

SU Kevin
MAM David
Encadrant : Stéphane Lopes

Données de justice en Open Data



2023-2024

Table des matières

Introduction

1 État de l'art

1.1 API Judilibre

1.1.1 Utilisation

1.2 Pseudonymisation

1.2.1 Conséquence de la violation des droits

1.3 Exploitation des données juridiques

2 Extraction des données

2.1 Prétraitement des textes

2.1.1 Définition

2.1.2 Les textes bruts

2.1.3 Le Natural Language Processing (NLP)

2.2 Topic modeling

2.3 Cluster TF-IDF

2.3.1 Définition

2.4 Application

Introduction

Données de justice en Open Data consiste à rendre accessibles gratuitement et sous forme électronique aux personnes du public les décisions rendues publiquement par les juridictions judiciaires et les administrations.

Cette disposition est menée par La cour de cassation et le Conseil d'État, et a pour but de mettre en avant la jurisprudence afin d'assurer la transparence de la justice et de renforcer la confiance des citoyens dans leurs justices, ainsi favoriser l'accès aux données, qui profite à la fois aux citoyens et aux professionnels qui pourront se servir de ce registre pour prendre leurs décisions en prenant en compte les pratiques jurisprudentielles qui ont eu lieu dans des situations similaires à leurs propres cas.

Les premières décisions ont été publiées le 30 septembre 2021. Ce sont des décisions rendues par la Cour de cassation, aujourd'hui, on y trouve des décisions de cours d'appel, civil, social et commercial qui représente environ 230 000 décisions par an, et l'objectif est d'inclure l'ensemble des décisions de l'ordre judiciaire qui représente environ 3 000 000 décisions par an. Avec ces données massives, l'enjeu principal de ce projet est la protection de la vie privée et des données à caractère personnel des parties prenantes aux décisions publiées afin de trouver un équilibre pour maximiser les avantages de l'open data judiciaire.

Dans un premier temps, nous allons présenter l'API judilibre, une interface de programmation d'application, qui permet d'accéder aux décisions et nous allons voir d'utilisation de ce dernier.

Ensuite, nous allons traiter la protection de la vie privée, les différents algorithmes pour effectuer la pseudonymisation.

Et pour finir, nous allons expérimenter différents algorithmes d'extraction de connaissances sur ces données

État de l'art

1.1 API Judilibre

API (Application Programming Interface), est une interface informatique visant à connecter un logiciel ou une application à d'autres systèmes distincts afin qu'ils puissent échanger leurs fonctionnalités, leurs services, leurs technologies et leurs données. L'API se matérialise comme une passerelle d'accès à une fonctionnalité détenue par une entité indépendante.

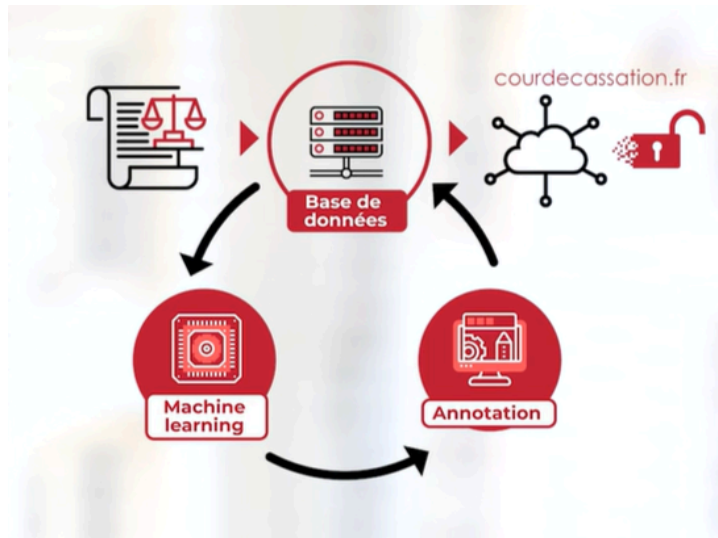
Grâce à un langage de programmation universel, elle favorise l'interaction entre utilisateurs et fournisseurs par l'envoi de requêtes d'accès aux services du fournisseur. Elle facilite la création et l'intégration de fonctionnalités afin que les développeurs n'aient pas besoin de maîtriser complètement le programme qu'ils souhaitent exploiter.

