Pseudonymiser des documents grâce à l'IA

Etalab

etalab gouv.fr

11/02/2021

Table des matières

1 Introduction

1.1 À quoi sert ce guide ?

De nombreuses administrations publiques sont confrontées à des problèmes de pseudonymisation dès lors qu'elles ont à publier des **documents textuels contenant des données à caractère personnel**. Ces documents recouvrent par exemple des décisions de justice, des actes administratifs, des procès-verbaux, des notes, etc.

C'est dans ce cadre qu'Etalab a développé un outil d'intelligence artificielle de pseudonymisation pour le Conseil d'État, qui publie en open data des décisions de justice administrative. Cet outil est open-source et peut donc être librement réutilisé pour d'autres projets de pseudonymisation.

Pour accompagner la publication de cet outil technique de pseudonymisation, nous pensons qu'il est nécessaire de publier également un **guide qui expose ce qu'est la pseudonymisation de documents textuels et, lorsque c'est possible, l'utilisation de l'intelligence artificielle (IA) pour la mettre en œuvre**.

1.2 À qui s'adresse ce guide ?

Ce guide s'adresse principalement **aux organismes publics**, et plus particulièrement **aux personnes chargées du traitement et de la protection de données à caractère personnel** dans ces organismes. Ces personnes peuvent être des agents publics, internes à l'administration, mais aussi des prestataires. Dans ce dernier cas, le sous-traitant devra veiller au respect des obligations relatives à la sous-traitance imposées par le RGPD (voir le guide de la CNIL sur ce sujet).

Ce guide pourra également intéresser d'autres acteurs faisant face à un besoin de pseudonymisation de documents textuels, dans le cadre de développements de services ou de produits à partir de données à caractère personnel.

1.3 Sommaire

Ce guide est composé de trois parties et d'un lexique :

- La première partie permet de **découvrir ce qu'est la pseudonymisation, pourquoi elle est utile dans les administrations publiques** et présente les méthodes de pseudonymisation existantes.
- La deuxième partie expose une vue d'ensemble de la **méthode basée sur l'IA** que nous avons développée à Etalab.

— La troisième partie s'adresse à un public plus technique, comme des data scientists, et **présente de manière plus détaillée la mise en œuvre de cette approche** basée sur l'IA.
— Le lexique proposé en annexe vous permet finalement de retrouver une définition des termes techniques mentionnés tout au long du guide.

> **Ce que ce guide n'est pas**
>
> — un guide juridique sur la protection des données à caractère personnel ;
> — un guide sur la pseudonymisation de données autres que textuelles (en particulier tabulaires) ;
> — un guide sur la sécurité des données et des systèmes d'information.

1.4 Comment contribuer ?

Ce document est un outil évolutif et ouvert. Vous pouvez contribuer à l'améliorer en proposant une modification sur GitHub ou en contactant directement l'équipe du Lab IA d'Etalab.

2 Pourquoi et comment pseudonymiser dans l'administration

2.1 Qu'est-ce que la pseudonymisation ?

2.1.1 Quelle différence entre anonymisation et pseudonymisation ?

Anonymisation et pseudonymisation sont deux notions parfois difficile à distinguer, et qui concernent toutes deux la **protection des données à caractère personnel**.

> **Lexique : Donnée à caractère personnel**
>
> Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un nom, un numéro d'identification (par exemple le numéro de sécurité sociale) ou à un ou plusieurs éléments qui lui sont propres.

Le guide de la CNIL sur l'anonymisation des données présente bien la différence entre anonymisation et pseudonymisation :

> **Lexique : Anonymisation**
>
> « **L'anonymisation** est un traitement qui consiste à utiliser un ensemble de techniques de manière

> à rendre **impossible, en pratique, toute identification de la personne** par quelque moyen que ce soit et ce de **manière irréversible**. »

> **Lexique : Pseudonymisation**
>
> « **La pseudonymisation** est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires. En pratique **la pseudonymisation consiste à remplacer les données directement identifiantes** (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro dans un classement, etc.). […] En pratique, il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces. C'est pourquoi des données pseudonymisées demeurent des données personnelles. **L'opération de pseudonymisation est réversible, contrairement à l'anonymisation**. »

La différence entre anonymisation et pseudonymisation réside ainsi dans le **caractère réversible ou non** de la dissimulation des données à caractère personnel.

> **Un exemple de différence entre pseudonymisation et anonymisation**
>
> Supposons qu'une caisse d'allocations familiales (CAF) dispose d'une base de données contenant les noms, dates de naissance et adresses des demandeurs d'allocation logement en 2019, ainsi que les montants des allocations reçues et le nombre de personnes dans le foyer.
>
> Si la CAF souhaite **anonymiser** ces données, elle pourra supprimer les informations potentiellement identifiantes comme les noms, dates de naissances et adresses, puis aggréger les montants des allocations en ne publiant par exemple que la moyenne par commune. Impossible d'identifier qui se cache derrière les allocations reçues, ce qui garantit la protection totale des données personnelles. Mais impossible aussi de comparer les bénéficiaires des années 2018 ou 2020 avec ceux de 2019, puisque l'on ne dispose pas des données à l'échelle individuelle.
>
> Si elle souhaite **pseudonymiser** ces données, elle remplacera les noms et dates par un identifiant unique (au lieu de supprimer les colonnes) et remplacera les adresses complètes par les seules communes. On peut cette fois-ci comparer les identifiants entre bases pour retrouver les allocataires communs, sans pour autant être en mesure de connaître directement leur identité. Cependant, pour les communes avec un faible nombre d'habitants, les informations sur la composition du foyer pourraient être suffisantes pour réidentifier certains bénéficiaires et ainsi connaître le montant qu'ils perçoivent.

Ainsi, si l'anonymisation seule garantit une totale protection des données à caractère personnel, elle implique parfois une importante perte d'information, nécessaire à empêcher la réidentification mais limitant les réutilisations possibles des données. La pseudonymisation est donc une alternative attractive, à condition de garantir une protection suffisante.

2.1.2 Pourquoi pseudonymiser des documents administratifs ?

La loi n°2016-1321 pour une République numérique fait de **l'ouverture des données publiques la règle par défaut**. Etalab propose par ailleurs un guide détaillé sur l'ouverture de ces données.

