Notes:

Pour la problématique :

# « Digitalisation du processus de recrutement :

# Etude du développement d’un analyseur de CV dans le cadre d’une approche orientée Data Science »

# Table des matières

[Table des matières 2](#_Toc63359177)

[Sources 3](#_Toc63359178)

[Introduction 4](#_Toc63359179)

[La digitalisation du processus de recrutement 4](#_Toc63359180)

[Définitions 4](#_Toc63359181)

[Le recrutement 5](#_Toc63359182)

[La digitalisation 5](#_Toc63359183)

[Quels avantages ? 5](#_Toc63359184)

[Gain de temps et d’argent 5](#_Toc63359185)

[Simplification et optimisation du processus de recrutement 6](#_Toc63359186)

[Meilleure attraction des candidats 6](#_Toc63359187)

[Une expérience candidat améliorée 6](#_Toc63359188)

[Comment digitaliser la fonction recrutement d’une entreprise ? 6](#_Toc63359189)

[Les outils pour digitaliser le recrutement 6](#_Toc63359190)

[Recruter avec l’intelligence artificielle : 8](#_Toc63359191)

[L’extraction d’information et le Traitement Automatique du Langage Naturel 11](#_Toc63359192)

[L’extraction d’information 11](#_Toc63359193)

[Définition 11](#_Toc63359194)

[Historique 12](#_Toc63359195)

[Le Traitement Automatique du Langage Naturel 12](#_Toc63359196)

[Définition 12](#_Toc63359197)

[Historique 13](#_Toc63359198)

[Un projet de type « Data Science » 15](#_Toc63359199)

[L’analyse automatique de CV 16](#_Toc63359200)

[Définition 16](#_Toc63359201)

[Un analyseur générique, ou spécifique ? 17](#_Toc63359202)

[Deux mesures clés 18](#_Toc63359203)

[Problématiques 18](#_Toc63359204)

[Quatre grands types d’approches 19](#_Toc63359205)

[Différentes étapes successives 20](#_Toc63359206)

# Sources

[1] « Recrutement », *Wikipédia*. janv. 25, 2021, Consulté le: janv. 31, 2021. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Recrutement&oldid=179193823.

[2] J. Ready, « Comment digitaliser ses recrutements ? Astuces & Conseils ! », *Culture RH*, juin 24, 2020. https://culture-rh.com/comment-digitaliser-processus-recrutement/ (consulté le janv. 24, 2021).

[3] « Que signifie Données semi-structurées? - Definition IT de Whatis.fr », *Whatis.com/fr*. https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees (consulté le févr. 02, 2021).

[4] « Information extraction », *Wikipedia*. déc. 05, 2020, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Information\_extraction&oldid=992409176.

[5] A. Turing, « Computing machinery and intelligence », *Oxford University Press*, vol. 59, no 236, oct. 1950, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://www.csee.umbc.edu/courses/471/papers/turing.pdf.

[6] « Georgetown–IBM experiment », *Wikipedia*. déc. 12, 2020, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Georgetown%E2%80%93IBM\_experiment&oldid=993811492.

[7] « ELIZA », *Wikipédia*. sept. 30, 2020, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=ELIZA&oldid=175170901.

[8] « Universal grammar », *Wikipedia*. janv. 03, 2021, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Universal\_grammar&oldid=998100233.

[9] « SHRDLU », *Wikipédia*. déc. 09, 2020, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=SHRDLU&oldid=177446296.

[10] « Apprentissage automatique », *Wikipédia*. août 06, 2020, Consulté le: sept. 27, 2020. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Apprentissage\_automatique&oldid=173595588.

[11] « Why we need a methodology for data science », *IBM Big Data & Analytics Hub*. https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science (consulté le janv. 17, 2021).

[12] J. Atlan, « Le CV parsing, c’est quoi ? » https://www.easyrecrue.com/fr/blog/parsing-cv-candidatures (consulté le janv. 16, 2021).

[13] « PHRASÉOLOGIE : Définition de PHRASÉOLOGIE ». https://www.cnrtl.fr/lexicographie/phras%C3%A9ologie (consulté le févr. 02, 2021).

[14] S. Sonar et B. Bankar, « Resume Parsing with Named Entity Clustering Algorithm », p. 6.

<https://culture-rh.com/comment-digitaliser-processus-recrutement/#La_digitalisation_definition>

<https://culture-rh.com/recruter-intelligence-artificielle-avantages-inconvenients-exemples-application/>

# Introduction

**TODO : Retravailler l’introduction**

Le recrutement, processus charnière du monde du travail, connaît aujourd’hui une transformation en profondeur : sa digitalisation pour un nombre sans cesse plus grand d’organisations. La « transformation numérique » des entreprises ne passe pas seulement par l’utilisation d’outils tels que Microsoft Teams ou Trello. Le monde de l’entreprise accueille chaque jour des approches nouvelles, imaginés en accord avec des habitudes digitales contemporaines.

L’un des enjeux réside donc dans l’adaptation des méthodes professionnelles à ce monde « 2.0 ».

La digitalisation des processus du recrutement, par exemple, permet d’appréhender des axes déjà connus, comme la cooptation, sous un angle différent et des méthodes comme la « gamification » recherchent à améliorer les approches « de terrain ». Les progrès en matière d’apprentissage automatique ouvrent des portes jusque-là fermées (algorithmes).

Avec plus de 15 ans d’ancienneté, les réseaux sociaux professionnel sont une pierre angulaire de la digitalisation des processus de ressources humaines. Ils ont révélé des possibilités nouvelles et c’est sur leurs modèles que se sont développé un certain nombre de plateformes de recrutement digital.

Ils facilitent la prise de contact entre recruteurs et candidats, font jouer les notions de réseaux et permettent aux flux d’informations de sortir des filières traditionnelles du recrutement, les faisant grossir et se croiser. L’une des unités d’information de ces flux est le CV.

Environ 1,4 milliard de CV sont analysés chaque année. Avec le grossissement des flux de données, et l’émergence des Big Datas, l’une des problématiques émergentes dans le domaine des ressources humaines est la structuration des données de CV.

En effet, qu’il s’agisse de recherche de talents ou de création d’outil d’appariement entre candidats et emplois, les projets nécessitant de telles structurations se révèlent souvent cruciaux pour le bon développement d’une entreprise. Les bénéfices peuvent aller de la simple amélioration de l’expérience utilisateur, à divers progrès liés aux stratégies de recrutement ou de positionnement.

De ce fait, la digitalisation et l’automatisation des processus de recrutement connait aujourd’hui une véritable montée en popularité. Les différents acteurs de l’industrie embrassent uns à uns, avec plus ou moins d’enthousiasme, les technologies qui permettent cette évolution.

L’analyse automatique de CV se propose comme un élément de solution à ce problème. Elle est d’autant plus cruciale qu’aucune recherche ne pourra renvoyer le meilleur candidat si les bonnes informations n’ont pas toutes été chargées correctement dans les bases de données.

