ... et bénéficier des dispositions légales qui sont prises

11

Evolution  
juridique

2016



### Loi pour une République numérique:

#### L'article 38 : Exceptions au code de la propriété intellectuelle

« Conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche publique. »

Introduction d'une **exception au droit d'auteur** ainsi qu'une **exception au droit sui generis des producteurs de bases de données**

2019



### Directive européenne sur le droit d'auteur et les droits voisins dans le marché unique du numérique ou Directive

#### « Copyright »:

Les **articles 3 et 4 de la directive**, portent sur la "fouille de textes et de données à des fins de recherche scientifique" ; la pratique du TDM (text and data mining). Ces articles prévoient une exception au droit d'auteur "pour les reproductions et les extractions effectuées par des organismes de recherche et des institutions du patrimoine culturel, en vue de procéder, à des fins de recherche scientifique, à une fouille de textes et de données sur des œuvres ou autres objets protégés auxquels ils ont **accès de manière licite**"

2021



### Ordonnance de transposition en droit français de la Directive européenne sur le droit d'auteur:

<https://www.vie-publique.fr/loi/282569-ordonnance-completant-transposition-directive-droits-dauteur>

" L'ordonnance consacre ou adapte tout d'abord des **exceptions au droit d'auteur et aux droits voisins** afin de favoriser la **fouille de textes et de données**, l'utilisation d'extraits d'œuvres à des fins **d'illustration dans le cadre de l'enseignement** et la reproduction des œuvres dans un souci de conservation du patrimoine culturel."

2022



### Décret n°2022-928 du 23 juin 2022:

<https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000045960058>

Ce décret fait suite à l'ordonnance du 24 novembre 2021 ci-dessus. Il introduit des modifications du code de la propriété intellectuelle et formalise les **modalités d'application de l'exception en vue de la fouille de textes** et de données (conditions de détention des copies numériques nécessaires à la fouille de textes entre autres)

# Quand le droit et la politique s'allient...

12



Budapest Open Initiative

Déclaration de Berlin

2001



2003



2016



2018



Directive  
européenne sur  
le droit d'auteur



2019



2e Plan national pour la science ouverte (2021-2024)

Ordonnance de  
transposition en droit  
français de la DE

2021



Plateforme Recherche Data Govv

2022



Décret n°  
2022 du 23  
juin 2022

2025



LLM open source  
européens

80 M € - 3ans

Rapport Villani sur l'I.A

1er Plan national pour la Science ouverte

5 M € /an

Le Grand Débat

Feuille de route pour la Science ouverte du CNRS

Engagement des universités

Loi pour une  
République  
numérique





**Les enjeux du  
TDM**

# DES DIFFICULTES ET DES SOLUTIONS

Le TDM repose sur:

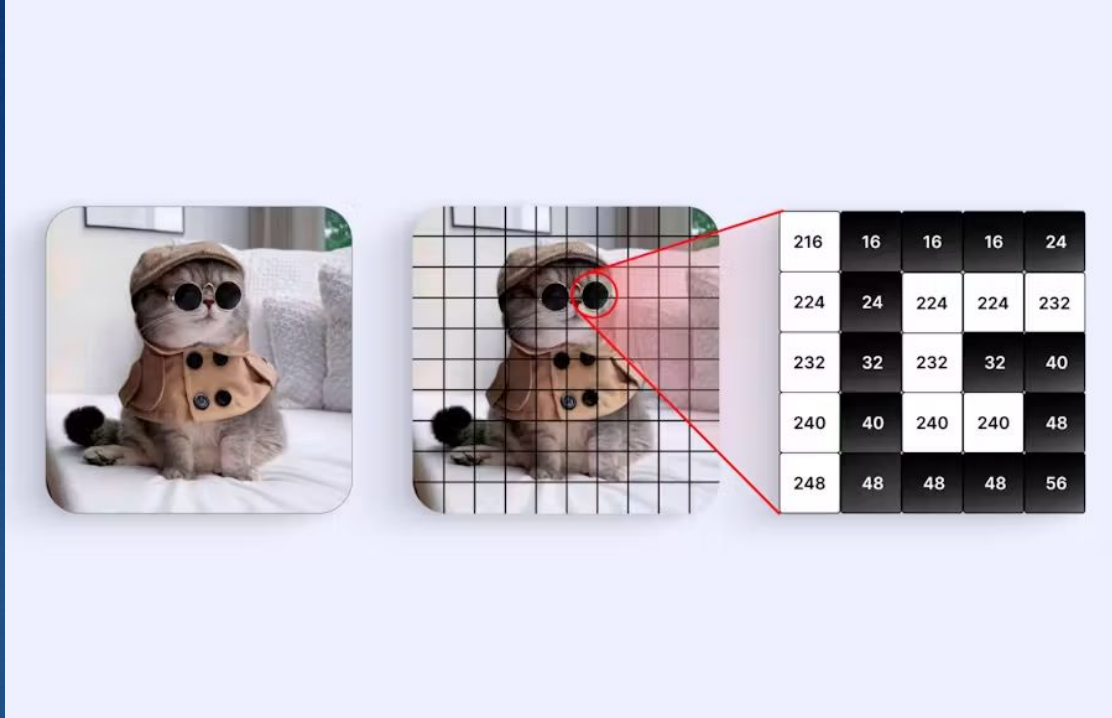
1. l'exploitation de **texte**
2. des traitements automatiques du **langage naturel**
3. des traitements informatiques et mathématiques basés sur des outils d'**intelligence artificielle**

**Première difficulté : modélisation  
mathématique des données**



# Le cas particulier du texte ...

15



[encord](#), 9/sept/2025

Représenter vectoriellement une image ou un son peut être assez intuitif.



[Codelabs](#), 9/sept/2025

... et pour le **texte** ?