Lorsque les administrations diffusent dans ce cadre des documents contenant des données personnelles, **l'occultation préalable des éléments à caractère personnel est généralement une obligation** qui s'impose à elles en application de l'article L. 312-1-2 du Code des relations entre le public et l'administration, sauf dans certains cas particuliers.

2.1.3 Quelles données personnelles dois-je retirer de mes données ?

Cela dépend du contexte réglementaire, le même cadre ne s'appliquant pas à tous les documents. Néanmoins, il conviendra la plupart du temps de **pseudonymiser toute information se rapportant à une personne physique identifiée ou identifiable**. Une « personne physique identifiable » est une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale.

Pour plus de détails sur les **différents cadres légaux**, vous pouvez consulter le guide juridique de la publication des données publiques élaboré par la CNIL et la CADA.

Pour satisfaire à l'obligation d'occultation, **la CNIL préconise d'anonymiser** les documents administratifs avant de les diffuser, garantissant ainsi une parfaite impossibilité de réidentification. Néanmoins, pour les documents qui contiennent des données non structurées, en particulier du texte libre, le curseur de la « quantité d'information » à retirer d'un jeu de données pour éviter tout risque de réidentification est difficile à évaluer. De fait, **une complète anonymisation est difficile à atteindre et à évaluer** et peut aboutir à une trop grande perte d'informations.

Un bon exemple de document administratif pseudonymisé sont les décisions de justice, diffusées notamment sur le site Légifrance. Y sont retirés notamment les noms, prénoms, adresses, dates civiles (naissance, décès, mariage) des personnes physiques. D'autres catégories d'informations, comme les noms d'entreprises, la description de faits (dates et montants d'une transaction par exemple) pourraient permettre, en les recoupant avec d'autres informations, de réidentifier une personne physique. Cependant, retirer trop de catégories d'informations reviendrait à perdre beaucoup d'informations et appauvrirait le contenu d'une décision.

Par acte d'huissier de justice du 28 janvier 2009, Mme H... Y... a notifié à M. W... P... et Mme G... E... B..., épouse P... (les époux P...), titulaires d'un bail à usage professionnel sous seing privé du 21 juin 1995, portant sur un appartement de sept pièces au 3e étage d'un immeuble sis [...], son intention de vendre les locaux qu'ils occupaient, au prix de en vertu de l'article 10 I. de la loi no 75-1351 du 31 décembre 1975, s'agissant de la première vente consécutive à la division initiale de l'immeuble par lots. Par lettre reçue le 2 a fait savoir qu'elle se portait "acquéreur aux prix et conditions notifiés, sous la seule réserve de l'obtention d'un prêt". Par une lettre non datée adressée à Mme P..., Mme Y... a

Un extrait de décision de justice pseudonymisée

Il y a donc un arbitrage à faire entre la minimisation du risque de réidentification et la préservation de l'utilité des données. Trouver le bon curseur n'est pas simple et doit passer par une double analyse des risques de réidentification, à la fois **juridique** (pour évaluer par exemple quelles données ne doivent pas pouvoir être réidentifiées) et **technique** (pour estimer la possibilité technique de réidentifier ces données). Juger de l'utilité de conserver ou non certaines catégories de données **dépendra aussi des usages envisagés** de ces données.

> **Quelle quantité de données retirer ? Un exemple fictif**
>
> Prenons l'exemple d'un extrait de décision de justice fictive : *« Monsieur Dupont est accusé d'avoir cambriolé l'établissement "Café de la Paix" à Gentioux-Pigerolles, en Creuse, situé en face de son domicile, et d'avoir dérobé la recette de la semaine évaluée à 1 000€ »*.
>
> — **Cas 1** : on conserve le plus d'information possible, en supprimant néanmoins les noms des personnes physiques et morales. La pseudonymisation sera par exemple : *« Monsieur X. est accusé d'avoir cambriolé l'établissement "Café E." à Gentioux-Pigerolles, en Creuse, situé en face de son domicile, et d'avoir dérobé la recette de la semaine évaluée à 1 000€ »*. Le problème, c'est que s'il n'y a qu'un seul café dans ce petit village, il est très aisé de comprendre de quel établissement on parle, de sa localisation et donc celle du domicile de l'accusé, et ainsi de réidentifier ce dernier si l'on est familier du village. La pseudonymisation est donc inutile et ne protège pas suffisamment les données à caractère personnel.
> — **Cas 2** : on conserve le moins d'information possible. On pourra alors obtenir la pseudonymisation suivante : *« Monsieur X. est accusé d'avoir cambriolé l'établissement "E." à Y., en Z., situé en face de son domicile, et d'avoir dérobé la recette de la semaine évaluée à N€ »*. Le problème c'est qu'il n'y a là plus beaucoup d'information utile. Par exemple, comment réaliser une cartographie du crime sans localisation ? Comment estimer les préjudices moyens des cambriolages pour un assureur ?

Un rapport du groupe de travail du G29 sur la protection des personnes à l'égard du traitement des données à caractère personnel présente une analyse détaillée des risques de réidentification après pseudonymisation, d'un point de vue juridique et technique, et des bonnes pratiques en fonction des types de données.

2.2 Quelles sont les différentes méthodes de pseudonymisation ?

2.2.1 Dans le cas où les données à caractère personnel sont tabulaires

Lorsque les données à caractère personnel sont contenues dans un jeu de données tabulaire (c'est-à-dire, pour faire simple, sous forme d'un tableau dont les lignes sont des entrées et les colonnes des

catégories d'information), il est aisé de procéder directement à des traitements visant à pseudonymiser ou anonymiser, en **supprimant les colonnes concernées ou en chiffrant leur contenu**. Ce cas de figure n'est pas l'objet de ce guide. Pour plus d'informations à ce sujet, on se référera aux ressources de la CNIL sur l'anonymisation.

2.2.2 Dans le cas où les données à caractère personnel apparaissent dans du texte libre

Lorsque les données à caractère personnel sont contenues dans du texte libre, le ciblage précis des éléments identifiants dans le texte est une tâche souvent complexe. Encore largement réalisée par des humains, **cette tâche est coûteuse en temps et peut requérir une expertise spécifique dans la matière traitée** (dans les textes juridiques par exemple). L'intelligence artificielle et les techniques de traitement du langage naturel peuvent permettre d'automatiser cette tâche souvent longue et fastidieuse.