API agit en tant qu'intermédiaire entre deux systèmes informatiques indépendants afin qu'ils puissent échanger et exploiter les fonctionnalités d'une application et le contenu d'une base de données ou d'un système de fichiers.

API Judilibre est développé par les services de la Cour de cassation pour la mise en open data des décisions de justice de l'ordre judiciaire. Elle a pour but de permettre une récupération et une réutilisation facilitées des décisions rendues publiquement par la Cour de cassation et par les chambres civiles, sociales et commerciales des cours d'appel, éventuellement enrichies, et pseudonymisées.

L'accès à l'API est ouvert à toute personne, l'utilisation de la base de données est libre et gratuite après inscription. Elle est hébergée par la plateforme PISTE (Plateforme d'intermédiation de services pour la transformation de l'État), un outil utilisé pour la publication des API des ministères économiques et financiers et est également proposée à la communauté interministérielle. Cette plateforme a surtout fortement contribué au développement de la culture API dans l'administration française, ainsi qu'à la transformation des stratégies d'échanges de systèmes d'information.

Les données disponibles via l'API sont également celles de la version du site de la Cour de cassation (<https://www.courdecassation.fr/recherche-judilibre>). Dans la base de données, on trouve les décisions rendues publiquement par la Cour de cassation éventuellement enrichies et pseudonymisées. Au fur et à mesure du calendrier établi par l'arrêté du 28 avril 2021, la base de données sera enrichie de décisions rendues par d'autres juridictions de l'ordre judiciaire.



Cette figure illustre la mise en pratique des décisions envoyées dans la base de données.

Une fois que les données de justice sont rendues et numérisées, un moteur de machine learning effectue un premier traitement de manière à annoter et occulter certains éléments tels que, les noms, prénoms, adresse etc... un algorithme de pseudonymisation que nous allons traiter plus tard, ensuite un travail manuel sera effectué afin de vérifier le travail du moteur machine learning, puis de le stocker dans la base de données et à disposition de tout le monde

1.1.1 Utilisation de l'API

L'api Judilibre offre de nombreuses fonctionnalités pour optimiser et organiser les recherches juridiques, comme des recherches par mot clés ou bien limiter ses recherches à certaines juridictions voire même à une partie de la décision.

Et les recherches peuvent être téléchargées de manière totalement gratuite en format PDF.

Les prérequis avant d'accéder à l'API :



- Création d'un compte sur la plateforme PISTE
<https://piste.gouv.fr/component/apiportal/registration>
- Validé les conditions générales d'utilisations à travers l'onglet API -> consentement CGU API et cocher Judilibre

Sélectionnez les API

Valider mes choix CGU

judilibre

Consentement CGU

<input type="checkbox"/>	Nom de l'API	Environnement	CGU
<input checked="" type="checkbox"/>	JUDILIBRE	PROD	 CGU_open_data_V8.pdf (pdf, 845.9Ko)
<input checked="" type="checkbox"/>	JUDILIBRE	SANDBOX	 CGU_open_data_V8.pdf (pdf, 845.9Ko)

lignes 1 à 2 sur 2 (Filtrer un maximum de 46) Précédent 1 Suivant 10 Afficher lignes

- Raccorder l'application Sandbox en appuyant sur l'onglet Application puis cliquer sur "Modifier l'application" et cocher Judilibre

Ingres Noyau v2	2.0.0	-	SANDBOX	<input type="checkbox"/>	Demander l'accès	Oui
JUDILIBRE	1.0.0	-	SANDBOX	<input checked="" type="checkbox"/>		Oui
Légitime	2.0.0	-	SANDBOX	<input type="checkbox"/>		Oui

À présent, nous avons la possibilité d'utiliser les différentes fonctionnalités de l'API que nous allons détailler par la suite.