15

# 1/ Le texte est une donnée mais avec des caractéristiques spécifiques...

16

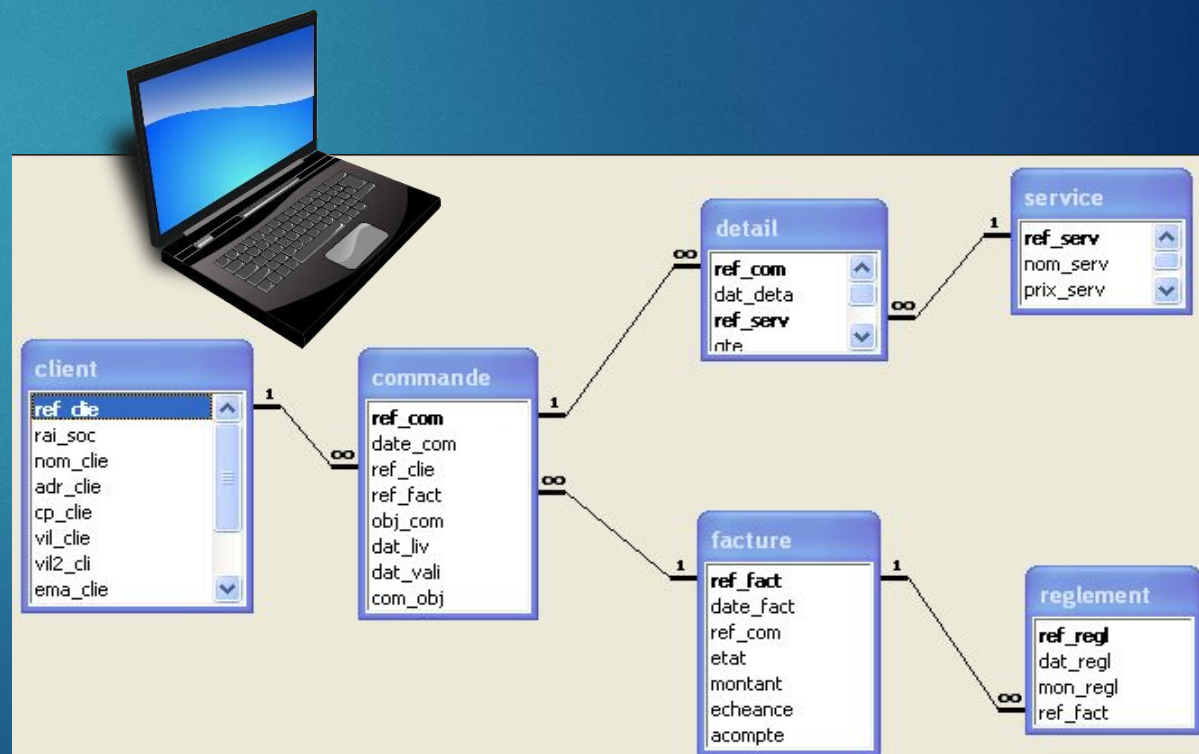
Le texte est une **donnée non structurée**



Un ordinateur interprète de la **donnée structurée**

« Vous trouverez par la présente le courrier de M. Dupont qui honore le règlement de sa commande du 22 mai 2019 au sujet de l'achat d'une caisse de 12 bouteilles de Bourgogne »

**QUESTION:** la facture de M. Dupont est-elle payée ?





## 2/ la langue est complexe

17

Pour interpréter et comprendre...

|                       |                                 |
|-----------------------|---------------------------------|
| <b>Paris</b>          | capitale de la France, ville US |
| <b>ne... pas...</b>   | négation                        |
| <b>Orange</b>         | couleur, fruit, société, ville  |
| <b>Labrador</b>       | hyponymie (chien)               |
| <b>Boire un verre</b> | 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** : comment sont construites les phrases

La **sémantique** des mots: désambiguïsation

### 3/ Quelques techniques de TDM



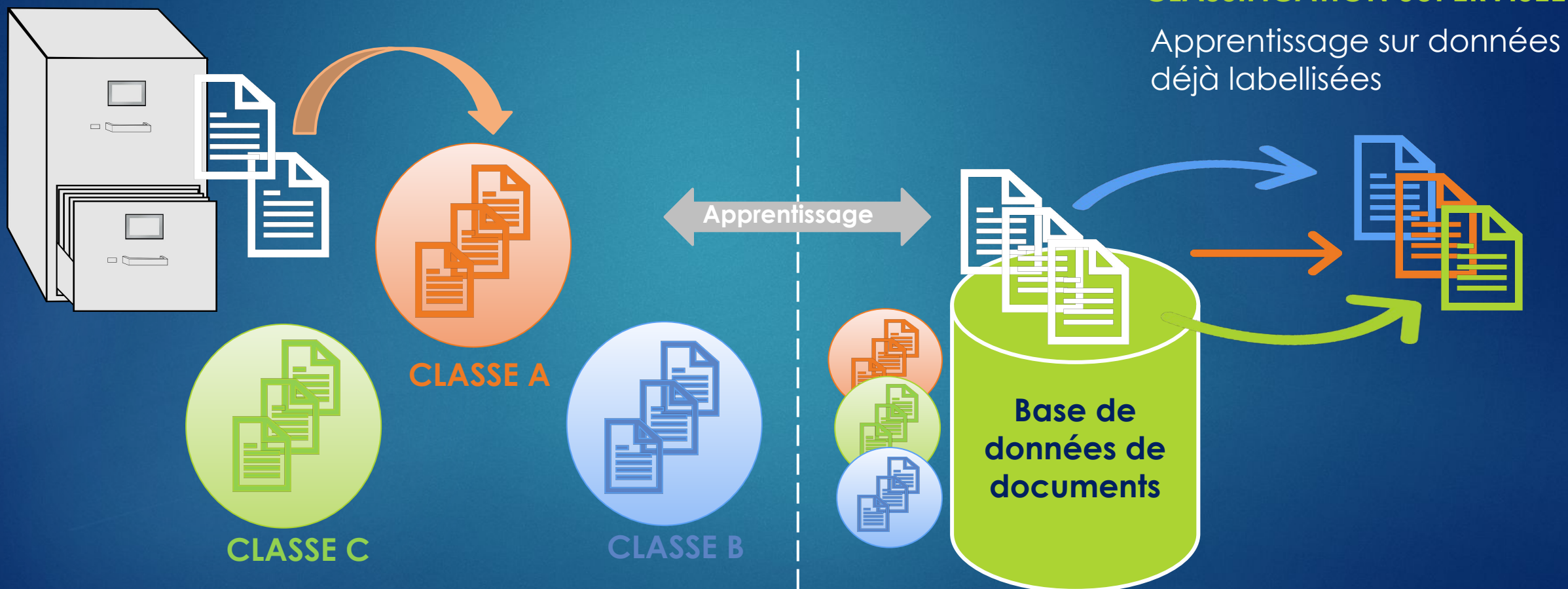
### 3/ Quelques techniques de TDM - CLASSIFICATION