Une méthode d'automatisation simple : les moteurs de règles Il existe des méthodes assez intuitives permettant d'automatiser la tâche de pseudonymisation, comme **les moteurs de règles**. Les moteurs de règles sont un ensemble de règles prédéfinies « à l'avance ». Par exemple, une règle de pseudonymisation pourrait être : « si le mot qui suit "Monsieur" ou "Madame" commence par une majuscule, alors ce mot est un prénom ». La complexité du langage naturel et la diversité des formulations qui se trouvent dans du texte libre fait que ce type de moteur de règles a de forte chance de faire beaucoup d'erreurs dans des textes administratifs dont la forme varie souvent. Il est cependant bien adapté à des textes simples, ou lorsque la méthode a besoin d'être parfaitement explicable et simplement modifiable.

Une méthode plus complexe : l'intelligence artificielle (IA) L'utilisation de l'IA pour automatiser la pseudonymisation de documents peut être plus ou moins pertinente. Les solutions d'IA pour pseudonymiser des données textuelles sont en grande majorité des modèles supervisés. **Ces modèles d'IA dits d'apprentissage supervisé se sont beaucoup développés ces dernières années**, en particulier les modèles de réseaux de neurones profonds (ou *deep learning*) qui sont aujourd'hui les plus performants.

Ces modèles supervisés sont des algorithmes qui prennent en entrée des données avec des *labels* (ou étiquettes en français), dont ils vont chercher à « apprendre » la logique d'attribution. L'objectif est ainsi que lorsqu'on leur présente une nouvelle donnée « non labelisée », l'algorithme soit capable de retrouver seul le bon label.

Dans le cas de la pseudonymisation, les données d'entrées sont **chacun des mots du document à pseudonymiser** et le label qu'on leur attribue est la **catégorie sémantique** à laquelle il se rattache :

nom, prénom, adresse, etc. Ces catégories varient selon la nature du document et le degré de pseudonymisation souhaité. En traitement du langage naturel, ce type de tâche s'appelle la **reconnaissance d'entités nommées** (*Named Entity Recognition (NER)* en anglais).

Mais pour que ces modèles puissent arriver à de bonnes performances, ils nécessitent de remplir un certain nombre de prérequis. Assez exigeants, ils sont pourtant indispensables au succès de l'utilisation de l'IA appliquée à la pseudonymisation. Nous vous proposons de les passer en revue dans la section suivante.

2.3 Quels sont les prérequis pour utiliser l'intelligence artificielle pour pseudonymiser ?

2.3.1 Disposer de données brutes de qualité

La qualité des données brutes (c'est-à-dire avant tout traitement) que l'on souhaite utiliser est un premier critère essentiel qui sera déterminant pour la performance de l'algorithme. Cette qualité fait souvent référence à **la facilité d'utilisation du format utilisé**. En effet, les données textuelles brutes peuvent se présenter sous différents formats, plus ou moins lisibles. Idéalement, les documents textuels sont stockés au **format *txt* ou *json***. Des formats moins standards (*doc*, *pdf*, *png*, etc..) nécessiteront des conversions afin de pouvoir être traités. Lorsque les documents sont au format image (car résultant par exemple d'une numérisation de documents papiers), la mise en place d'une brique de reconnaissance optique de caractères sera nécessaire afin de les convertir au format texte, et complexifie donc le traitement en amont du projet. La qualité des données brutes est évaluée par les data scientists en amont du projet.

2.3.2 Disposer d'un grand volume des données

Essentiel également, le volume de documents annotés nécessaires dépendra de la complexité de la tâche de pseudonymisation, qui sera fonction notamment du nombre de catégories d'entités nommées retenues et de la complexité du langage utilisé dans les documents. Il est en général nécessaire de disposer de l'ordre de **plusieurs milliers de documents afin d'obtenir des résultats optimaux**.

2.3.3 Avoir la possibilité d'annoter ses données

Puisque la tâche de notre IA consiste à reconnaître la catégorie sémantique de chaque mot, il est nécessaire en amont de tout projet de **disposer « d'exemples » que l'on souhaite montrer à l'algorithme pour qu'il s'entraîne**. Il sera donc nécessaire de constituer au préalable, à la main (humaine), une base d'exemples corrects. **Cette tâche consistant à attribuer des labels à certains mots ou groupes**

de mots d'un document s'appelle l'annotation. Cette tâche pourra nécessiter des compétences spécifiques en fonction de la nature des documents et des catégories à annoter.