Plusieurs dirigeants de l’industrie des logiciels de recrutement décrivent les trois principaux atouts des analyseurs de CVs comme ceci :

- Ils permettent un gain de rapidité ;

- Ils offrent une meilleure expérience aux candidats ;

- Ils permettent d’obtenir des données utilisables ou mieux utilisables par un ATS

# La digitalisation du processus de recrutement

## Définitions

Il faut pour commencer s’attarder sur les définitions de “recrutement” et “digitalisation” :

### Le recrutement

* Le recrutement, c’est “*l’ensemble des actions mises en œuvre pour trouver un candidat censé correspondre aux besoins et compétences considérées comme requises à un poste de travail donné, dans une organisation données.”*[1]

Il s’agit donc, à l’instar d’un projet, d’un processus regroupant des activités, coordonnées et rapprochées pour accomplir un objectif précis, en l’occurrence l’embauche du meilleur candidat possible pour un poste donné.

### La digitalisation

* La digitalisation, c’est “*le procédé qui vise à transformer des processus traditionnels, des objets, des outils, ou encore des professions par le biais de technologies digitales dans le but de les rendre plus performants.”*[2]

La digitalisation du processus de recrutement est donc le procédé par lequel on va chercher à transformer une ou plusieurs des activités intervenant dans le processus d’embauche d’un candidat pour obtenir un gain de performance par le biais des technologies digitales.

Souvent apparentée à la “transformation digitale[[1]](#footnote-1)”, la digitalisation a commencé dès les débuts d’internet : les courriers postaux sont devenus les emails ; les magasins et leurs vitrines, les sites d’e-commerce.

La digitalisation est un processus presque naturel, en adéquation avec l’évolution de notre société vers toujours plus de numérique.

Dans le domaine du recrutement, la digitalisation permet :

* De gagner du temps, en automatisant les tâches répétitives effectuées tant par les candidats que par les recruteurs ;
* De s’affranchir des limitations en matière de distance : avec l'appui des technologies digitales, il est devenu beaucoup plus simple de d’effectuer un recrutement à distance. Il est en effet de moins en moins nécessaire de se déplacer en train ou en avion pour une rencontre ;
* De mieux collaborer : le temps passé à joindre la bonne personne du bon service, ou même à s’y rendre. Un email, ou une plateforme de recrutement permet de regrouper les principaux acteurs d’un recrutement ;
* De limiter les erreurs : il devient plus facile et plus rapide de détecter les erreurs et donc de les corriger
* De maximiser les profits : la digitalisation du processus de recrutement permettra de faire gagner de l’argent au service de recrutement des organisations qui entreprendront de telles démarches.

**TODO : TRANSITION**

## Quels avantages ?

Pour le recrutement, les avantages de la digitalisation sont nombreux :

### Gain de temps et d’argent

Le premier avantage à noter est le gain de temps et d’argent procuré par cette digitalisation. Le recruteur peut automatiser une grande partie de son processus de recrutement : la gestion des candidats ou des offres (sauvegarde, tri, filtrage, etc.) ou les tests de recrutements peuvent être digitalisés à leurs différents niveaux.

Cette digitalisation permettra au recruteur de s’affranchir d’un certain nombre de tâches répétitives et fastidieuses et ce gain de temps sera rapidement transformé en un gain d’argent.

De plus, les technologies digitales proposent un panel large de manières d’attirer les candidats potentiels (gamification).

### Simplification et optimisation du processus de recrutement

Avec l’automatisation due à la digitalisation, le dépôt de candidature est facilité pour le candidat, et la gestion des candidatures est simplifiée pour le recruteur.

A cela s’ajoute le nombre sans cesse grandissant d’outils analytiques digitaux, qui permettent au recruteur de s’assurer de la bonne adéquation de son processus de recrutement. Il peut, ou pourra, l’adapter et l’optimiser en temps réel.

L’un des principaux sujets de préoccupation des recruteurs est la marque employeur[[2]](#footnote-2). Avoir une bonne “marque employeur” permet à la fois d’attirer plus facilement les candidats, mais aussi de limiter le turnover.

Grâce aux outils digitaux, le service de recrutement pourra améliorer et renforcer en permanence sa marque employeur : mise en place de page carrières, entre autres.

### Meilleure attraction des candidats

La digitalisation du recrutement permet de mieux cibler les candidats visés, tout en renforçant l’attractivité de l’entreprise auprès de ces derniers.

### Une expérience candidat améliorée

Chaque jour de plus en plus en tension, le marché de l’emploi est devenu de plus en plus exigeant. Attirer les candidats peut parfois se révéler compliqué, et pouvoir offrir une expérience améliorée, au travers d’un processus attractif, se révélera être un véritable atout dans ce contexte.

**TODO : TRANSITION**

## Comment digitaliser la fonction recrutement d’une entreprise ?

L’ensemble du processus doit être digitalisé : de la phase de lancement jusqu'à la validation du futur collaborateur.

Toutes les étapes peuvent être digitalisées au travers de diverses technologies digitales. Des outils “clé en main” existent déjà, qui ne demandent parfois même pas d’installation sur un poste.

### Les outils pour digitaliser le recrutement

#### Les logiciels de recrutement

“*L’Applicant Tracking System*” (ATS) est un outil relativement indispensable, qui va permettre de digitaliser le recrutement, et de suivre l’ensemble des parcours des candidats qui postulent aux offres proposées.

Les ATS permettent également de prendre du recul sur le recrutement et son processus. Les candidatures sont centralisées, il n’est plus nécessaire pour le recruteur de naviguer entre les divers services logiciels (mail, tableurs, dossiers…).

#### Les tests vidéo en direct ou différé

L’entretien vidéo est de plus en plus utilisé par les recruteurs. Le test vidéo permet de réellement gagner du temps, notamment dans le fait qu’il évite un déplacement, au recruteur comme au candidat.

Avec les tests vidéo *“différés*”, les candidats doivent s’enregistrer face caméra, tandis qu’ils répondent à une suite de questions définie au préalable. Les recruteurs pourront alors ensuite visualiser les vidéos des candidats à loisir. L’une des vocations du test vidéo différé est de remplacer les entretiens téléphoniques.

Le test vidéo différé peut même être automatisé au maximum et venir remplacer l’étape de l’entretien : il permet d’éviter d’avoir à prévoir 45 minutes par candidat, pour poser toujours les mêmes questions

#### Les tests de personnalité et tests de recrutement

Il est également possible de digitaliser l’évaluation des candidats en utilisant des tests de personnalité ou de “recrutement”. S’ils ne sont pas infaillibles, ces tests permettent tout du moins de s’assurer de l’adéquation entre le besoin en recrutement et un aperçu de la personnalité du candidat.

#### Le chatbot

Ce système permet d’automatiser la totalité, ou presque, des échanges entre candidats et recruteurs. Il résout les problèmes de disponibilité : Un candidat veut poser une question alors que le recruteur n’est pas disponible ? Le chatbot peut alors prendre le relai, et assurer la disponibilité du service.