19

#### Problématique

Classer les documents suivant (par exemple):

- les thèmes de ces documents
- les zones géographiques considérées...



### 3/ Quelques techniques de TDM - Reconnaissance d'entités nommées

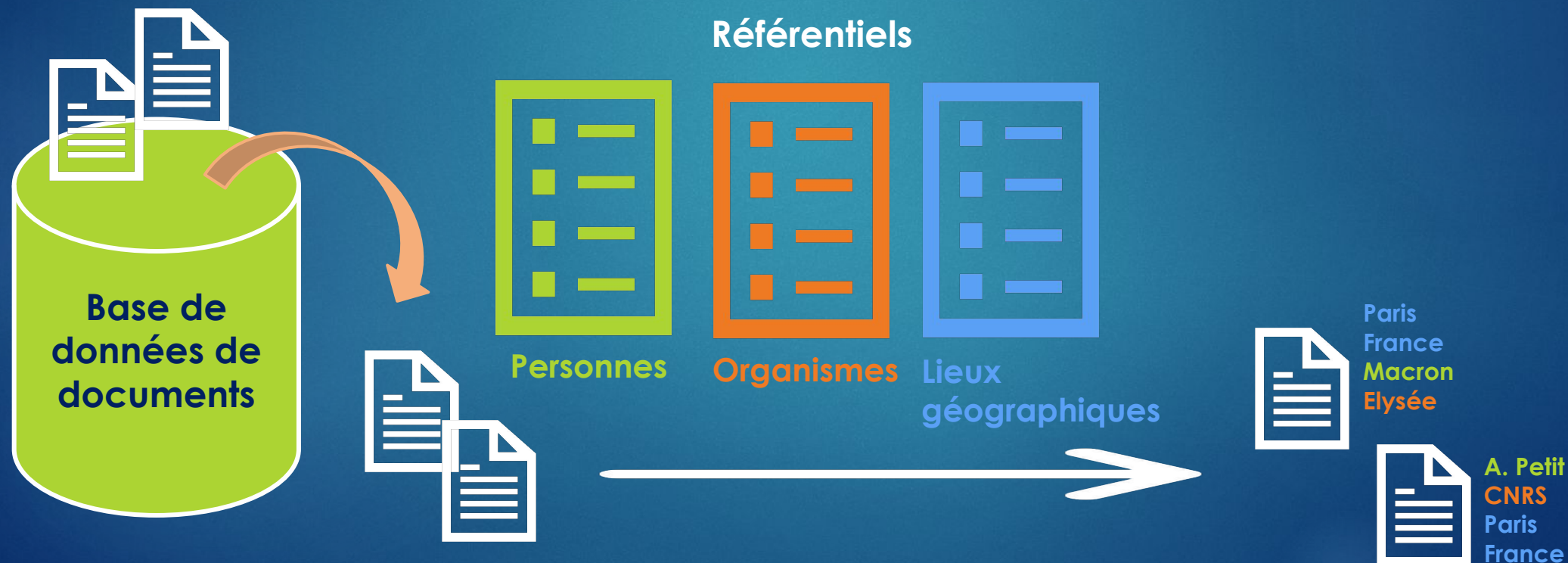
#### Problématique

Repérage de :  
Personnes, lieux géographiques, institutions, sociétés,  
microorganismes...

#### Réponses possibles

Diverses méthodes possibles : statistiques, utilisation de  
référentiels, deep learning, ...

## EXTRACTION D'INFORMATION RECONNAISSANCE D'ENTITES NOMMEES







**Quelques outils  
de TDM**

# 1/ Structuration des données

22

« 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)

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

ADV VERB PRON DET NOUN CCONJ PRON NOUN PREP NOUN PUNCT

## Lemmatisation

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

transformer chiffre





Corpus :

1. « Comment transformer le contenu d'un document en chiffres ? »
2. « Je l'ai transformé en chiffres ! »

| <u>doc</u> | comment | transformer | le | contenu | de | un | document | en | chiffre | je | avoir |
|------------|---------|-------------|----|---------|----|----|----------|----|---------|----|-------|
| <u>1.</u>  | 1       | 1           | 1  | 1       | 1  | 1  | 1        | 1  | 1       | 0  | 0     |
| <u>2.</u>  | 0       | 1           | 1  | 0       | 0  | 0  | 0        | 1  | 1       | 1  | 1     |



Corpus :