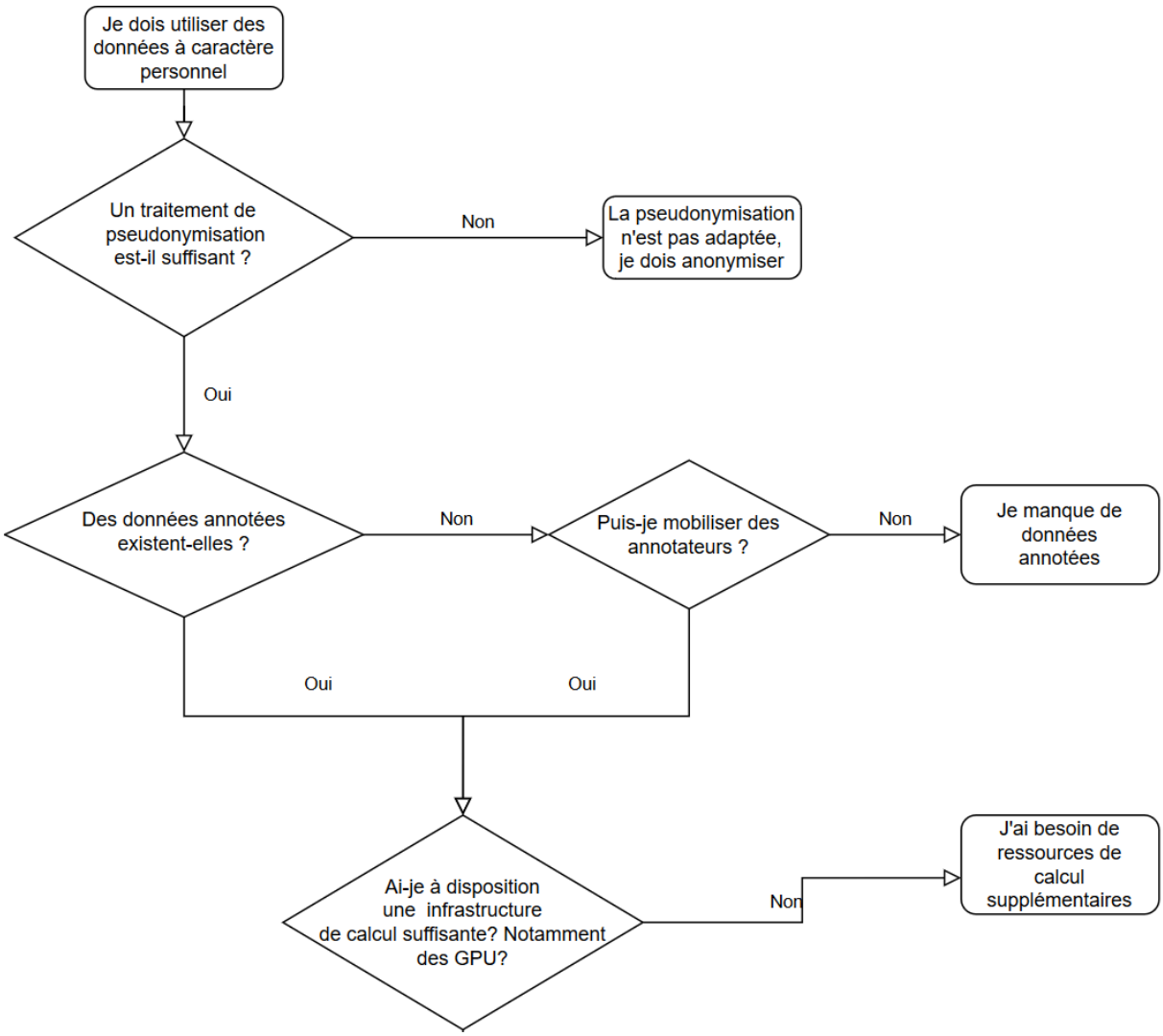
> **L'annotation, un processus exigeant et chronophage**
>
> Le processus d'annotation requiert de mobiliser des équipes souvent nombreuses (pour aller plus vite) mais aussi qualifiées. Par exemple, si vous cherchez à identifier les noms, prénoms et adresses dans un recours administratif, une simple maîtrise du français suffira. En revanche, si vous cherchez à identifier les moyens et les conclusions juridiques mentionnées, il vous faudra disposer d'une équipe de juristes expérimentés ! Pour des documents complexes, il pourra ainsi être nécessaire de mobiliser longuement des experts métiers pour obtenir une quantité d'annotation suffisante et de qualité (avec le moins de mauvais labels). On parle même de campagnes d'annotation !

Afin de constituer un ensemble de documents annotés, il est nécessaire d'utiliser un logiciel d'annotation qui permet d'enregistrer les annotations réalisées par les annotateurs. Il existe de nombreux logiciels d'annotation, dont beaucoup sont open source comme Doccano.

2.3.4 Avoir accès à des infrastructures de calcul adéquates

L'entraînement de modèles de traitement du langage récents, basés sur des réseaux de neurones profonds (deep learning), **nécessite des ressources dédiées et exigeantes**. D'une part, la volumétrie de données nécessaires pour l'entraînement peut mener à la constitution de corpus de plusieurs giga voire téraoctets et peut nécessiter des infrastructures de stockages dédiées, comme des **serveurs de stockage**. D'autre part, l'entraînement des modèles est pour sa part très gourmand en capacités de calcul, et s'appuie notamment des **processeurs graphiques** (*GPU* en anglais) qui permettent d'accélérer considérablement le temps de calcul. Même en disposant de GPU de dernières générations, il faut compter plusieurs jours voire plusieurs semaines pour un apprentissage complet du modèle.

En résumé, de nombreuses conditions doivent être remplies avant de se lancer dans un projet d'utilisation de l'IA pour la pseudonymisation. Voici résumé le questionnement logique à suivre :



2.4 Ressources externes

- Guide de l'anonymisation pour les données ouvertes de la CNIL
- Guide pratique (juridique) de la publication en ligne et de la réutilisation des données publiques élaboré par la CNIL, la CADA et Etalab
- Guide RGPD du développeur de la CNIL
- Avis sur les techniques d'anonymisation du Groupe de travail du G29 sur la protection des personnes à l'égard du traitement des données à caractère personnel
- Guide « Pseudonymisation techniques and best practices » de l'Agence Européenne de Cybersécurité (ENISA)

3 Les étapes d'un projet de pseudonymisation grâce à l'IA

Nous proposons dans cette partie de passer en revue les différentes étapes d'un projet de pseudonymisation de documents textuels. À ce stade, nous faisons l'hypothèse que nous disposons déjà de données nettoyées, prêtes à l'emploi. Le pré-traitement des données brutes évoqué plus tôt (OCR, conversion en format standard, etc.) n'est pas détaillé ici.

3.1 Annoter ses données

Nous l'avons vu plus tôt, le premier prérequis pour mener une pseudonymisation automatique de données textuelles par l'IA est de **disposer de données annotées**, qui serviront « d'exemples » de pseudonymisation correcte à l'algorithme. Si vous ne disposez pas de telles données en l'état, il sera nécessaire que vous les annotiez à la main. **C'est un processus assez long**, surtout si l'on cherche à disposer de plusieurs milliers de documents ! Il faudra donc rassembler deux éléments :

— un logiciel d'annotation, comme nous l'avons déjà vu, comme les outils open source Doccano, WebAnno ou Universal Data Tool ;
— une équipe d'annotateurs dévoués et patients, qui disposent d'une expertise métier adéquate si vous travaillez sur des documents spécifiques, comme par exemple des décisions de justice.

Gardez à l'esprit que l'annotation est un processus chronophage et peut **nécessiter plusieurs mois de travail pour des documents complexes**. Ceux-ci sont en effet long et fastidieux à lire puis à annoter et il sera donc impossible de procéder d'une traite. Une solution alternative est de procéder **par séances courtes mais fréquentes, et rassemblant de nombreux annotateurs**. L'évaluation précise du temps d'annotation nécessaire ne peut se faire qu'**après quelques séances d'annotation**, afin de pouvoir évaluer la vitesse moyenne d'annotation d'un document.