Une fois entrées sur l'application, nous avons découvert deux sections : default et schemas.

Dans la partie default, on peut trouver différentes saisies qui servent à extraire des informations spécifiques, on trouve /decision, /taxonomy, /stats, /search, /export et /healthcheck.

Dans la partie schemas, on peut observer différents types d'objets qui sont utilisés pour afficher les résultats des différentes requêtes saisies.

- **/decision** permet de récupérer le contenu intégral d'une décision à savoir l'identifiant de sa juridiction, l'identifiant de sa chambre, sa formation, son numéro de pourvoi, son ECLI (« European Case Law Identifier » : identifiant européen de la jurisprudence), son code NAC, son niveau de publication, son numéro de publication au bulletin, sa solution, sa date, son texte intégral, les délimitations des principales zones d'intérêt de son texte intégral (introduction, exposé du litige, moyens, motivations, dispositif et moyens annexés), ses éléments de titrage, son sommaire, ses documents associés (communiqué, note explicative, traduction, rapport, avis de l'avocat général, etc.), les textes appliqués, les rapprochements de jurisprudence.

Paramètre :

id : Identifiant de la décision

resolve_references : true or false, si ce paramètre vaut true elle retournera aussi les informations telles que chamber, date, publication etc.

query : les termes à surligner dans le texte intégrale

operator : or, and, exact

- **/taxonomy** permet de récupérer les listes des termes employés par le processus de recherche comme la liste des types de décision, la liste des juridictions dont le système intègre les décisions, la liste des chambres, la liste des formations, la liste des niveaux de publication, la liste des matières, la liste des solutions, la liste des champs et des zones de contenu des décisions pouvant être ciblés par la recherche, la liste des zones de contenu des décisions, etc.

Paramètre :

id : liste des taxonomie : "cc", "ca", "tj", "all", "chamber", "date_type", "field", "filetype", "formation", "jurisdiction", "location", "operator", "order", "publication", "solution", "sort", "theme", "type"

key : la clé de l'intitulé complète

value : l'intitulé complète pour obtenir la clé

context_value : valeur pouvant contextualiser certaines listes

- **/stats** permet de récupérer des statistiques sur le contenu de la base JUDILIBRE, ce point ne requiert aucun paramètre et affiche les statistique suivant : nombre de décisions indexées (au total, par année, par juridiction), Nombre de requêtes (par jour, par semaine, etc.), date de la décision la plus ancienne, date de la décision la plus récente.
- **/search** Permet d'effectuer une recherche dans les données ouvertes des décisions de justice, elle affichera des texte en saisie libre, lequel sera mis en correspondance avec tout ou partie du contenu des décisions ; le mode de mise en rapport des termes de la recherche (*ou, et*, expression exacte) ; contenu ciblé par la recherche : décision intégrale, zones spécifiques de la décision (exposé du litige, moyens, motivations, dispositif), sommaire, titrages, numéro de pourvoi, etc. ; nature de décision ; matière ; chambre et formation; juridiction et commission; niveau de publication; type de solution; intervalle de dates; pertinence et date; nombre de résultats par page et index de la page de résultats affichée.

Paramètre :

query : chaine de caractère correspondant à la recherche

field : liste des champs, métadonnées ou zones de contenu ciblés par la recherche

operator : or, and, exact

type : nature des décisions,liste en tapant dans /taxonomy id = type

theme : matière relative aux décisions, liste /taxonomy id = theme

chamber : la liste dans /taxonomy id = chamber

formation : la liste dans /taxonomy id = formation

jurisdiction : la liste dans /taxonomy id = jurisdiction

location : code du siège ayant émis les décisions la liste dans /taxonomy id = location et context_value = ca par exemple

publication : la liste dans /taxonomy id = publication

solution : la liste dans /taxonomy id = solution

date_start

date_end

sort : on peut trier par date, score et scorepub

order : asc ou desc

page_size : nombre de résultat par page

page : le numéro de page à retourner

resolve_references

withFileOfType : type de document /taxonomy id = filetype
particularInterest : true ou false

- **/export** permet d'effectuer un export par lot de décisions de justice.