1. « Comment transformer le contenu d'un document en chiffres ? »
2. « Je l'ai transformé en chiffres ! »

| <u>doc</u> | comment | transformer | le | contenu | de | un | document | en | chiffre | je | avoir |
|------------|---------|-------------|----|---------|----|----|----------|----|---------|----|-------|
| <u>1.</u>  | 1       | 1           | 1  | 1       | 1  | 1  | 1        | 1  | 1       | 0  | 0     |
| <u>2.</u>  | 0       | 1           | 1  | 0       | 0  | 0  | 0        | 1  | 1       | 1  | 1     |

↓  
Après suppression des mots vides

| <u>doc</u> |  | transformer |  | contenu |  |  | document |  | chiffre |  |  |
|------------|--|-------------|--|---------|--|--|----------|--|---------|--|--|
| <u>1.</u>  |  | 1           |  | 1       |  |  | 1        |  | 1       |  |  |
| <u>2.</u>  |  | 1           |  | 0       |  |  | 0        |  | 1       |  |  |

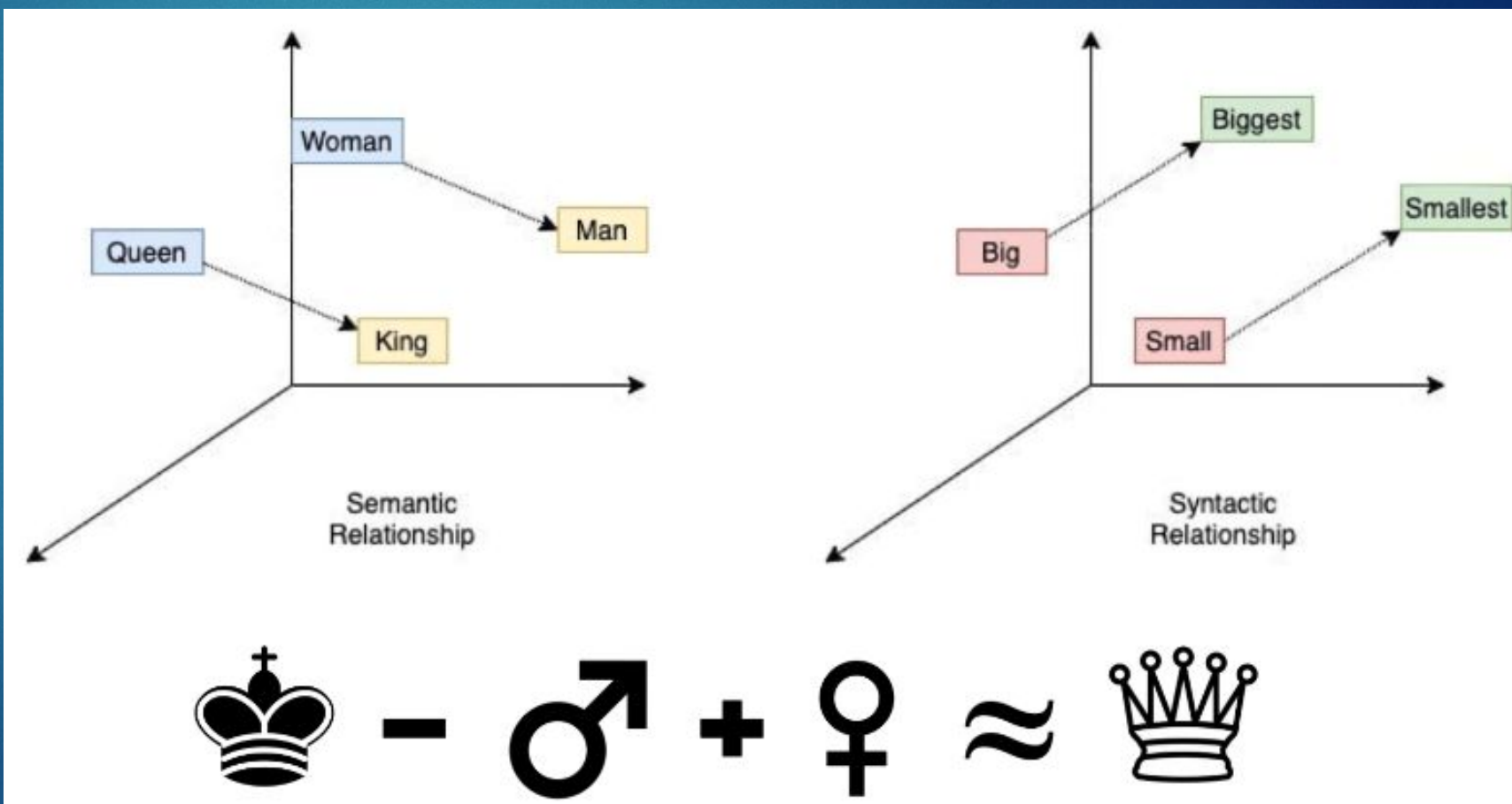
## 2/ Vectorisation des données

26

### embedding

Phrase : On attribue ici à chaque mot un vecteur fixe, calculé pour chaque mot.