L'annotation se devra d'être de qualité pour garantir de bonnes performances de l'algorithme. **Une annotation de mauvaise qualité se caractérise par des omissions fréquentes d'entités nommées, ou l'attribution de la mauvaise catégorie d'entité à des mots**, etc. La phase d'entraînement d'un algorithme d'IA consiste en effet à « mimer » la labélisation qu'on lui présente, et de mauvaises annotations vont logiquement conduire l'algorithme à mal prédire les catégories sémantiques dans de nouveaux documents. De plus, une autre partie des données annotées va servir à évaluer la performance de l'algorithme, en comparant les labels prédits par l'algorithme à ceux déterminés « manuellement ». **Si les labels issus de l'annotation par des humains ne sont pas fiables, l'évaluation de la performance de l'algorithme ne sera pas non plus fiable.** La qualité des annotations doit donc être vérifiée par des experts métiers tout au long de la campagne d'annotation.

3.2 Organiser ses données

Dans le cas de la pseudonymisation, les données sont constituées de l'ensemble des documents (texte libre) desquels il faut occulter des éléments identifiants. Dans un projet d'apprentissage supervisé, on distingue deux grands ensembles de données : les **données annotées**, que l'on consacre à la phase d'apprentissage du modèle, et les **données à labéliser** (ou non annotées) sur lesquelles on appliquera le modèle une fois celui-ci entraîné.

Les données annotées sont elles-mêmes séparées par la suite en deux groupes, souvent aléatoirement, entre **données d'entraînement** et **données de test**. Les données d'entraînement vont permettre à l'algorithme **d'apprendre à reproduire la tâche de reconnaissance des labels de chacun des mots du texte**. Afin d'évaluer la performance de l'algorithme, il est nécessaire d'utiliser d'autres données annotées, que l'algorithme n'a pas « vu » en entraînement. C'est le rôle du jeu de données de test.

3.3 Formater ses données

Certains arbitrages doivent être effectués à cette étape, afin de **choisir quels prétraitements vont être appliqués au texte**. Par exemple, dois-je transformer toutes les majuscules en minuscules ? Dois-je conserver la ponctuation ? Et quid des « stop words », ces mots peu porteurs de sens comme « et », « ou », « mais » ? Le but de ces prétraitements est de supprimer l'information inutile, mais d'en conserver assez pour que l'algorithme atteigne son but.

Une fois le texte prétraité, il est **transformé grâce à un modèle de langage**, qui permet, pour faire simple, d'obtenir pour chaque mot une représentation sous forme vectorielle. C'est en effet ce type d'input qu'utilisent les algorithmes d'apprentissage automatique, et de nombreux modèles de langages peuvent être utilisés en fonction de la tâche ou de la langue, comme par exemple CamemBERT pour le français.

3.4 Entraîner son modèle

Une fois les données formatées et mises sous forme de vecteurs, elles peuvent être utilisées pour entraîner l'algorithme dont la tâche sera de reconnaître le label de chacun des mots du texte. Là encore, **de nombreux arbitrages sont possibles pour choisir l'architecture à utiliser**. Les solutions les plus performantes aujourd'hui s'appuient sur des réseaux de neurones profonds. L'un des modèles le plus utilisé pour la tâche qui nous intéresse porte par exemple le nom un peu barbare de Bidirectional Long Short Term Memory - Conditional Random Fields (ou BiLSTM-CRF). Une fois l'architecture définie, **l'apprentissage consiste à « donner à voir » à l'algorithme les données d'entraînement**, souvent de nombreuses fois d'affilée, afin que celui-ci ajuste ses paramètres pour effectuer au mieux la tâche de reconnaissance du label correspondant à chaque mot.

3.5 Valider ses résultats

La validation des performances du modèle est un double processus qui repose à la fois sur l'analyse de métriques statistiques et sur l'expérience humaine. Les métriques statistiques, souvent communes à tous les modèles et indifférentes au type de projet concerné, **permettront d'obtenir un résumé synthétique des performances générales de l'algorithme** et d'apprécier par exemple ses progrès au cours du projet. Cependant, **l'appréciation humaine sera indispensable** pour vérifier dans le détail que les résultats sont réellement satisfaisants sur la tâche particulière que l'on cherche à automatiser. Par exemple, pour la pseudonymisation des décisions de justice, une métrique « métier » consistera à calculer le ratio « nombre de décisions avec au moins une erreur » sur « nombre de décisions analysées ».

D'autres critères **plus « matériels » peuvent entrer en jeu** dans la validation du modèle. Par exemple pour la pseudonymisation, combien de temps faut-il pour pseudonymiser entièrement un document ? Combien de temps faut-il pour réentraîner le modèle avec de nouvelles données annotées ? Quelles ressources de calcul et de stockage cela nécessite-t-il ? Si les réponses à ces questions semblent à chaque fois trop élevées, il est peut-être nécessaire de **questionner les choix de modèles réaliser et de chercher un meilleur compromis** entre performance et exigences matérielles.

3.6 Pseudonymiser de nouveaux documents

Une fois que vous estimez les résultats de votre algorithme convaincants, le tour est joué ! Vous pouvez maintenant lui présenter de nouveaux documents, **que vous n'avez pas annotés**. Si votre algorithme est bien entraîné, il sera capable de reconnaître seul le label de chaque mot. Ainsi, si vous **ajoutez la règle simple de remplacer par un alias tous les mots dont le label est une donnée à caractère personnel** (par exemple les noms par X, Y ou Z, les prénoms par A, B ou C, etc.), vous obtenez un outil de pseudonymisation automatique !

Cette méthode de pseudonymisation par l'IA n'est évidemment jamais exempte d'erreurs. Il est donc nécessaire de prévoir, avant toute ouverture au grand public, une étape de vérification voire de reprise manuelle pour s'assurer que l'ensemble des documents est bien pseudonymisé. Il n'est pas forcément nécessaire de relire l'intégralité des documents, mais plutôt de convenir d'un protocole de contrôle à appliquer avant toute publication. Celui-ci peut par exemple prendre la forme de l'échantillonage d'un sous-ensemble représentatif des documents à publier, sur lequel la relecture se limitera. Ce protocole permettra en particulier de s'assurer que la qualité de la pseudonymisation automatique ne décroit pas au cours du temps.