Paramètre :

type
theme
chamber
formation
jurisdiction
location
publication
solution
date_start
date_end
abridged : version abrégée (sans métadonnée et texte intégral)
date_type : creation ou update
order
batch_size : nombre de résultat par lot
batch : le numéro du lot à retourner
resolve_references
withFileOfType
particularInterest

- **/healthcheck** permet de vérifier la disponibilité du service, aucun paramètre n'est nécessaire

1.2 Pseudonymisation

Le droit au respect de la vie privée et à la protection des données à caractère personnel sont des droits fondamentaux reconnus par des instruments internationaux et nationaux. Ils bénéficient en premier lieu aux personnes physiques.

Pour ce faire, une équipe de la Cour de cassation a mis au point un moteur d'apprentissage automatique pour pseudonymisation des décisions qu'on appelle aussi moteur de traitement automatique des langues (TAL). Dans un premier temps, le modèle du langage (aspect du TAL) permet d'obtenir des

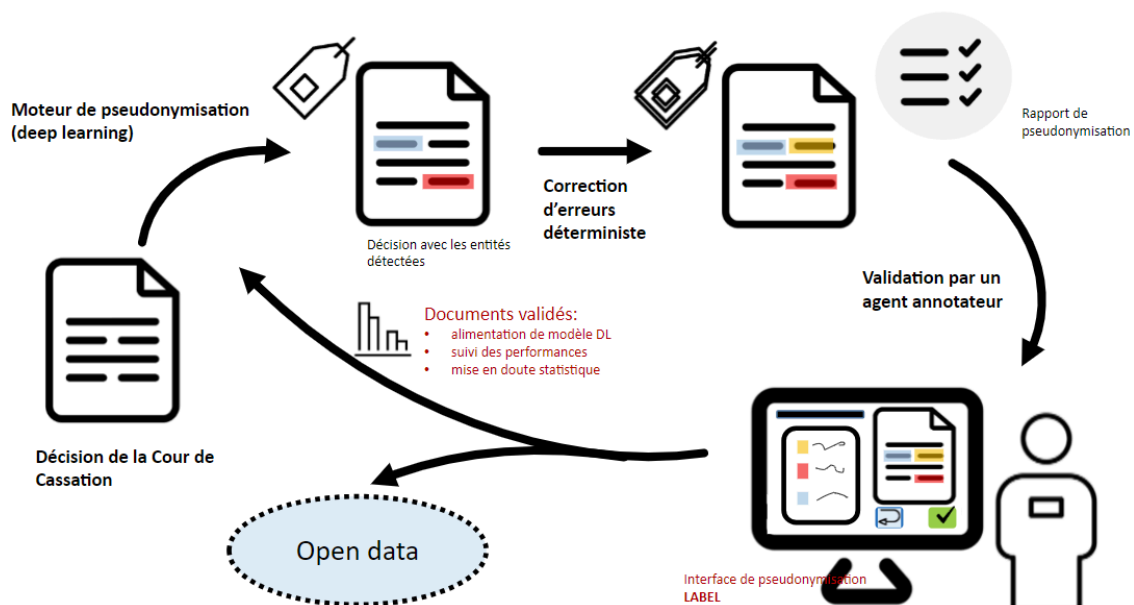
embeddings, des mots représentés par un vecteur dans un espace à plusieurs dimensions.

Dans leur modèle de production il utilise une combinaison de BytePairEmbeddings et de FlairEmbeddings..

Les **BytePairEmbeddings** sont des embeddings de mots précalculés au niveau des sous-mots. Cela signifie qu'ils sont capables d'intégrer n'importe quel mot en divisant les mots en sous-mots et en recherchant leurs embeddings.