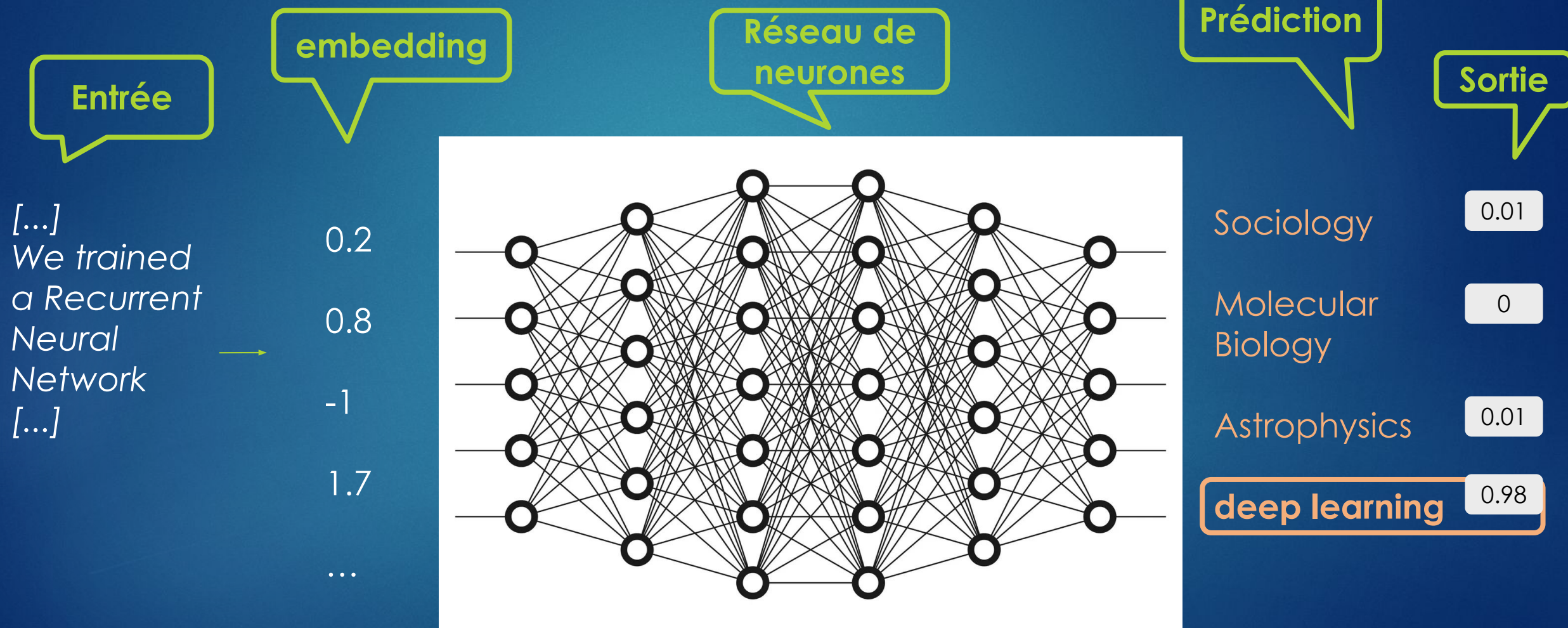
roi  $\longrightarrow$  [0.01, 0.8, -0.1, ..., 0.2, -1.4]





### 3/ Qu'est-ce qu'un réseau de neurones artificiel ? (deep learning)

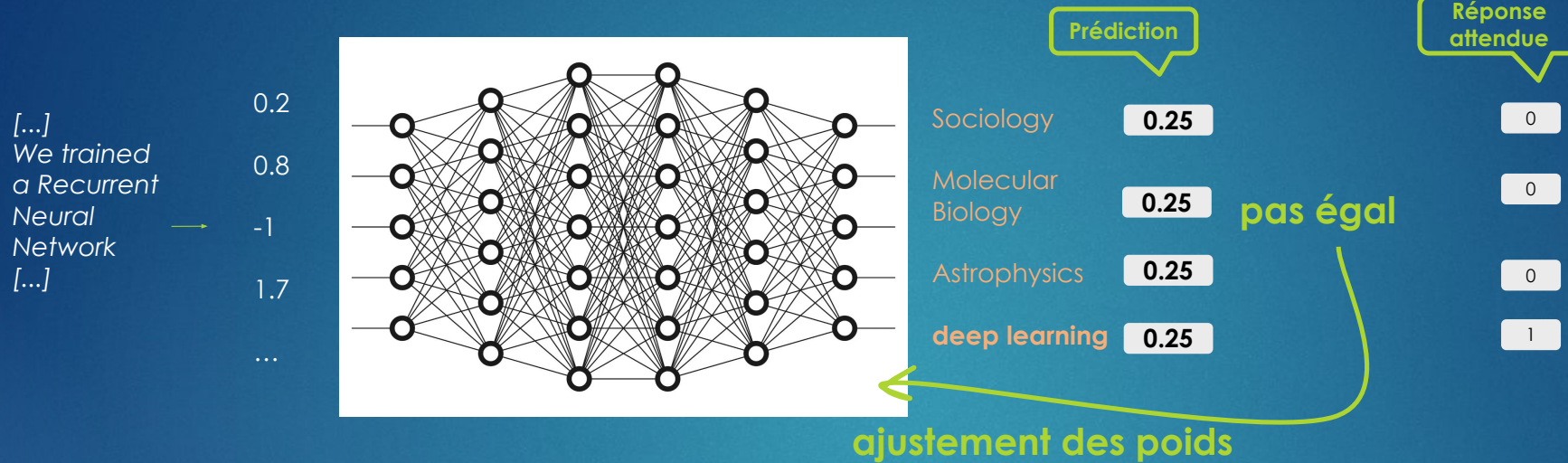
Exemple : Classification en domaine scientifique



### 3/ Qu'est-ce qu'un réseau de neurones artificiel ? (deep learning)

28

Avant l'entraînement



Entraîner un réseau de neurones requiert un jeu de données labellisées ...



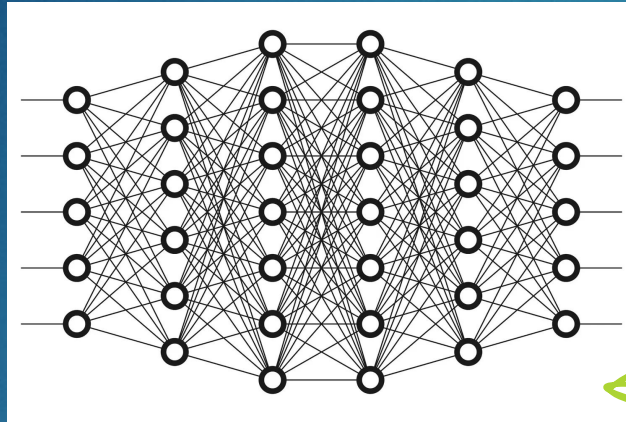
### 3/ Qu'est-ce qu'un réseau de neurones artificiel ? (deep learning)

29

#### Avant l'entraînement

[...]  
We trained  
a Recurrent  
Neural  
Network  
[...]