4 La pseudonymisation par l'IA en pratique

Après avoir vu dans les grandes lignes les étapes d'un projet de pseudonymisation grâce à l'IA, nous revenons plus en détails dans cette partie sur ces différentes étapes pour présenter les choix, arbitrages et préconisations techniques que nous avons tirés de nos travaux pour la création d'un moteur de pseudonymisation pour les décisions du Conseil d'État. Ceux-ci sont disponibles sur GitHub.

4.1 Formater ses données annotées

Afin de pouvoir utiliser les données annotées pour l'entraînement d'un algorithme d'apprentissage, **celles-ci doivent être converties dans un format spécifique**. Dans l'exemple ci-dessous, un document textuel (ici « Thomas CLAVIER aime beaucoup Paris. ») est alors structuré en un tableau, avec un mot par ligne, et deux colonnes, une pour le mot (ou *token*) et une pour l'annotation linguistique. Ce type de format s'appelle **CoNLL**.

Token	Label
Thomas	B-PER
CLAVIER	I-PER
aime	O
beaucoup	O
Paris	B-LOC

Plus particulièrement, nous utilisons le format IOB2, très commun pour les tâches d'apprentissage séquentiel comme la reconnaissance d'entités nommées, pour labéliser nos données. Ce format permet d'aider l'algorithme d'apprentissage à mieux repérer les entités. Le préfixe B- avant un label indique que le label est le début d'un groupe de mots, le préfixe I- indique que le label est à l'intérieur d'un groupe de mots, et le label O indique que le token n'a pas de label particulier. Il existe d'autres formats similaires à IOB2, tels que IOB/BIO, BILOU, et BIOES.

> **Le format CoNLL**
>
> CoNLL, pour « Conference on Natural Language Learning », est un format général, dont il existe de nombreuses versions, couramment employé pour les tâches de traitement du langage naturel, décrivant des données textuelles en colonne selon un nombre d'attributs (catégorie d'entité

> nommée, nature grammaticale, etc.). Le format IOB2 que nous utilisons est l'une des méthodes de labélisation du format CoNLL.

Il existe de très **nombreux logiciels ou solutions d'annotation de données textuelles** et les données annotées en sortie peuvent donc avoir différents formats (il existe en effet de multiples formats de données annotées). Pour transformer vos données annotées, un développement spécifique sera probablement nécessaire afin de les convertir au format IOB2, le format des données d'entrée de l'algorithme de reconnaissance d'entités nommées que nous avons choisi. Plusieurs exemples de fonctions et de librairies développées pour le Conseil d'État constitueront néanmoins un point de départ dans le répertoire GitHub de notre projet.

4.2 Tokeniser le texte

Afin de mettre nos données sous format CoNLL, nous avons besoin d'abord d'identifier les mots individuels dans nos documents. Si l'on considère un document, composé de blocs de caractères, **la tokenisation est la tâche qui consiste à découper ce document en éléments atomiques**, en gardant ou supprimant la ponctuation. Par exemple :

Phrase
Mes amis, mes enfants, l'avènement de la pseudonymisation automatique est proche.

La phrase ci-dessus pourrait être tokenisée de cette manière :

Token1	Token2	Token3	Token4	Token5	Token6	Token7	Token8	Token9	Token10	Token11	Token12	Token13	Token14	Token15	Token16
Mes	amis	,	mes	enfants	,	l	'	avènement	de	la	pseudonymisation	automatique	est	proche	.

Les tokens correspondent généralement aux mots, mais il est important de comprendre qu'une autre unité peut être choisie, par exemple les caractères. D'autres choix dans la façon de tokeniser peuvent avoir un impact sur les résultats de l'algorithme. Par exemple, le choix de conserver ou non la ponctuation a son importance. De manière pratique, il est important de bien comprendre la méthode de tokenisation utilisée par les algorithmes, afin de prendre en compte ces choix lors de l'étape finale d'occultation d'éléments identifiants dans le texte.

4.3 Entraîner son modèle

Dans le code que nous avons développé, nous utilisons la librairie Open Source Flair. Celle-ci permet en effet d'utiliser de nombreux modèles de langage, par exemple les modèles Flair, Bert et CamemBERT et même de combiner plusieurs de ces modèles. **Un modèle de langage permet pour chaque mot d'obtenir une représentation vectorielle** (ou *embedding*). Ces embeddings sont ensuite passés à un classificateur BiLSTM-CRF qui attribue à chaque mot une des classes du jeu de données d'entraînement.

L'entraînement d'un tel classificateur nécessite de choisir la valeur d'un certain nombre d'**hyper-paramètres**. Les hyper-paramètres sont les paramètres de l'algorithmes qui sont fixés avant l'apprentissage, par opposition aux paramètres de l'algorithmes qui sont fixés de manière itérative au cours de l'apprentissage. Des exemples de configuration avec des explications des différents hyper-paramètres et de leur impact sont disponibles dans la section correspondante du répertoire GitHub.

Nous proposons un exemple de module permettant d'entraîner un algorithme de reconnaissance d'entités nommées via la librairie Flair à partir d'un corpus annoté. Enfin, pour aller plus loin, la librairie Flair propose un module très pratique permettant de fixer les valeurs optimales des hyper-paramètres optimaux pour l'apprentissage.

4.4 Valider ses résultats

La validation des performances du modèle et la mise en production est un double processus reposant sur **l'analyse de métriques et sur l'expérience humaine**, comme nous l'avons vu dans la partie précédente. Comme illustré ci-dessous dans le cas de notre outil de pseudonymisation des décisions de justice, cette validation des résultats est charnière pour décider ou non du passage en production, c'est-à-dire du déploiement pour utilisation par les métiers.