Les **FlairEmbeddings** sont des embeddings puissants qui capturent des informations syntaxiques et sémantiques latentes qui vont au-delà des intégrations de mots standard. Les principales différences sont les suivantes : (1) ils sont formés sans aucune notion explicite de mots et modélisent donc fondamentalement les mots comme des séquences de caractères. Et (2) ils sont contextualisés par le texte qui les entoure, ce qui signifie que le même mot aura des intégrations différentes en fonction de son utilisation contextuelle.

Le modèle de reconnaissance d'entités nommées, ce modèle



permet de rechercher les mots à annoter.

La figure illustre le processus de la pseudonymisation des décisions avant de les mettre en open data.

Une fois que les décisions ont été validées par le moteur de pseudonymisation, une équipe d'annotateur vérifie les décisions pré-annotées

par le moteur avec le logiciel d'Anonymisation d'une Base Enrichie Labellisée intitulé LABEL puis envoyé dans la base de données.

1.2.1 Conséquence de la violation des droits

Le droit au respect de la vie privée et à la protection des données à caractère personnel sont des droits fondamentaux reconnus par des instruments internationaux. Au sein de l'Union européenne, le Règlement Général sur la Protection des Données (RGPD), entré en vigueur en France en 2018, met en œuvre ce droit fondamental en encadrant les traitements de données à caractère personnel.

Les différents principes du RGPD sont donc applicables aux informations sur des personnes physiques contenues dans les décisions, de telle sorte que le législateur français a entendu mettre en œuvre l'open data en opérant une conciliation avec les droits fondamentaux des justiciables.

La Commission nationale de l'informatique et des libertés (CNIL) a précisé les modalités d'anonymisation des décisions dans un avis du 29 novembre 2001. Elle rappelle que doivent être anonymisés le nom et l'adresse des parties et des témoins dans tous les jugements et arrêts librement accessibles sur internet, quels que soient l'ordre ou le degré de la juridiction et la nature du contentieux. Il n'est en revanche pas nécessaire d'occulter l'identité des magistrats ou membres des juridictions, ni celle des auxiliaires de justice ou experts. Cependant, les noms des juges non professionnels en matière pénale sont systématiquement occultés (jurés d'assises, assesseurs des tribunaux pour enfants, etc).

Avec l'entrée en vigueur du RGPD, la pseudonymisation remplace l'anonymisation qui consiste à remplacer les données directement identifiantes (nom, prénom, etc.) par un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.), un projet récent par la Cour de cassation cité plus haut.

En cas de manquement à l'obligation de respect de la vie privée ou d'insuffisance d'anonymisation de la décision, les personnes victimes, en cas d'atteinte disposent de plusieurs recours.

- Sur le fondement de l'article 38 de la loi du 6 janvier 1978, elles peuvent s'opposer à la poursuite du traitement les concernant en faisant valoir un motif légitime devant la CNIL ou un juge.

- L'article 17 du RGPD prévoit l'obtention dans les meilleurs délais de l'effacement des données à caractère personnel les concernant (droit à l'oubli).
- L'article 21 du même règlement permet à la personne concernée de s'opposer à tout moment aux traitements de ses données personnelles (droit d'opposition).

La diffusion d'une décision non anonymisée est en outre susceptible de recevoir une qualification pénale. Le manquement à l'obligation d'anonymisation d'une décision diffusée s'analyse en un traitement automatisé de données personnelles, punissable sur le terrain du droit à la protection des données personnelles. Le responsable de ces agissements encourt jusqu'à cinq ans d'emprisonnement et de 300 000 euros d'amende vertu des articles 226-16 et suivants du Code pénal sanctionnant les infractions à la législation sur les données personnelles. Le législateur a également interdit la réutilisation des données relatives à l'identité des magistrats et greffiers afin de réaliser des analyses statistiques permettant de prédire ou de comparer la manière de traiter les contentieux en fonction des juridictions.

1.3 Exploitation des données juridiques

De nombreuses recherches ont exploité les données juridiques pour étudier divers aspects du droit et de la justice.

Parmi ceux-ci, on peut citer l'habilitation à diriger des recherches (HDR) sur "De quoi la 'justice prédictive' est-elle le nom ? Algorithmes, décision et jugement" présenté par Laurence DUMOULIN.