0.2  
0.8  
-1  
1.7  
...



Prédiction

Sociology 0.25  
Molecular Biology 0.25  
Astrophysics 0.25  
deep learning 0.25

Réponse  
attendue

0  
0  
0  
1

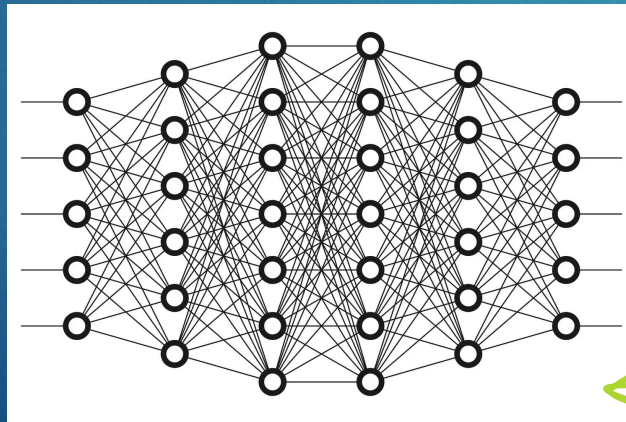
pas égal

ajustement des poids

#### Pendant l'entraînement

[...]  
We trained  
a Recurrent  
Neural  
Network  
[...]

0.2  
0.8  
-1  
1.7  
...



Prédiction

Sociology 0.12  
Molecular Biology 0.08  
Astrophysics 0.2  
deep learning 0.6

Réponse  
attendue

0  
0  
0  
1

Mieux mais  
pas égal

ajustement des poids

Entraîner  
un réseau  
de  
neurones  
requiert un  
jeu de  
données  
labellisées  
...

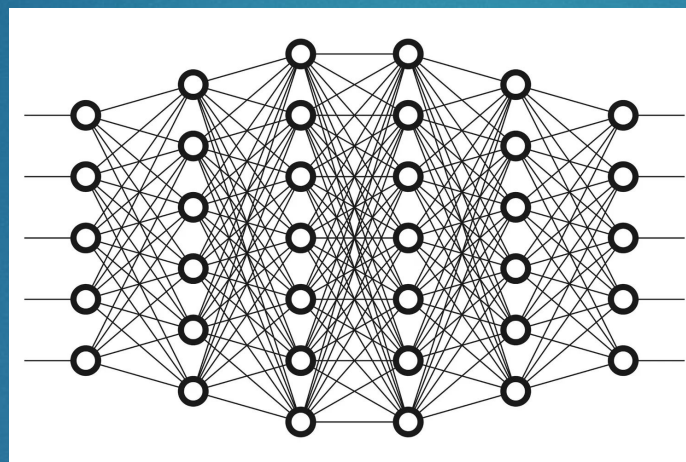
## 4/ Comment fonctionne l'IA générative ?

30

This  
is  
a  
<MASK>  
on  
Artificial  
Intelligence

Donnée à  
prédire  
(ici, un mot)

LLM



Prédiction

presentation

0.58

draft

0.15

advance

0.05

...

water

0

Classification  
parmi plus de 50000  
classes  
(= taille du  
vocabulaire)



## 4/ Comment fonctionne l'IA générative ?

### PROMPT :

"Give me the definition of AI."



### Ce que reçoit le modèle :

Give me the definition of AI. <MASK>

### Ce que retourne le modèle :

The

On réitère...

### Ce que reçoit le modèle :

Give me the definition of AI. The  
<MASK>

### Ce que retourne le modèle : definition

On réitère...

31

## Nouvelles technologies implique nouvelles problématiques

- Toutes les problématiques sur les données évoquées précédemment

- Toxicité du modèle

Dépend de la provenance et de la qualité des données

- Biais du modèle

Exacerbés vu le fonctionnement des LLM

- Coût écologique et économique :

- Pour l'entraînement

"ferme de GPU"

- A l'utilisation

1 requête chatgpt ⇔ 10 requêtes google



**L'INIST-CNRS et le  
TDM**



## Qu'est ce qu'un web service ?

33

Web service (WS): 1 outil = 1 tâche

□ **Aucun besoin de connaissance a priori, aucun paramétrage nécessaire**



## Qu'est ce qu'un web service ?

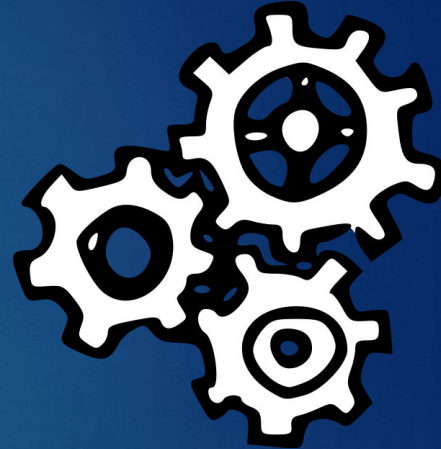
34

Web service (WS): 1 outil = 1 tâche

□ **Aucun besoin de connaissance a priori, aucun paramétrage nécessaire**

La complexité de nos WS dépend de plusieurs facteurs :

- Type de tâches (classification, extraction, indexation ...)
- Types de données (résumé, métadonnées, texte intégral ...)