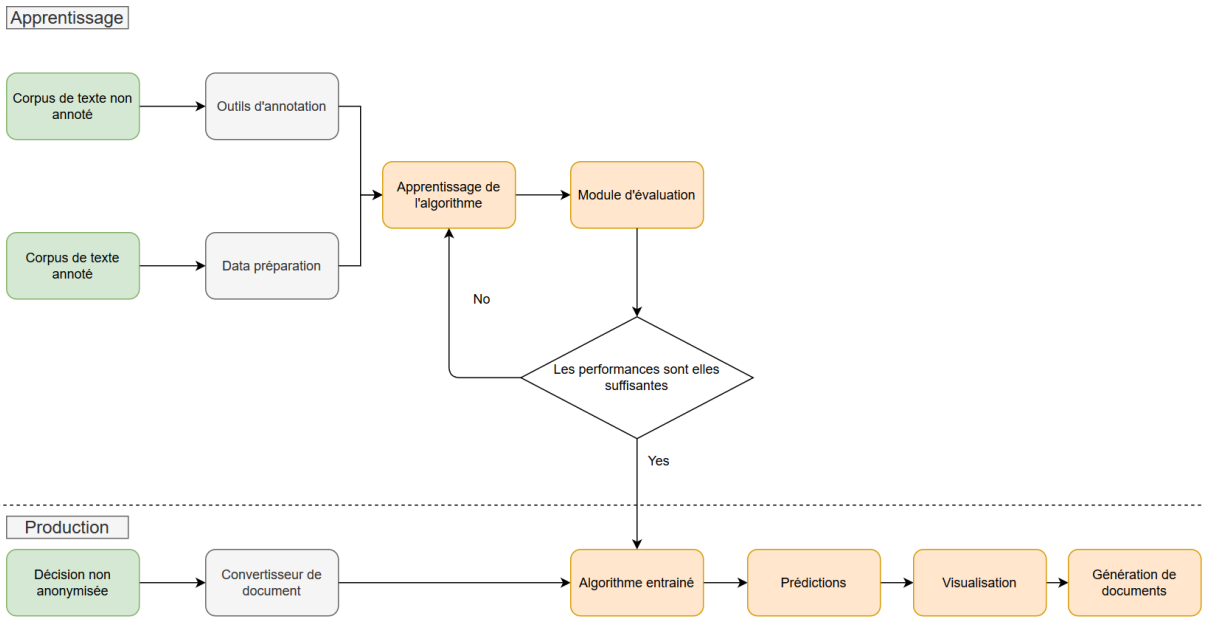


FIGURE 1 – Le processus de conception de notre outil de pseudonymisation

Pour permettre cette double analyse métriques/métiers, notre module de génération de documents pseudonymisés permet de produire en sortie des fichiers labélisés, au format CoNLL, et ainsi :

— d'utiliser des **métriques permettant de comparer**, pour un document ayant été annoté manuellement, la pseudonymisation par l'algorithme à celle réalisée manuellement. On utilise généralement le score F1 pour mesurer la performance du modèle ;
— de charger dans notre outil d'annotation basé sur Doccano un fichier **mettant en avant les différences entre les annotations provenant de sources différentes**, indiquant en rouge les labels en désaccord et en vert les labels en accord.

4.5 Pseudonymiser de nouveaux documents

Le modèle entraîné permet d'attribuer une catégorie à chaque token du document à pseudonymiser. **Les sorties de l'algorithme de reconnaissance d'entités nommées ne permettent donc pas d'obtenir directement le document peudonymisé**, mais est nécessaire d'ajouter une brique pour **remplacer les mots identifiés comme des données à caractère personnel par un alias**. Pour le bon fonctionnement de cette étape, il est très important de fournir à l'algorithme un document tokenisé selon une méthode identique à celle utilisée pour entraîner l'algorithme.

4.6 Quelles ressources disponibles pour pseudonymiser ?

4.6.1 Les librairies

De nombreuses librairies open-source permettent d'entraîner et d'utiliser des algorithmes de reconnaissance d'entités nommées. Parmi celles-ci, Flair et SpaCy présentent l'avantage de proposer des algorithmes à l'état de l'art tout en facilitant l'expérience utilisateur.

— Flair est un framework simple pour le NLP. Il permet d'utiliser des modèles de NLP à l'état de l'art sur des textes de tout genre, en particulier des algorithmes de reconnaissance d'entité nommées et des embeddings pré-entraînés
— SpaCy est un module Python à forte capacité d'industrialisation pour le NLP rédigée en Python et Cython. Il implémente les toutes dernières recherches dans le domaine du traitement du langage naturel et a été conçu pour être utilisé en production. Il possède des modèles statistiques et des embeddings pré-entraînés.
— Stanza est une librairie Python de l'Université de Stanford qui utilise la très connue librairie CoreNLP comme moteur NLP. Ses composants permettent un entraînement et une évaluation efficace avec vos propres données annotées. La boîte à outils est conçue pour être parallèle entre plus de 70 langues, en utilisant le formalisme des dépendances universelles.

Si SpaCy est la librairie la plus rapide, Flair est celle que nous avons choisie pour le développement de notre outil de pseudonymisation, et ce pour la performance de son algorithme de reconnaissance d'entités.

4.6.2 Outils d'annotation

Comme évoqué dans la partie précédente, il existe de nombreuses interfaces d'annotation, notamment en open source comme Doccano, WebAnno et Universal Data Tool. Ces outils fournissent des fonctionnalités d'annotation pour la classification de texte, la labélisation de mots et d'autres tâches classiques de traitement du langage naturel. Il est ainsi possible de créer rapidement des données d'entraînement pour l'analyse des sentiments, la reconnaissance d'entités nommées, la synthèse de texte, etc.

4.7 Voir la pseudonymisation en action

Vous pouvez explorer notre démo pseudonymisation en ligne et retrouver le code de cet outil sur notre dépôt Git.

5 Lexique des termes techniques

Dans cette annexe, nous vous proposons de retrouver les définitions des termes techniques évoqués dans ce guide.

5.0.1 Annotation

L'annotation est la tâche manuelle (et donc humaine) qui consiste à attribuer à chaque donnée le label qui lui correspond. Par exemple, à attribuer le label « chien » ou « chat » à une base de photographies d'animaux. Ou encore à attribuer le label correct entre « nom », « prénom », « adresse », « date » ou « aucun » à chacun des mots d'un document. On conçoit ainsi une base de données annotée, utile pour l'apprentissage supervisé.