La 'justice prédictive' désigne un ensemble disparate de plateformes en ligne qui, à partir d'un grand nombre de décisions de justice, dont le traitement est assuré par différentes méthodes mathématiques, statistiques et/ou algorithmiques, simulent différentes décisions de justice possibles sur un cas donné et évaluent leur probabilité respective. La montée en puissance de la "justice prédictive" est interprétée comme la manifestation de l'émergence d'une nouvelle configuration de réforme administrative qui s'est mise en place en France depuis les années 2010 pour favoriser l'open data des décisions de justice et encourager l'innovation dans le droit. Cependant, il existe des préoccupations de l'utilisation de l'analyse prédictive dans le domaine de la justice.

Ces préoccupations incluent des questions telles que la transparence des algorithmes utilisés, le risque de biais algorithmique, la protection de la vie privée des individus concernés par les affaires judiciaires et l'impact potentiel sur l'équité et l'accès à la justice.

On peut citer aussi une revue “ L'accès numérique au droit “ par Roseline LETTERON Professeur de droit public, Sorbonne Université.

Cette revue évoque également la justice prédictive qui permettrait de lutter contre l'engorgement des tribunaux en offrant la possibilité de régler rapidement les contentieux de masse par des décisions parfaitement standardisées. On pourrait ainsi l'envisager pour définir les prestations compensatoires en matière de divorce, l'indemnisation des licenciements abusifs ou encore le contentieux des obligations de quitter le territoire.

Elle cite que l'accès aux données de justice en open data contraint les juges à motiver soigneusement leurs décisions. S'il n'est pas contrôlé, il risque de conduire à une justice automatisée et conservatrice, incitant les requérants à se tourner vers des modes alternatifs de règlement des litiges, souvent très onéreux et moins fiables que la décision juridictionnelle.

Extraction des données

2.1 Prétraitement des textes

2.1.1 Définition

Dans cette partie, nous allons plus parler algorithme. L'API Judilibre permet de faire une recherche par mot-clé, mais il existe également des moyens pour extraire du texte. Cependant, il est important de prendre seulement les informations pertinentes, et c'est donc là que vient le prétraitement.

Le prétraitement est une étape qui cherche à standardiser du texte, et ainsi rendre son usage plus facile, ce qui nous permet d'avoir les informations qu'on veut plus facilement.

Dans le prétraitement, on traite chaque mot d'un texte différemment, pour ainsi pouvoir chercher les informations pertinentes plus facilement. Mais pas seulement, on peut aussi corriger certaines erreurs comme les fautes d'orthographe évidentes ou les incohérences typographiques. Traiter chaque

mot différemment peut aussi nous aider à expliciter des informations manquantes grâce à des ressources externes.

2.1.2 Les textes bruts

Traiter des textes bruts peut s'avérer difficiles dû à deux raisons: les incohérences et les ambiguïtés. Celles-ci doivent être impérativement traitées avant de passer à l'extraction, et la correction de ces problèmes permet un traitement unifié lors des étapes ultérieures d'extraction d'informations. Pour citer des exemples d'incohérences, nous avons le tableau ci-dessous:

1. (Grefenstette et Tapanainen, 1994)
Supprimer les étiquettes SGML
Recoller les césures
2. (Adda et al., 1997)
Encodage des accents et autres diacritiques
Prétraitement des nombres et unités
Correction du formatage et des ponctuations
Traitement des ponctuations non ambiguës
Pas de distinction de casse
Pas de diacritiques
3. SXPIPE 1.0 (Sagot et Boullier, 2005)
Réaccentuation et recapitalisation
4. Notre modèle (Amrani et al., 2004)
Remplace les caractères non ASCII et entités
Convertit le document en format texte linéaire
Normalise le paragraphe - Incohérences

Le format variable d'encodage des textes (tels que des caractères accentués encodés de manière différente d'un texte à l'autre) peut être vu comme un premier type d'incohérence. C'est un problème qui se situe au niveau du

caractère, et un encodage unique est nécessaire avant de passer au traitement des mots.