Plusieurs approches sont possibles, comme les systèmes à base de règles - un ensemble de questions/réponses prédéfinies ou la réponse du candidat sera comparée à la réponse préenregistrée - ou des systèmes plus complexes fondés sur des systèmes d’intelligence artificielle.

#### Les outils de collaboration

Ces dernières années, le nombre d’outils de collaborations destinés aux organisations a décuplé. Qu'il s'agisse de systèmes de messagerie simples (Slack, Pidgin ...) ou plus élaborés (Microsoft teams), la plupart de ces outils permettent la création de “fils de discussion[[3]](#footnote-3)” qui permettent d’échanger directement ou indirectement avec d’autres collaborateurs.

Les échanges peuvent se fluidifier, et on verra ainsi diminuer le taux d’erreur.

#### Les outils d’intelligence artificielle

L’intelligence artificielle est l’ensemble des techniques et théories qui s’appliquent à développer des systèmes informatiques capables de simuler, à diverses échelles et divers niveaux, certains traits de l’intelligence humaine.

Avec ces systèmes, reposant sur ce qu’on appelle “l’apprentissage”, il est possible de procéder à des analyses poussées sur chaque candidat. Cela va de la recherche de compétences dans un CV à l’analyse du ton voix et du débit de paroles).

Ces outils apportent une aide qu’il ne faut pas sous-estimer aux recruteurs, principalement dans la présélection des candidats.

### Recruter avec l’intelligence artificielle :

Notion très peu utilisée dans le monde des RH et du recrutement il y a quelques années encore, l’intelligence artificielle offre de nombreux avantages à celui qui souhaite se lancer dans l’aventure, comme l’automatisation de certaines tâches (notamment le sourcing[[4]](#footnote-4) ou le traitement des candidatures)

Elle permet également dans certains cas de mieux identifier les "talents" qui pourraient correspondre avec le poste à pouvoir, ou même l’entreprise en général.

#### L’IA, c’est quoi ?

L’intelligence artificielle est l’ensemble des techniques et théories qui s’appliquent à développer des systèmes informatiques capables de simuler, à diverses échelles et divers niveaux, certains traits de l’intelligence humaine.

#### Pourquoi l’IA dans le recrutement ?

Il s 'agit dans un premier temps d’une tendance. Les progrès spectaculaires dans le domaine de l’apprentissage automatique ont propulsé l’Intelligence Artificielle sur le devant de la scène ces dernières années, et le monde des RH n’a pas fait exception.

De nombreux recruteurs ont été étonnés et séduits par des exemples de recrutements entièrement automatisés grâce à l’intelligence artificielle.

Cette dernière permet d’effectuer des analyses poussées sur les candidats (débit de parole, analyse des mots utilisés …).

Il s’agit avant tout d’un outil d’aide à la décision, permettant de gagner un temps considérable dans le tri, l’évaluation et la présélection des candidats.

#### Ses avantages

##### Faciliter le sourcing

L’Intelligence Artificielle va principalement faciliter le “sourcing”.

L’identification et la présélection des CVs pertinents est une tâche fastidieuse dans une certaine mesure.

L’intelligence artificielle permet au recruteur de gagner un temps précieux, en utilisant par exemple un système de sourçage “autonome” : il se “nourrira” automatiquement des nouveaux CVs et la classifier automatiquement dans des catégories correspondant aux offres d’emplois ou de poste qui pourraient leur correspondre, ou mettra en avant les “meilleurs talents” en fonction de données passées en entrée.

##### L’identification des talents

Identifier les candidats (les “talents”) qui correspondent le plus aux postes à pourvoir est une mission délicate qui peut se montrer difficile.

Automatiser le tri de l’ensemble des candidatures reçues permet au recruteur d’augmenter ses chances de trouver le candidat idéal.

##### Analyse poussée

Les Big Data ont initié un changement de paradigme en matière de quantité de données. L’intelligence artificielle, de par son lien étroit avec les Big Data et sa capacité à extraire des corrélations à partir de ces structures de données “impraticables” pour des méthodes analytiques traditionnelles, permet des analyses plus “poussées” que les analyses traditionnelles.

L’analyse de l’intonation de la voix, du débit de parole ou des *curriculums vitae* des candidats forment une somme d’informations dont le recruteur pourra tenir compte pour que sa décision soit plus efficace.

##### La neutralité et la distance émotionnelle

L’utilisation de l’intelligence artificielle permet de s’affranchir des préjugés desquels peuvent être sujets les recruteurs. Un recrutement automatisé par une intelligence artificielle sera donc en théorie plus objectif.

Mais les systèmes d’apprentissage automatique peuvent eux aussi être victimes de biais, ne serait-ce que du choix des données en alimentation à l’IA, qui pourrait fausser l’objectivité du système.

Plus on accordera de responsabilités à une IA, plus le calibrage des outils devra être précis et rigoureux.

#### Ses inconvénients

Ces avantages trouvent leurs symétries dans les inconvénients suivants :

##### Une certaine déshumanisation

De par son essence, l’humain restera toujours au centre du recrutement. L’expertise humaine est nécessaire à ce processus.

Cependant une partie des actifs, notamment les plus jeunes, est demandeuse d’une continuité entre les outils de leur quotidien et l'agilité du cœur de l’entreprise. L’intelligence artificielle, de par sa flexibilité et son adaptabilité, peut permettre d’automatiser et de digitaliser au maximum le processus de recrutement pour répondre à cette tendance. Or, il faut veiller à ce que, bien qu’il se digitalise, le processus de recrutement ne se déshumanise pas.

Digitaliser au point de déshumaniser serait prendre le risque de passer à côté du cœur même du recrutement : l’humain.

La “rencontre”, à l’avantage de parfois venir heurter notre zone de confort et nous permettre de nous dépasser. Certains profils apparaîtront parfois comme inintéressant ou instables pour un système d’intelligence artificielle, alors qu’ils pourraient révéler une personnalité qui ne ferait qu’enrichir, dans tous les sens du terme, l’entreprise.

##### Les biais

L’entraînement du modèle d’intelligence artificielle, le bloc logiciel qui va produire des prédictions du type “quel candidat correspond le mieux à quelle fiche de poste”, est une étape cruciale et doit être effectuée avec beaucoup de rigueur.

Si l’échantillon fourni en guise de données d’apprentissage n’est pas, ou pas assez aléatoire, on risque de voir se dégager des propositions biaisées.

Ce qui est arrivé à [ENTREPRISE A RETROUVER] en est l’exemple même : dans le but de digitaliser leur processus de recrutement, cette entreprise est passé par un analyseur de CV qui devait proposer les meilleurs candidats. Cet analyseur avait été entrainé avec les CVs des ingénieurs de l’entreprise, qui étaient majoritairement des hommes. Le modèle par son entrainement à généré une règle qui s’apparenterait à “un bon ingénieur est un homme”, et s’est mis à ne proposer que des profils masculins.

Dans une moindre mesure, le risque est aussi de voir l’entreprise acquérir toujours le même type de profil, se privant ainsi d’une diversité qui pourrait être bienvenue.

