# Parsing de CV

## Définition

L’analyse de CV permet le stockage et l’analyse automatisés de données issues de CV.

## Description

Le Parser de CV analyse un CV, extraie les informations désirées, et sauve les informations dans une base de données avec un identifiant unique pour chaque candidat. Une fois que le CV a été analysé, un recruteur peut effectuer une recherche dans la base de données sur des mots ou des phrases, et obtenir une liste de candidats pertinents. Beaucoup de parser prennent en compte la recherche sémantique, qui ajoute un contexte aux termes recherchés, et essaie de comprendre le sens du texte afin de rendre les résultats plus fiables et complets.

## Machine Learning

Le Machine Learning est d’une importance capitale pour le parsing de CV. Chaque bloc d’information doit se voir affecter une étiquette et être positionné dans une catégorie, qu’il s’agisse du parcours « scolaire », de l’expérience professionnelle ou encore les coordonnées.

Les parsers basés sur des règles utilisent un ensemble de règles prédéfinies pour analyser le texte. Cela ne fonctionne pas pour les CV car le parser à besoin de comprendre le contexte dans lequel les mots interviennent, et les relations qui existent entre eux.

Un parser à base de règles nécessiterait des règles d’une très grande complexité pour tenir compte de toue l’ambiguïté d’un CV, et fournirait une couverture limitée ou incomplète.

Cela conduit à l’utilisation du Machine Learning et particulièrement aux techniques du Traitement Automatique du Langage Naturel (NLP). Le NLP est une branche de l’intelligence artificielle qui utilise le Machine Learning pour comprendre le contenu et le contexte d’un texte, tout en offrant la possibilité de faire des prédictions. Beaucoup des fonctionnalités du NLP sont extrêmement importantes dans le cas de l’analyse de CV :

* La normalisation et le marquage des acronymes prend en compte les différents formats d’acronymes et les normalise ;
* La lemmatisation réduit les mots à leurs racines en utilisant un dictionnaire de langage ;
* Le Stemming va venir retirer les « s », ou les terminaisons verbales.
* L’extraction d’entité utilise des expressions régulières, des dictionnaires, de l’analyse statistique et une extraction complexe basée sur de la reconnaissance de schéma pour identifier les personnes, les endroits, les organisations, les numéros de téléphone, les adresses email et bien plus.

## Efficacité

Les analyseurs de cv atteignent une précision d’environ 87% (article Medium), ce qui fait référence à la précision de la saisie des données et la catégorisation correcte des données. La précision humaine n’est généralement pas supérieure à 96%, de sorte que les analyseurs de CV ont atteint une précision « quasi humaine ».

## Bénéfices

L'analyse des CV permet aux candidats d'être classés en fonction d'informations objectives et peut aider à prévenir les préjugés qui apparaissent si facilement dans le processus de recrutement. Le logiciel peut être programmé pour ignorer et masquer les facteurs qui contribuent aux préjugés tels que le nom, le sexe, la race, l'âge, l'adresse, etc.

Cette technologie peut être extrêmement rentable et permettre d’économiser des ressources. Plutôt que de demander aux candidats d’entrer manuellement l’information, ce qui pourrait les décourager de postuler, la saisie des données se fait automatiquement.

Les informations de contacts, les compétences pertinentes, l’historique de l’expérience professionnelle, le parcours de formation et d’autres informations plus spécifiques à propos des sont a priori plus facilement accessible.

Le processus de sélection des candidats et potentiellement beaucoup plus rapide et efficace. Au lieu d’avoir à regarder en détail chaque CV, les recruteurs peuvent les filtrer sur différentes caractéristiques, les trier et faire des recherches dans la base de données. Cela leur permet de suivre le processus d’entrevue et de pouvoir des postes plus rapidement.

## Défis

Les parsers de CV doivent s’appuyer sur des règles complexes et des algorithmes statistiques pour capturer correctement les informations contenues à l’intérieur des CVs. Il y a de nombreuses variations dans le style d’écriture, le choix des mots, la syntaxe, etc. et le même mot peut avoir de multiples sens ? Une simple date peut être écrite de plus de 100 manières différentes. Cela reste un défi, pour les analyseurs de CV, de tenir compte de toute cette ambiguïté.

Le traitement automatique du langage naturel et l’intelligence artificielle ont encore du chemin à parcourir pour comprendre les informations basées sur le contexte et ce que l’Homme cherche à dire en écrivant.