Nous avons ensuite les marqueurs incohérents de présentation ou de sens (tels que des balises XML). Ces marqueurs doivent tout simplement être supprimés (ou transformés) car ils sont souvent incohérents d'un standard à l'autre, voire d'une version à l'autre.

Et en dernier, nous avons la présence de fautes d'orthographe ou d'incohérences typographiques, comme par exemple des majuscules incorrectement utilisées.

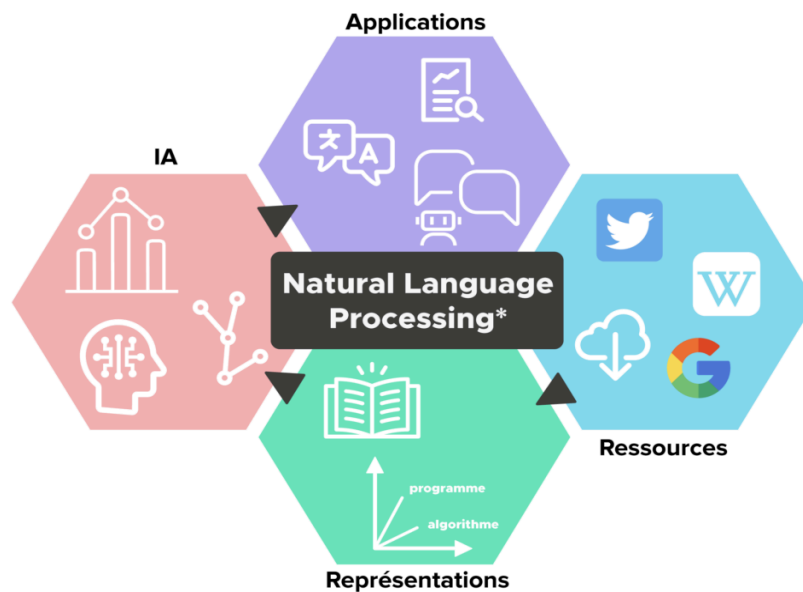
Après les incohérences, on parle d'ambiguïtés. Ils doivent être explicités afin que les traitements ultérieurs de niveau syntaxique ou sémantique ne soient pas trop pénalisés par le manque d'informations des niveaux inférieurs. Pour citer quelques-uns, nous avons d'abord les ambiguïtés lexicales, et ensuite la structure des textes non encodée ou de façon partielle (comme par exemple la fin des phrases et des paragraphes).

Ainsi, avec cette étape du prétraitement, on peut déterminer par exemple si un mail est un spam ou non en transformant le texte brut du mail en des données exploitables. Ce qui est aussi utile, bien évidemment, pour l'extraction sur les données de justice de Judilibre.

2.1.3 Le Natural Language Processing (NLP)

Le NLP (ou Traitement du Langage Naturel) est une discipline qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines. C'est ce qu'utilisent par exemple des assistants personnels IA tels que Siri ou Cortana, ou alors les correcteurs d'orthographe pour détecter les erreurs de syntaxes.

Le schéma suivant montre les problématiques à laquelle le NLP répond:



**Traitement Automatique du Langage Naturel*

 DataScientest • com

Sans rentrer dans les détails, si on veut comprendre son fonctionnement, le NLP fait évidemment un prétraitement des textes qu'il traite.

On passe donc par plusieurs étapes:

- Tokenisation: c'est tout simplement le fait de traiter chaque mot d'un texte différemment. Dans ce cas, on dit qu'on découpe en plusieurs *tokens*.
- Stemming: on découpe un mot afin de conserver uniquement sa racine. Par exemple, "décisions" devient alors "décis". Ceci permet à un mot avec plusieurs variantes d'être considéré comme un seul et même mot.
- Lemmatisation: c'est à peu près la même chose que le Stemming mais il utilise un vocabulaire et une analyse fine de la construction des mots, ce qui permet donc de supprimer uniquement les terminaisons inflexibles. Par exemple, "décidez" devient "décider".