##### Une présélection trop … Sélective

En se concentrant sur les compétences opérationnelles, techniques ou managériales, l’extraction d’informations à partir d’un CV, par exemple, ne saura que mettre en évidence les “*hard skills*[[5]](#footnote-5)” d’un candidat, et fera donc l’impasse sur ce qu’on appelle les “*soft skills*[[6]](#footnote-6)”. Ce serait passer à côté de candidats qui pourraient potentiellement apporter une véritable plus-value à l’entreprise.

#### Ses utilisations

##### La rédaction d’offres d’emploi

L’une des principales et premières utilisations de l’intelligence artificielle dans le recrutement est la rédaction des offres d’emploi. Grâce aux algorithmes de génération automatique de texte[[7]](#footnote-7), il devient possible de rédiger des offres contenant par exemple les mots clés les plus attirant pour un public visé. Pour aller plus loin, des outils comme l’analyse de sentiments[[8]](#footnote-8) peuvent permettre de déterminer le meilleur ton à adopter pour une offre.

##### La gestion du sourcing

Le sourçage est une autres des activités longues et chronophages des recruteurs. L’intelligence artificielle peut apporter un soutien non négligeable dans cette tâche. Un grand nombre de solutions logicielles spécialisées dans le recrutement permettent d’automatiser au maximum le sourçage, et donnent la possibilité aux recruteurs d’identifier les profils les plus pertinents d’internet, ou d’une base de données de plusieurs milliers de CVs.

##### La préqualification et le traitement des candidatures

Avec des modèles correctement entrainé (il faut savoir éviter les différents écueils de l’apprentissage automatique), il devient possible de donner en entrée à un modèle des données concernant la société ou les postes à pourvoir pour qu’il en ressorte, après analyse de l’ensemble des candidatures, une liste des candidats censés être les plus adaptés au besoin en recrutement correspondant.

**TODO : TRANSITION**

# L’extraction d’information et le Traitement Automatique du Langage Naturel

La recherche et l’extraction d’information dans des CV, aussi appelé « parsing », permet l’extraction d’informations pertinentes à partir de CV possédant une forme semi structurée. Il s’agit d’une sous discipline du Natural Language Processing (NLP, en français « Traitement Automatique du Langage », TAL). Le TAL est une technique couramment utilisée dans le domaine de l’extraction d’informations.

## L’extraction d’information

### Définition

L’extraction d’informations désigne la recherche, dans un texte ou un ensemble de textes (corpus) automatisées de données informatives en liens avec un sujet donné et plus ou moins précis.

Ces outils (d’extraction d’information) donnent la possibilité d’extraire des informations à partir de documents textuels, de bases de données, de sites web…

Ces sources d’informations peuvent être non structurées[[9]](#footnote-9), semi structurées[[10]](#footnote-10) ou structurées[[11]](#footnote-11).

Pour donner un exemple, un document de traitement de texte écrit avec Word est de manière générale considéré comme un ensemble de données non structurées : pragmatiquement, il s’agit d’une suite de caractères plus ou moins formatée et plus ou moins longue. Ces données ne sont pas à proprement parler « structurées ». On peut néanmoins leur associer une certaine somme d’informations qu’on appelle « métadonnées [[12]](#footnote-12)».

Dans les faits, les délimitations de ces trois catégories ( structurées, semi structurées ou non structurées) ne sont pas clairement définies. On parlera, pour les considérer dans leur ensemble, de « *continuum des données* »[3].

Les techniques de l’extraction d’information sont principalement utilisées dans le cadre du Traitement Automatique du Langage Naturel, dont leur but est d’extraire du texte structuré à partir d’un texte qui ne l’est pas.

### Historique

L’extraction d’information est une discipline qui a vu le jour dans les années 1970 avec l’apparition du Traitement Automatique du Langage Naturel. L’un des premiers systèmes commercialisés était « JASPER »[4], développé dans le but de fournir des informations financières en temps réel aux trader financiers.

Quelques-unes des problématiques actuelles de ce domaine dont l’annotation automatique ou la reconnaissance et l’extraction de contenu depuis des fichiers d’images ou de vidéos.

Le très grand nombre de règles linguistiques issues de la complexité du langage est la raison principale pour laquelle le développement de systèmes d’extraction d’information de haute qualité se révèle être un véritable enjeu, notamment pour les systèmes d’Intelligence Artificielle.

## Le Traitement Automatique du Langage Naturel

### Définition

Le Traitement Automatique du Langage (Naturel) ou TAL, est un domaine de recherche qui travaille au développement de systèmes permettant la compréhension, la manipulation et la génération du langage dit « naturel », par des machines.

Son importance réside dans le fait qu’il est l’interfaces entre la science informatique et la linguistique, et traite donc de la capacité des systèmes informatiques à interagir avec l’humain, plus ou moins directement.

Le terme TAL est en fait un terme générique qui va regrouper sous un même nom divers champs applications : traduction automatique, analyse de sentiment (également appelée « Opinion Mining »), analyse de comportement, agent conversationnel, et autres domaines d’applications, comme la classification de texte ou le résumé automatique.

Tout problème de TAL possède au moins deux aspects:

* L’approche « linguistique », consiste en la transformation après prétraitement, des données d’entrées en données exploitables.
* Suit le traitement par apprentissage automatique, ou l’on va appliquer des modèles de Machine Learning ou de Deep Learning au jeu de données obtenu à l’étape précédente.

C’est dans ces derniers domaines, la classification de texte et le résumé automatique, que viennent s’inscrire les analyseurs automatiques de CV.

### Historique

Pour saisir les enjeux de ce domaine de recherche, il me paraît important d’en dresser un bref historique.

#### L’approche « symbolique »

Les principes du TAL symbolique peuvent être résumés ainsi : étant donné une somme de règles -linguistiques- l’ordinateur tend à simuler une forme de compréhension du langage naturel, en appliquant ces règles aux données auxquelles il est confronté.

##### Les années 1950-1960

Les premiers travaux dans le domaine du traitement automatique du langage naturel débutent dans les années 50, dans un contexte politique de guerre froide qui favorise le travail de recherche sur la traduction automatique.

Les premiers éléments développés ont trait au traitement automatique des conversations. Le test de Turing, exposé en 1950 dans l’article « *Computing machinery and intelligence* »[5] par le mathématicien du même nom, propose une méthode d’évaluation du degré « d’intelligence » d’une machine. Cette mesure est effectuée à partir de la capacité d’un programme à se faire passer pour un être humain dans le contexte d’une conversation : lors d’un échange de messages écrits, un individu humain doit évaluer son interlocuteur et décider de s’il s’agit d’un être humain ou d’un ordinateur.

(Les progrès dans des sciences comme la psychologie ont cependant démontré que cette base de test est relativement fragile pour évaluer une intelligence « artificielle », notamment dans le fait que l’impression d’un utilisateur unique dépende de bien trop de facteurs pour être objective. )

Les travaux de Turing sont le point de départ de nombre d’expériences, qui ont tantôt pour but la traduction de phrases (expérience « *Georgetown-IBM* »[6]), tantôt la mise au point d’automates conversationnels (« *ELIZA* »[7] de Joseph Weizenbaum).