## Qu'est ce qu'un web service ?

Web service (WS): 1 outil = 1 tâche

□ **Aucun besoin de connaissance a priori, aucun paramétrage nécessaire**

La complexité de nos WS dépend de plusieurs facteurs :

- Type de tâches (classification, extraction, indexation...)
- Types de données (résumé, métadonnées, texte plein...)

- TDM Factory : interface web de l'INIST pour utiliser un WS simplement
- Lodex : outil de data visualization (nos web services peuvent y être utilisés)



# Web services : de la simplicité à la complexité !

36

Quelques exemples de web services :

- **Extraction de termes avec Teeft**
- Détection de genre d'un auteur
- Classification en domaines scientifiques Pascal-Francis

## Extraction de termes d'un texte via Teeft

Le service web teeft extrait les termes les plus pertinents d'un texte en anglais ou en français. Il permet d'avoir une idée de ce dont parle le texte. Idéalement, le texte doit contenir plusieurs paragraphes. Par défaut teeft extrait 5...

| AVANT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | APRES                                                                                                                                                                                                                                 |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>[{   "id":   "https://fr.wikipedia.org/wiki/Mars_Exploration_Rover",   "value": "Mars Exploration Rover (MER) est une mission double de la NASA lancée en 2003 et composée de deux robots mobiles ayant pour objectif d'étudier la géologie de la planète Mars, en particulier le rôle joué par l'eau dans l'histoire de la planète. Les deux astromobiles ont été lancés au début de l'été 2003 et se sont posés en janvier 2004 sur deux sites martiens susceptibles d'avoir conservé des traces de l'action de l'eau dans leur sol. Chaque astromobile, piloté par un opérateur depuis la Terre, a alors entamé un périple en utilisant une batterie d'instruments embarqués pour analyser les roches les plus intéressantes (...)" }]</pre> | <pre>[  {   "id":   "https://fr.wikipedia.org/wiki/Mars_Exploration_Rover",   "value": [     "deux robots",     "panneaux solaires",     "mars exploration rover mer",     "mission double",     "deux robots mobiles"   ]  } ]</pre> |



# Web services : de la simplicité à la complexité !

37

Quelques exemples de web services :

- Extraction de termes avec Teeft
- **Détection de genre d'un auteur**
- Classification en domaines scientifiques Pascal-Francis

| AVANT                                                                                                                                                                                                  | APRES                                                                                                                                                                                 |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>[<br/>  {"id": "1", "value": "Jean Christophe,<br/>Dupont"},<br/>  {"id": "2", "value": "Amke"},<br/>  {"id": "3", "value": "Seong-Eun Park"},<br/>  {"id": "4", "value": "James A."}<br/>]</pre> | <pre>[<br/>  {"id": "1", "value": "masculin"},<br/>  {"id": "2", "value": "mixte_feminin"},<br/>  {"id": "3", "value": "feminin"},<br/>  {"id": "4", "value": "masculin"}<br/>]</pre> |

## Détection de genre

Ce web service permet de détecter le genre à partir d'une liste de prénoms genrés. Cette liste est un mélange entre les données issues de la librairie python gender-guesser et des données issues de la plateforme Kaggle. Elles ont été...

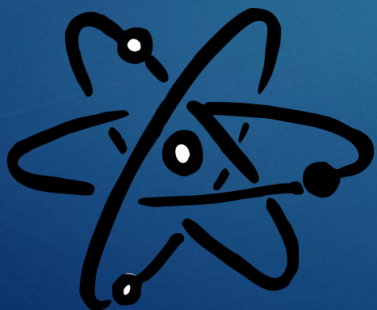


Quelques exemples de web services :

- Extraction de termes avec Teeft
- Détection de genre d'un auteur
- **Classification en domaines scientifiques Pascal-Francis**

## Classification en domaines scientifiques

Le web service de classification automatique permet de classer 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 possédera un domaine scientifique homogène, dans...



| AVANT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | APRES                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>[{"idt": "08-040289", "value": "Planck 2015 results. XIII. Cosmological parameters. We present results based on full-mission Planck observations of temperature and polarization anisotropies of the CMB. These data are consistent with the six-parameter inflationary LCDM cosmology. [...] However the amplitude of the fluctuations is found to be higher than inferred from rich cluster counts and weak gravitational lensing. Apart from these tensions, the base LCDM cosmology provides an excellent description of the Planck CMB observations and many other astrophysical data sets."}]</pre> | <pre>[{"idt": "08-040289", "value": [{"code": {"id": "001", "value": "Sciences exactes et technologie."}, "confidence": 1.0000057220458984, "rang": 1}, {"code": {"id": "001E", "value": "Terre, océan, espace."}, "confidence": 0.9999549388885498, "rang": 2}, {"code": {"id": "001E03", "value": "Astronomie."}, "confidence": 1.0000100135803223, "rang": 3}]}]</pre> |



[← ACCÈS ISTEX.FR](#)

ACCUEILLOTERRERACTUALITÉSCORPUS SPÉCIALISÉS

**ISTEX TDM**  
Les **services Istex** pour la fouille de textes

**Rechercher un web service**

**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

Actuellement **42** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

 **Besoin d'aide ?**

← ACCÈS ISTEX.FR

ACCUEILLOTTERREALITÉS

CORPUS SPÉCIALISÉS

ISTEX TDM

Les services Istex pour la fouille de textes

Accueil > Web-services

Recherche de web-services

genre

RECHERCHER

OBJET TRAITÉ

☐ Auteurs (1)

☐ Résumés (3)

☐ Texte intégral (3)

LANGUES (3)

TRAITEMENT (4)

TYPE DE DONNÉES (1)

genderDetect

Détection du genre de l'auteur

Ce web service retourne le genre d'un auteur ou d'une autrice à partir d'un prénom.

speciesTag

Extraction de noms d'espèces

Ce service web détecte dans un texte les noms scientifiques d'espèces animales, végétales (ainsi que les virus, bactéries, champignons, chromistes, protistes, etc.). Ce service fonctionne quelle que soit la langue à condition qu'elle soit dans un alphabet latin.

textNormalize

Normalisation d'un texte ou d'un terme

Teeft

Extraction de termes d'un texte via Teeft

Detect-Gender - Détection du genre de l'auteur

Description

Utilisation

Niveau d'utilisation : Débutant

Niveau de validation : Expérimental

Objectif

Ce web service retourne le genre d'un auteur ou d'une autrice à partir d'un prénom.