Avec ces étapes (entre autres) du prétraitement, le NLP peut donc traiter efficacement un texte donné.

(Petite note: nous avons testé le prétraitement avec des bibliothèques comme SpaCy ou nltk dans le github prévu à cet effet.)

2.2 Topic modeling

Après avoir fait du prétraitement dans le texte qu'on veut extraire, il est maintenant temps d'attaquer le coeur du sujet d'extraction. Une des méthodes pour extraire des données est le Topic modeling.

Le Topic modeling est un type de modélisation statique qui utilise le "Machine Learning" pour identifier des clusters ou des groupes de mots similaires dans un corps de texte.

Cette méthode de recherche peut extraire des données sans avoir besoin de tags prédéfinis, en utilisant des structures sémantiques. Ainsi, il peut identifier exactement le contenu d'un document (comme un contrat, une facture, etc.) grâce à son analyse.

L'analyse sémantique latente et l'analyse Dirichlet latente sont deux méthodes principales du topic modeling et ils analysent de grands fichiers texte pour catégoriser les sujets, fournir des informations précieuses et faciliter la prise de décision.

Pour faire court et être plus précis ce que sont ces deux méthodes:

- le LSA (Latent Semantic Analysis) est une technique statistique permettant d'extraire et de représenter les idées principales d'un texte. L'analyse sémantique latente repose sur le principe selon lequel les mots dont le sens est proche ont tendance à être utilisés ensemble dans le contexte.

- le LDA (Latent Dirichlet Analysis) est une méthode qui permet de découvrir la structure cachée d'un ensemble d'observations en examinant les relations entre les mots d'un document et en les regroupant en thèmes.

Ensuite, si on veut comprendre le fonctionnement du topic modeling, il faut savoir que c'est en fait "simple": il déduit les mots, en regroupant les modèles de mots similaires dans des thèmes pour créer des groupes de thèmes.

Par exemple, le topic modeling peut analyser d'énormes quantités de données non structurées pour trouver des modèles basés sur la fréquence des mots, l'ordre, la distance et le sens, et regrouper divers éléments dans des catégories pertinentes sans formation prédéfinie.

2.3 Cluster TF-IDF

2.3.1 Définition

Le TF-IDF (Term Frequency-Inverse Document Frequency) est aussi une autre méthode de pondération qu'on utilise pour la fouille de texte et qui utilise le "clustering". Cette mesure statistique permet d'évaluer l'importance d'un mot dans un corps de texte. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la répartition du mot dans le corps de texte.

Le TF-IDF possède deux "définitions" formelles, comme on peut le voir dans son nom. Nous allons expliquer pour chacun de ces deux:

Term Frequency: La fréquence du terme est tout simplement le nombre d'occurrences de ce terme dans le texte qu'on traite, et qu'on divise par le nombre de mots du document. Il existe cependant d'autres variantes pour exprimer la fréquence du terme (le plus simple étant juste de prendre le nombre d'occurrences de ce terme).

Inverse Document Frequency: La fréquence inverse de document est une mesure de l'importance du terme dans l'ensemble du corps de texte. Elle vise à donner un poids plus important aux termes les moins répartis, considérés comme plus discriminants. Le calcul de cette fréquence utilise des notions logarithmiques: plus la valeur approche de 0, plus le mot est commun.

2.3.2 Différences entre Topic modeling et TF-IDF

Bien que ces deux méthodes regroupent des documents et analysent chaque mot d'un document, leur manière de faire est différente. Le topic modeling permet d'identifier le sujet et le contenu d'un document, tandis que le TF-IDF calcule la fréquence et l'importance d'un mot dans un document. Pour le cas de TF-IDF, comme on parle de "clustering", il faut donc savoir que la fréquence des mots est calculée selon un groupe avec des données similaires. Pour le topic modeling, il utilise une approche statistique pour trouver les sujets cachés dans une collection de documents.

2.4 Application

<https://github.com/Galaboost/NLP>