En 1957, le travail de Noam Chomsky sur les structures syntaxiques va révolutionner » la linguistique avec la « grammaire universelle »[8], un système de structures syntaxiques fondé sur des règles.

Vers la fin des années 1960, un chercheur du MIT du nom de Terry Winograd parvient à programmer « *SHRDLU* »[9], un programme qui donne la possibilité à son utilisateur de dialoguer avec la machine pour manipuler des cubes affichés sur un écran.

Le fait est notable car il s’agit du premier programme qui comprenne et exécute des ordres complexes en langage naturel. Cependant les opérations qu’il était capable de réaliser étaient limités.

De manière générale, les progrès sont minimes et décevant : le monde la recherche prend conscience de l’extrême complexité des langues.

##### Les années 1970-1980

De nombreux chercheurs de cette période ont débuté l’écriture « d’ontologies conceptuelle ». Leur but était de définir une ou des structures d’information pour la (l’information) rendre intelligible par la machine. Roger Schank, par exemple, est un linguiste et mathématicien qui travaille travailla à l’université de Stanford. Il cherche alors à construire un modèle[[13]](#footnote-13) du langage indépendant de la syntaxe. Il propose dans cette optique de donner aux phrases une représentation par des primitive conceptuelles (des actions, des rôles), le but étant de donner un type aux situations et aux entités pour diminuer les ambiguïtés de la phrase.

#### L’approche « statistique »

A la fin des années 1980, la plupart des systèmes de traitement du langage naturel sont fondés sur de complexes ensembles de règles composées manuellement. Une véritable révolution va cependant vernir heurter le monde du TAL avec les premières utilisations d’algorithmes d’apprentissage automatique[[14]](#footnote-14) pour ce type de traitement.

On peut noter deux causes majeures de ce progrès :

* L’augmentation de la puissance de calcul des machines ;
* La diminution progressive de la domination des théories issues du travail de Noam Chomsky sur la linguistique, dont la théorie décourageait de l’utilisation d’approches « Apprentissage Automatique ».

##### Les années 1990

Dans les années 1990, c’est essentiellement le travail et les recherches d’IBM qui va permettre de réels succès via les méthodes statistiques, essentiellement dans la traduction automatique. Ces systèmes vont tirer parti de la présence de nombreux corpus de textes multilingues, comme ceux produit par le parlement du Canada et l’Union Européenne à la suite de lois demandant la traduction des procédures gouvernementales dans toutes les langues officielles des systèmes de gouvernement concernés.

Cependant, ces systèmes étaient limités aux « métiers » des corpus desquels ils étaient dépendants, ce qui représenta une limitation majeure dans le succès de ces systèmes. La recherche s’est donc orientée sur le développement de systèmes et des méthodes permettant d’apprendre plus efficacement à partir de quantités de données plus limitées.

##### Les années 2000

Avec la démocratisation du web, des quantités de plus en plus importantes de données brutes (non étiquetées) sont disponibles depuis le milieu des années 1990. De ce fait, la recherche s’est majoritairement concentrée sur le développement d’algorithmes d’apprentissage non-supervisé[[15]](#footnote-15) ou semi-supervisé[[16]](#footnote-16) : de tels algorithmes peuvent « *apprendre* » à partir de flux de données qui n’ont pas été manuellement étiquetées.

Cette approche donnait en général de moins bons résultats en matière de « précision [[17]](#footnote-17)» du fait des tailles relativement réduites des structures de données disponibles, mais l’avènement des Big Data[[18]](#footnote-18) (l’intégralité du World Wide Web peut par exemple être analysée) a réellement amorcé un changement de paradigme en matière d’apprentissage et de programmation, et le problème s’est déporté sur la recherche d’une complexité temporelle assez petite pour que l’utilisation de tels algorithmes soit envisageable.

##### Aujourd’hui

Les années 2010 ont vu augmenter considérablement la popularité des méthodes d’apprentissage automatique de type réseau neuronal et, par conséquent, d’apprentissage dit « profond ». Ces méthodes se sont très naturellement généralisées dans le domaine du traitement du langage naturel.

En effet, les résultats montrent que des telles techniques atteignent de très bons niveaux de performance, pour de nombreuses tâches du TAL : modélisation du langage, parsing[[19]](#footnote-19), etc.

### Un projet de type « Data Science »

#### La nature du projet

Un projet de parsing de CV, en tant que projet de TAL, s’inscrit dans la lignée des projets dits de « Data Science » et requiert une méthodologie adaptée. En effet, malgré l’augmentation de la puissance de calcul et l’amélioration de l’accès aux données, la capacité d’utilisation des données dans le processus de décision est soit perdue, soit non maximisée. En effet, tout comme les scientifiques « traditionnels », les scientifiques des données ont besoin d’une méthodologie fondamentale qui leur servira de stratégie dans la résolution des problèmes.

Cette méthodologie doit être indépendant des technologies ou des outils, et doit fournir un cadre pour l’utilisation des méthodes et des processus qui seront utilisé dans le but d’obtenir réponses et résultats.

#### Méthode CRISP

La société IBM a, dans les années 60, proposé la méthode CRISP. Elle se décompose en 6 étapes :

##### Compréhension du problème métier

Ou comprendre les éléments métiers que le projet essaie de résoudre, ou améliorer, au travers de la science des données.

##### Compréhension des données

Cette étape a pour but de déterminer les données à analyser le plus précisément possible. Durant cette phase, la qualité des données disponibles est mesurée et identifiée. C'est aussi à ce moment-là que le lien entre les données et leur signification métier

##### Préparation des données

Cette phase va voir la transformation des données brutes en un ensemble structuré de données à analyser.

Les données sont classées en fonction de critères choisis, nettoyées, et normalisées pour les rendre entièrement compatibles avec les algorithmes composant la suite du traitement.\\\\

##### Modélisation

C'est le « cœur » de la science des données. Cette phase comprend le processus de prise de décisions concernant les algorithmes qui constitueront le modèle.

Cette phase se déroule en trois temps :

A. D’abord intervient la « description » de ce qu’il s’est produit, dans le but de comprendre pourquoi tel ou tel phénomène est survenu.

B. Ensuite intervient une phase de « prédiction » qui va chercher à expliquer ce qu’il va se passer.

C. Enfin, une phase de « prescription », qui va rechercher l’optimisation d’une situation future.

##### 6. Evaluation

Il s’agit de comparer les modèles ou les connaissances obtenus avec les objectifs décidés au début du processus. C’est dans cette phase qu’est prise, ou non, la décision de déployer un modèle, et si besoin, l’améliorer.

##### 7. Déploiement

C’est la mise en production. Dans cette étape finale, les utilisateurs finaux utilisent les modèles construits précédemment. La connaissance générée par la modélisation est rendue accessible, et intégrée au processus de décision.