Méthode

Les formats de prénoms pris en compte sont les suivants :  
"prénom"  
"prénom nom"  
"prénom, nom"  
Plusieurs sorties sont possibles :  
- masculin : le prénom est masculin  
- féminin : le prénom est féminin  
- mixte\_masculin : le prénom est mixte mais majoritairement porté par des hommes  
- mixte\_feminin : le prénom est mixte mais majoritairement porté par des femmes  
- mixte : le prénom est mixte  
- unknown : le prénom n'est pas dans nos données ou mal formé (ex: une initiale)  
Notre liste "genre-prénom" est un mélange entre les données issues de la bibliothèque python [gender-quesser](#) et des données issues de la plateforme [Kaggle](#) :  
- Gender-quesser : regroupe plus de 40000 prénoms internationaux avec le genre associé et  
- Kaggle : regroupe les données des prénoms des bébés français et leur genre de 1900 à 2018 (INSEE)  
Ces données ont été fusionnées dans un pré-traitement et enregistrées sous la forme d'un dictionnaire avec les prénoms en clé et les genres en valeurs :  
{ "Jean-Claude": "masculin", "Amke": "mixte\_feminin" }  
Le genre d'un prénom peut être différent selon le pays. Ainsi nous avons fait le choix de sélectionner le genre le plus fréquent dans le monde.



[← ACCÈS ISTEX.FR](#)

ACCUEILLOTTERREALITÉS

ISTEX TDM

Les services Istex pour la fouille de textes

Accueil > Web-services

Recherche de web-services

OBJET TRAITÉ

☐ Auteurs (1)

☐ Résumés (3)

☐ Texte intégral (3)

LANGUES (3)

TRAITEMENT (4)

TYPE DE DONNÉES (1)

genre

RECHERCHER

genderDetect

Détection du genre de l'auteur

Ce web service retourne le genre d'un auteur ou d'une autrice à partir d'un prénom

speciesTag

Extraction de noms d'espèces

Ce service web détecte dans un texte les noms scientifiques d'espèces animales, végétales (ainsi que les virus, bactéries, champignons, chromistes, protistes, etc.). Ce service fonctionne quelle que soit la langue à condition qu'elle soit dans un alphabet latin.

textNormalize

Normalisation d'un texte ou d'un terme

Teeft

Extraction de termes d'un texte via Teeft

Detect-Gender - Détection du genre de l'auteur

Description

Utilisation

URL DU WEB SERVICE À RENSEIGNER DANS LODEX EST :

<https://authors-tools.services.istex.fr/v1/first-name/gender>

Exemple textuel du traitement

Le format d'entrée :

Copier l'exemple

[{"id": "1", "value": "Jean Christophe, Dupont"}, {"id": "2", "value": "Anke"}, {"id": "3", "value": "Seong-Eun Park"}, {"id": "4", "value": "James A."}]

Le résultat :

Copier l'exemple

[{"id": "1", "value": "masculin"}, {"id": "2", "value": "mixte\_feminin"}, {"id": "3", "value": "feminin"}, {"id": "4", "value": "masculin"}]

DÉMONSTRATION

CODE SOURCE

# Les Web-Services

<https://openapi.services.inist.fr/>

- permet de tester simplement un WS en ligne avec son propre exemple.

POST /v1/first-name/gender Renvoie le genre le plus probable d'un prénom

Try it out

| Name                         | Description                                                                                 |
|------------------------------|---------------------------------------------------------------------------------------------|
| path<br>string<br>(query)    | The path in each object to enrich with a Python script<br><input type="text" value="path"/> |
| indent<br>boolean<br>(query) | Indent or not the JSON Result<br><input type="text" value="--"/>                            |

Request body required

Example Value | Schema

```
[
  {
    "id": 1,
    "value": "Jean Christophe"
  },
  {
    "id": 2,
    "value": "Amke"
  },
  {
    "id": 3,
    "value": "Seong-Eun"
  },
  {
    "id": 4,
    "value": "James A."
  }
]
```



**Inist** | Institut de l'information scientifique et technique

Titre ⚙️

Evolution des parts de marché des grands pays industriels

Identifiant de l'article ⚙️

ecop\_0338-4217\_1974\_num\_14\_1\_2020

Auteurs ⚙️

• François Cellier

Collection ⚙️

Economie et prévision

Date ⚙️

1974

## Description bibliographique

Titre ▾

Collection ▾

Date ▾

Langue ▾

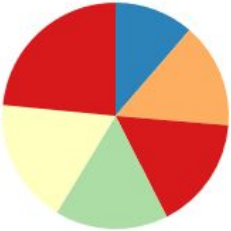
Editeur ▾

Présence de résumé en français ▾

## Filtres/Facettes

Graphe des collections ⚙️

Représentation



Economie appliquée

Economie et prévision

Géocarrefour

Paléorient

Psychologie clinique et projective

Revue d'écologie

PARCOURIR LES RÉSULTATS



## Ressources :

- Lien vers istex TDM : <https://services.istex.fr/>
- Site Lodex : <https://www.lodex.fr/>
- Documentation Lodex :  
<https://www.lodex.fr/docs/documentation/principales-fonctionnalites-disponibles/>