5.0.2 Anonymisation

Le guide de la CNIL sur l'anonymisation des données définit l'anonymisation comme « un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et ce de manière irréversible ».

5.0.3 Apprentissage supervisé

L'apprentissage supervisé est une catégorie de tâches de l'apprentissage automatique, principal domaine de l'intelligence artificielle aujourd'hui. Les tâches qu'il recouvre se définissent par l'apprentissage d'un label correspondant à chaque donnée. Elles nécessitent donc en entrée une base de données annotées. En opposition, l'apprentissage non supervisé se caractérise par une situation où les données d'apprentissage ne disposent pas de labels.

5.0.4 Données tabulaires

Un jeu de données tabulaire se présente sous forme d'un tableau dont les lignes sont des entrées et les colonnes des catégories d'information. Par exemple, un tableur CSV ou une base SQL dans lesquels chaque ligne correspond à un individu et chaque colonne à ses caractéristiques propres (nom, age, profession, etc.).

5.0.5 Données non structurées

À l'inverse des données tabulaires, les données non structurées sont des données qui ne sont pas stockées sous un format prédéfini et où l'information est donc plus dispersée. Cela recouvre par exemple le texte libre comme des documents administratifs, des lettres, des décisions de justice, mais aussi des images, des enregistrement sonores, des vidéos, etc.

5.0.6 Format CoNLL

Le format CoNLL, pour « Conference on Natural Language Learning », est un format général, dont il existe de nombreuses versions et déclinaisons, couramment employé pour les tâches de traitement automatique du langage naturel. Il décrit des données textuelles sous forme de colonne selon un nombre d'attributs : catégorie d'entité nommée, nature grammaticale, etc. Il permet ainsi de stocker un texte annoté.

5.0.7 Hyper-paramètre

En apprentissage automatique, un hyperparamètre est un paramètre qui ne peut pas être appris lors de la phase d'apprentissage de l'algorithme. Par exemple, pour un réseau de neurones profond, ce sera à l'utilisateur de fixer à la main le nombre de neurones qu'il souhaite mettre dans son réseau. Il ne s'agit cependant pas d'un choix arbitraire : chaque hyper-paramètre est ensuite optimisé par les data scientists pour sélectionner la valeur qui permet les meilleures performances.

5.0.8 Librairie (code)

En informatique, une librairie, aussi appelée « bibliothèque de code » ou « package », est un ensemble de code prêt à l'usage qui peut être facilement importé et réutilisé par un utilisateur pour que celui-ci n'ait besoin de réécrire ces portions de code. Par exemple, la librairie « NLTK » sous Python permet d'utiliser tout un ensemble de méthode pour pré-traiter des données textuelles. Les librairies permettent ainsi un énorme gain de temps en évitant que de nombreux développeurs ne codent des portions de code identiques chacun de leur côté.

5.0.9 Modèle de langage

Un modèle de langage est un modèle qui permet d'associer à chaque mot une représentation sous forme d'un vecteur, aussi appelé *embedding*. De tels modèles sont nécessaire puisque les algorithmes d'IA ne savent travailler qu'avec des données numériques. Ces modèles, entrainés sur des millions de

documents, ne sont pas spécifiques à chaque projet mais souvent conçus en amont pour être réutilisés par la suite. On pourra par exemple citer CamemBERT pour le français.

5.0.10 Moteur de règles

Un moteur de règles est un ensemble de règles prédéfinies « à l'avance ». Par exemple, une règle de pseudonymisation pourrait être « si le mot qui suit "Monsieur" ou "Madame" commence par une majuscule, alors ce mot est un prénom ». La complexité du langage naturel et la diversité des formulations qui se trouvent dans les documents fait que ce type de moteur de règles a de forte chance de faire beaucoup d'erreurs dans des textes complexes, ou dont la forme varie souvent.

5.0.11 Pseudonymisation

Le guide de la CNIL sur l'anonymisation des données définit l'anonymisation comme « un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires. En pratique la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro dans un classement, etc.) ».

5.0.12 Reconnaissance d'entités nommées

La reconnaissance d'entités nommées, ou *Named Entity Recognition* (NER) en anglais, est une tâche d'apprentissage supervisée où les données d'entrées sont chacun des mots d'un document et le label qu'on leur attribue est une catégorie sémantique à laquelle il se rattache : par exemple « nom », « prénom », « adresse », « date ».

5.0.13 Reconnaissance optique de caractères

La reconnaissance optique de caractères, plus connue sous son acronyme anglais « OCR » pour *Optical Character Recognition*, désigne le processus de reconnaissance de texte à partir d'images contenant du texte (comme par exemple des scans de lettres dactylographiées ou encore des PDF images) permettant d'extraire ce texte sous un format standard.

5.0.14 Réseaux de neurones profonds

Les réseaux de neurones profonds, appelés *deep learning* en anglais, désignent des architectures complexes de réseaux de neurones artificiels basés sur de nombreuses couches de neurones successives. Ces réseaux cherchent à modéliser des structures sous-jacentes des données avec un haut niveau d'abstraction pour des tâches d'IA complexes, comme la vision par ordinateur ou le traitement du langage naturel.

5.0.15 Stop word

Un *stop word*, ou « mot vide » en français, est un mot très commun et peu porteur de sens seul (on dit qu'ils ne sont pas significatifs). De bons exemple sont ainsi « le », « la », « de », « ou », etc. En traitement automatique du langage naturel, ces mots sont souvent retirés du texte puisqu'ils ne portent que peu d'information du fait de leur présence indistincte dans presque toutes les phrases d'un document.

5.0.16 Tokenisation

En traitement du langage naturel, la tokenisation désigne le fait de décomposer un texte sous forme d'éléments unitaires, des *tokens*, qui seront ensuite représentés sous forme de vecteurs (voir « Modèle de langage »). La tokenisation la plus répandue consiste simplement à tokeniser une phrase sous forme d'un token par mot.

5.0.17 Traitement automatique du langage naturel

Le traitement automatique du langage naturel (TAL), aussi connu sous l'acronyme « NLP » pour *Natural Language Processing*, désigne le domaine de l'intelligence artificielle qui s'intéresse au texte. Il regroupe plusieurs grands types de tâches, comme la reconnaissance d'entités, les agents conversationnels, la classification de documents, le question-answering, etc.