#### Méthodologie Data Science

John Rollins, scientifique des données chez IBM, propose dans un article intitulé *« Why we need a methodology for data science ?* »[11] propose une approche pour une stratégie directrice qui s’assurerait que les données utilisée dans la résolution du problème soient pertinentes, et fournissent un cadre adéquat à l’application des méthodes et processus utilisés pour la résolution du problème.

Cette méthodologie se décompose en 10 étapes, composantes d’un processus itératif menant de la conception au déploiement de la solution au retour d’expérience et au « raffinement ».

On y retrouve la compréhension du problème métier, la compréhension des données, la préparation des données, la modélisation, l’évaluation, et le déploiement.

Mais à ces 7 étapes viennent s’ajouter :

##### L’approche analytique

Une fois le problème métier clairement énoncé, il faut définir une approche analytique visant à résoudre le problème. Il s’agit de replacer le problème dans le contexte des techniques statistiques et d’apprentissage automatique, pour identifier le type de modèle qui apportera une réponse le plus efficacement.

##### Exigences des données

Une étape intermédiaire qui va permettre d’identifier de manière précise le contenu, les formats et les sources de données qui seront ensuite nécessaire à la collecte des données.

##### Collection des données

Ici intervient l’identification des ressources de données (structurées, semi-structurées et non structurées) pertinentes et disponibles pour la résolution du problème. Les sources de données sont fusionnées, et les informations redondantes sont supprimées.

##### Feedback

Une fois le déploiement effectué, les utilisateurs apportent un retour sur les performances du modèle. Son impact dans l’environnement de mise en œuvre est mesuré, et ces mesures servent à affiner le modèle, en augmentant sa précision.

**TODO : TRANSITION**

## L’analyse automatique de CV

### Définition

Un analyseur automatique de CV (ou « Parser ») est un bloc logiciel qui va prendre un CV en entrée, dans n’importe quel format (PDF, Word ou image) et le convertir en un format structuré de données comme le XML ou le JSON. Il sera ensuite potentiellement associé à d’autres blocs logiciels lui permettant de recevoir en entrée des structures de données contenant plusieurs CV.

La recherche porte une partie de ses efforts à développer un « Parser » qui va pouvoir extraire, stocker et analyser automatiquement les CVs ou les données des candidats de manière à ce que l’information produite soit ensuite prête à être catégorisée et étiquetée, triée ou encore parcourue lors de recherche.

Il faut cependant accepter qu’aucun analyseur ne peut donner de résultat exempt d’erreur : pour la plupart des tâches que réalise ce type de système, atteindre une qualité de 90% est un très bon résultat.

L’un des intérêts notables à l’obtention de telles données est la possibilité ensuite de s’en servir pour entraîner des modèles d’apprentissage automatique, ce qui agrandit nettement le panel des possibilités de traitement.

Directement ou indirectement, le parsing de cv permettrait ensuite d’accélérer tant le processus de candidature du candidat que traitement des données opéré par le recruteur :

L’utilisation des outils de gestion des candidatures crée le besoin d’harmoniser les données des candidats. Ce besoin amène les plateformes de recrutement à des systèmes où il faut en moyenne 30 clics pour postuler à une offre, ce à quoi s’ajoute, deux fois sur trois, la création d’un compte.[12]

A une heure ou les plateformes de réseaux sociaux permettent de visualiser parfois plusieurs dizaines de contenu avec moitié moins de clics, cela fait du processus de candidature une étape « longue » et « dissuasive ».

Le parsing de CV permet de simplifier au minimum cette tâche. Puisque les informations nécessaires à l’harmonisation des données sont extraites directement depuis les CVs, le processus peut s’affranchir de plusieurs étapes d’intégration d’informations par le candidat.

### Un analyseur générique, ou spécifique ?

Plusieurs approches sont possibles pour celui qui voudrait réaliser des analyses de documents : il est possible de passer par un analyseur générique, disponible prêt à l’emploi ; d’en créer un de toute pièces et sur mesure ou encore de combiner des analyseurs existants.

#### Un analyseur générique

Dans le cas où l’on souhaiterait pouvoir analyser une grande variété de textes, il faudra opter pour un analyseur générique. Le nombre croissant de corpus, et leurs tailles conséquentes permet de créer et « d’entrainer » des analyseurs génériques proposant une couverture et une précision jugée « acceptable », du moins pour certaines tâches.

L’intérêt de ce type d’analyseur est de pouvoir fonctionner sans avoir été paramétré. Cependant, ils n’analyseront correctement que les documents proches ou similaires à ceux avec lesquels ils ont été entrainés.

#### Un analyseur spécifique

Au contraire, si le domaine d’application du parser est défini et limité, si l’on connaît d’avance la nature du corpus à traiter, alors on peut choisir de privilégier la qualité de traitement en optant pour un analyseur spécifique.

Ce dernier va savoir tenir compte de chaque spécificité du corpus, de sa phraséologie[[20]](#footnote-20) et aux différents concepts qui pourraient y apparaître.

L’intelligence artificielle, avec l’apprentissage automatique, offre de nouveaux outils pour développer des systèmes d’analyse adapté, pour un cout relativement raisonnable et dans des délais réalistes.

L’apprentissage profond permet par exemple, en phase de début du projet, d’analyser le corpus pour en faire émerger des concepts pertinents. Obtenir ce type d’informations à cette étape du projet permet d’établir assez tôt une « taxonomie [[21]](#footnote-21)» qui va donner la possibilité d’amorcer déjà les travaux.

Durant le cycle de vie du projet, ce même apprentissage automatique mettra en évidence de nouveaux concepts à mesure qu’ils apparaîtront. Cette « veille » donnera lieu à une mise à jour de la taxonomie parallèle aux évolutions du corpus analysé. De plus, l’étude des signaux faibles[[22]](#footnote-22) offre la possibilité de visualiser et d’appréhender les phénomènes susceptibles d’émerger.

Pour terminer, il faut souligner que l’une des forces de l’apprentissage profond est que du moment qu’un corpus de grande taille dans une langue est disponible (exemple : Wikipedia), il devient possible d’entrainer des modèles non supervisés pour leur apprendre des modèles de langage automatiquement.

### Deux mesures clés

#### La couverture

Autrement dit, décrire ce qu’un analyseur tente d’extraire.

Tous les analyseurs essaient d’extraire les informations de contact des candidats. La plupart extraient aussi les compétences, l’historique des expériences professionnelles et les diplômes et qualifications.

Les analyseurs les plus avancés peuvent extraire le « résumé » du candidat, sa nationalité, son salaire souhaité, le statut de son visa, ses hobbies … Toutes ces informations sont nécessaires pour créer un enregistrement le plus complet possible destiné à être utilisé par le logiciel de gestion des candidats, donc en général, plus un analyseur extrait d’informations, mieux il fonctionne.

#### La précision

Autrement dit, décrire à quel point un analyseur est bon dans l’identification des informations dans un CV.

La précision rend compte de la fréquence à laquelle un analyseur a effectivement « raison ». Par exemple, une précision de 95% sur l’identification des noms signifie que l’analyseur de CV extrait correctement les noms des candidats à un pourcentage de 95 % pour tous les CV insérés.

En général, si un analyseur est en dessous d’une précision de 90%, le nombre d’erreurs sera trop large pour permettre de charger des données dans la base de données d’un logiciel de gestion des candidats sans demander une intervention ou une supervision humaine appuyée.

### Problématiques

#### La grande diversité des CV

Si nous prenons plusieurs CV au hasard de quelques recherches internet, l’une des premières choses que nous pouvons constater que le « langage » des CVs est une variante du langage naturel.

Cette variante est « semi-structurée », dans une forme ou les informations sont agglomérées en blocs. Chacun de ces blocs contient des informations relatives aux contacts d’une personnes, à son cursus de formation ou encore ses expériences professionnelles.[14]

On peut observer certaines récurrences, comme la présence du nom, du prénom, d’un bloc d’informations de contact et de blocs de rubriques (« Formation », « Expérience » …).

Cependant, malgré le domaine relativement restreint de ce processus et cette forme semi-structurée du langage des CVs, la tâche d’analyse automatique de ces derniers n’en demeure pas moins complexe (figure 1).

Les CVs diffèrent dans les types et dans l’ordre d’apparition des informations. Ils contiennent tant des phrases complètes que des abréviations ou de simples suites de mots, organisés dans des colonnes réparties au bon vouloir du créateur du document. Cela modifie le sens de lecture et impacte forcément le traitement et l’analyse.

Un autre élément à prendre en compte est la conversion du format d’origine vers un format exploitable par les algorithmes de l’analyseur.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure - Trois manières d'exprimer l'expérience professionnelle

En effet, la conversion en texte brut amènera à la perte d’une quantité non négligeable d’informations de style et de disposition, tandis que la conversion en un langage de balisage tel que le .xml ou le .html peut produire des résultats inattendus au niveau de la disposition.

Pour répondre correctement aux exigences métiers tout en obtenant des résultats acceptables pour les mesures clés, le système d’un analyseur doit être indépendant de l’ordre et de la disposition des informations dans un document.

Enfin, une dernière difficulté vient des choix graphiques des candidats. Un certain nombre de cv contiennent des images ou des logos pour représenter des entités, ou avec des systèmes de barres de progression et de symboles plus ou moins remplis comme jauges de la maîtrise de certaines compétences.

Si un humain pourra identifier sans trop de difficultés le sens de ces images, un système informatique, lui, se verra vite démuni.

### Quatre grands types d’approches

On peut noter quatre grandes approches dans l’analyse de CV pour l’extraction d’informations :

#### L’analyse fondée sur des entités nommées

Cette méthode tente d’identifier certains mots, certaines phrases ou schémas dans le texte, habituellement en utilisant des expressions régulières ou des dictionnaires de mots clés. Cette approche est généralement abordée en tant que seconde étape, après l’analyse lexicale d’un document.

L’analyse fondée sur des règles de grammaire

Le choix de cette méthode inclue la mise en place et le maintien d’un ensemble de règles de grammaires. Ces règles décrivent, pour un langage formel, ce qui définit une phrase valide. Elles seront ensuite appliquées aux documents dans le but d’en extraire de l’information.

#### L’analyse fondée sur les statistiques

Cette approche consiste en l’application de modèles numériques sur les texte « vectorisé » pour y identifier des structures sous-jacentes.

Les méthodes reposant sur des statistiques ont en commun le fait d’associer une probabilité à une règle de grammaire. Cette association permet par la suite d’obtenir les fréquences relatives de chaque règle de grammaire. A partir de ce concept, les parsers « statistiques » peuvent rechercher, en se fondant sur les analyses de tous les textes, l’analyse la plus probable pour une phrase.

L’analyse fondée sur l’apprentissage

Les méthodes utilisant un système d’apprentissage emploient généralement des algorithmes de classification pour extraire les informations d’un document.

Choisir un système ou un autre n’est pas une décision bloquante, dans le sens où de nombreux systèmes existant d’extraction d’information à partir de CV utilisent une approche hybride, combinant plusieurs méthodes différentes.

### Différentes étapes successives

Durant la phase de collection des données, il faudra extraire le contenu textuel de l’ensemble des CV.

Le format d’entrée est en général le format .PDF. Cependant, un certain pourcentage des fichiers de CV se trouve être des images, dans des formats comme le .jpg ou le .png. Il faut d’ores et déjà faire un choix dans la manière d’aborder cette problématique : ignorer ces « outliers » ou choisir de les traiter autrement avec un système différent.

Une fois cette décision prise intervient la phase de préparation et de nettoyage des données (vérification des données manquantes, des valeurs invalides, ou suppression des doublons.).

Dans le cadre précis de l’analyse de CV, le traitement des données textuelles passe généralement par les étapes suivantes :

#### « *Tokenisation* »

Il va s’agir ici de découper le texte brut du CV en « jetons » qui représenteront les mots ou les phrases contenus dans le texte (verbe, nom, connecteurs linguistiques, etc.)

##### Exemple :

La phrase :

« *J’ai 15 ans d’expérience dans le domaine du développement web.* »

Devient la liste suivante :

[« *j’* », « *ai* », « *15* », « *ans* », « *d’* », « *expérience* », « *dans* »,

« *le* », « *domaine* », « *du* », « *développement* », « *web* »]

#### Elimination des « *stop words* »

Cette étape a pour but de réduire le « bruit », en éliminant les jetons non significatifs, vides de sens (notamment les pronoms et certains connecteurs l’linguistiques.). Il est à noter que n’importe quel groupe de mots peut être choisi comme liste de « *stopwords* » pour un sujet donné.

L’une des problématiques en lien avec les *stopwords* est que leur élimination peut causer des problèmes lors de la recherche de phrases qui les incluent, particulièrement pour les noms propres contenant des pronoms.

##### Exemple :

La liste :

[« *j’* », « *ai* », « *15* », « *ans* », « *d’* », « *expérience* », « *dans* », « *le* », « *domaine* », « *du* », « *développement* », « *web* »]

Est filtrée et réduite à celle suivante :

[« *ai* », « *15* », « *ans* », « *expérience* », « domaine », « développement », « web »]

#### Le « Stemming » ou « racinisation » et la « lemmatisation

##### Le Stemming

Le « stemming » consiste en la réduction d’un mot à sa forme « racinaire », et en la production de variantes morphologiques fondées sur cette base.

Le but est de regrouper les variantes d’un mot sous un seul et même terme, pour élargir le champ des recherches.

###### *Exemple*:

|  |  |
| --- | --- |
| **Racine** : | **Variantes** : |
| «*continu* » | « *continu »*  « *continua »*  *« continuait »*  *« continuation »*  *« continue »* |

Tableau - Une racine et ses variantes

Le stemming est souvent considéré comme une méthode assez rudimentaire pour regrouper les mots apparentés. Un « *stem* », une racine obtenue par *stemming*, peut par exemple ne correspondre à aucun mot existant.

Les algorithmes sont généralement simples et consistent en la réduction lettre par la lettre en commençant par la fin du mot. Cette stratégie donne de bons résultats en anglais comme en français, mais les nombreuses exceptions présentes dans ces deux langues nécessitent souvent un processus plus sophistiqué.

Ainsi de nombreux systèmes de TAL préfèrent s’appuyer sur la lemmatisation

##### La Lemmatisation

Le terme « lemmatisation » désigne un traitement appliqué à un texte en prévision de son analyse, qui va permettre d’identifier les « lemmes » des « lexèmes » sujets à « flexion ».

Contrairement au *stemming*, la lemmatisation va chercher plus loin que la simple réduction d’un mot : elle considère le vocabulaire entier d’un langage pour appliquer une analyse morphologique à un mot en considérant également le texte qui « entoure » le mot.

###### Exemple :

|  |  |
| --- | --- |
| **Lemme :** | **Lexèmes :** |
| « *être* » | « *suis* »  « *es* »  « *est* »  « *sommes* »  « *êtes* »  « *sont* » |

Tableau - Un lemme et ses lexèmes

La lemmatisation donne généralement de meilleurs résultats lorsqu’elle est appliquée à un texte en français, car elle se base sur une analyse morphologique et lexicale des mots, alors que le *stemming* ne fera que les tronquer.

##### Conclusion

Du fait du besoin beaucoup plus important d’une connaissance de la structure du langage analysé, il est plus difficile de créer un bon algorithme de lemmatisation qu’un algorithme de *stemming*.

Le *stemming* présente l’avantage de proposer une suite d’opérations à effectuer par étapes sur le mot, ce qui lui offre un gain de temps. Avec la lemmatisation, il faut analyser le corpus pour fournir le lemme, ce qui prend nécessairement plus de temps.

Si la vitesse d’exécution est une priorité, c’est le *stemming* qu’il faudra privilégier, alors que dans le cas d’une recherche axée sur la qualité des résultats, on préfèrera la lemmatisation.

1. L’adoption stratégique de technologies digitales. [↑](#footnote-ref-1)
2. La marque employeur est l'image d'une entreprise auprès de ses employés et des candidats potentiels, et inclut par extension les efforts de marketing et de communication qui visent à l'améliorer et à la communiquer. [↑](#footnote-ref-2)
3. Fil de discussion : une suite de messages consécutifs sur un même thème, classés de manière arborescente [↑](#footnote-ref-3)
4. Le sourçage, en anglais sourcing, est le fait de trouver des fournisseurs ou des candidats dans les achats, les ressources humaines et les services informatiques (IT). [↑](#footnote-ref-4)
5. Comprend l’ensemble des compétences que l’on a généralement acquises par formation, et que l’on peut démontrer de manière concrète. [↑](#footnote-ref-5)
6. Aptitudes liées au caractère et à la manière d’interagir avec les autres [↑](#footnote-ref-6)
7. La génération automatique de texte (GAT) est une sous discipline de la linguistique computationnelle qui vise à exprimer sous une forme textuelle, syntaxiquement et sémantiquement correcte, une représentation formelle d'un contenu. [↑](#footnote-ref-7)
8. Processus qui permet de déterminer la tonalité émotionnelle qui se cache derrière une série de mots [↑](#footnote-ref-8)
9. Désignation générique qui vient décrire toute donnée qui n’appartient pas à une structure particulière. Les données textuelles sont majoritairement générées par les documents de traitement de texte (.doc, .rtf …), les logiciels de collaboration ou de messagerie instantanée (Teams, Azure DevOps … ) tandis que les données non textuelles correspondent plutôt à des supports tels que les images .jpeg, les fichiers audios .mp3 ou encore les fichiers vidéo. Il faut retenir que le format ne permet pas d’y accéder ni de les traiter « facilement ».

   Les données totalement « non structurées » sont en réalité assez rares, puisque la présence d’une structure peut être parfois observable à une autre échelle.

   En l’absence d’une gestion ou d’une administration définie, le volume généré annuel de données dites « non-structurées » peut se révéler être un véritable coût en termes de stockage.

   A cela s’ajoute que la présence de données non administrées dans des quantités plus ou moins importantes peut parallèlement poser des problèmes de responsabilité (exemple : cas d’une action en justice durant laquelle des informations ne pourraient pas être localisées)

   Extraire des informations de données non structurées se révèle souvent être une tâche difficile. Cela implique que les documents soient numérisés, afin de permettre à un système numérique d’en extraire des concepts en fonction de contextes spécifiques (recherche sémantique).

   [<https://www.lemagit.fr/definition/Donnees-non-structurees>] [↑](#footnote-ref-9)
10. Désignation générique qui désigne des ensembles de données qui ne sont pas organisés en référentiel spécialisé (exemple : en dans une base de données), mais qui comportent cependant des informations associées (exemple : des métadonnées) qui les différencient des données « brutes » et les rendent plus facile à traiter. L’organisation n’est pas « complexe », ce qui rend possible une analyse « poussée ».

    [<https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees>] [↑](#footnote-ref-10)
11. Ce terme désigne le contraire des données non structurées. Elles ont été formatées, et leurs éléments ont été réorganisés en fonction d’une structure leur permettant d’être traitées, organisées et manipulées selon diverses approches, avec diverses technologies.

    <https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees> [↑](#footnote-ref-11)
12. Ce terme signifie « donnée de/à propos de données ». Il désigne une donnée servant à définir ou décrire une autre donnée.

    <https://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e> [↑](#footnote-ref-12)
13. Il s’agit d’un modèle « statistique » qui vient modéliser, pour une langue naturelle, la distribution de séquences de symboles « discrets » (mots, phonèmes, lettres). [↑](#footnote-ref-13)
14. Un champ d’étude de l’intelligence artificielle, fondé sur des approches mathématiques et statistiques pour permettre aux ordinateurs « d’apprendre » à partir de flux de données. [10] [↑](#footnote-ref-14)
15. Dans le domaine informatique et de l'intelligence artificielle, l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées. Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. [↑](#footnote-ref-15)
16. Il a été démontré que l’apprentissage se trouve sensiblement amélioré lorsque l’on utilise conjointement à cette étape des données étiquetées et non étiquetées. [↑](#footnote-ref-16)
17. La précision est calculée en divisant le nombre de vrais positifs par la somme des nombres de vrais positifs et de faux positifs. [↑](#footnote-ref-17)
18. Des structures de données dont les caractéristiques en termes de variété, de volume et de vélocité demandent l’utilisation de techniques et de méthodes particulières pour générer de la valeur. [↑](#footnote-ref-18)
19. Le parsing est l’analyse d’un document textuel dans le but d’en extraire le sens. [↑](#footnote-ref-19)
20. La phraséologie étudie «*l’ensemble des tournures typiques d’une langue, soit par leur fréquence, soit par leur caractère idiomatique*»[13] [↑](#footnote-ref-20)
21. Une organisation structurée des concepts. [↑](#footnote-ref-21)
22. Des signaux « imperceptibles » à l’œil nu. [↑](#footnote-ref-22)