



M1 MIASHS : Big Data et Fouille de données

Interface d'automatisation du processus de recrutement

Organisme d'accueil
AMA Associates

Auteur
GHOUIBI Ghassen

Encadreur - Organisme d'accueil
Aurélien MICHEL

Encadreur - Université
Jean-Jacques Mariage

Résumé

La révolution digitale qu'on vit actuellement attire un nombre grand d'utilisateurs, qui se traduit un besoin important de spécialiste dans le domaine de l'informatique.

Ces dernières années l'apparition des ESN¹ prend de plus d'en plus d'ampleur pour chercher le candidat idéal qui à permis la naissance à plusieurs plateforme comme LinkedIn en 2002. Le processus de recrutement devient de plus en plus lourd surtout quand il s'agit de sélectionner un candidat parmi plusieurs de même les recruteurs ne peuvent pas faire un tri dans des métadonnées.

Le développement d'une interface qui automatise ce processus présente un véritable défi. En effet les recruteurs passent en moyenne 40 secondes pour lire un CV et 1 minute, 20 secondes pour décider si le candidat sera retenu dans la sélection.

Notre but est créer une interface qui pourrai nous présenter les meilleurs candidat par rapport à une fiche de poste. Les NLP présentent une solution pour résoudre ce problème, ces solutions donnent des résultats plutôt très correcte, en revanche les modèles qui existent actuellement trouve une difficulté à détecter des compétences générales par rapport à des compétences techniques.

En mettant l'accent sur un modèle qui pourrai à la fois détecter les compétences générales intéressantes qui pourrait contribuer à la montée en compétence de candidat, aussi que donnée l'accès à des formations serait un plus pour faire des économies au niveau de l'embauche aussi bien que donnée une chance équitable à chaque candidat.

Dans ce papier, notre but sera de créer une architecture qui pourrait prendre un nombre massive des données qui sera représenter par un banque de CV. Ensuite extraire les informations nécessaire à partir d'une fiche de poste nous allons essayer de coincider cette dernière avec une sélection des meilleurs CV notamment un modèle basée sur les réseaux de neurones ainsi utiliser Word2vec, Text2Vec pour atteindre notre objectif.

Mots-clefs : Word2vec, LSTM, Text2Vec, Naive Bayes, TF-IDF, NLP

1. Une entreprise de services du numérique

Table des matières

1	Introduction	3
2	État de l'art	6
3	Conclusion	7

Chapitre 1

Introduction

La dernière décennie a vu l'émergence d'internet, des nouveaux emplois ont vu le jour grâce à l'automatisation de plusieurs processus. Pour cela plusieurs entreprises essayent de trouver les bons candidats à leur entreprise mais la mission devient très difficile quand on parle des ESN.

L'apparition de plusieurs plateformes qui traitent notre curriculum vitae pour prédire si on est le meilleur candidat pour un poste comme LinkedIn. Notre but c'est concevoir une interface qui permet de correspondre une fiche de poste à notre base de données de candidats.

Les avancées de recherche dans le domaine du **text-mining** ne peuvent que présenter une solution idéale pour ce genre de problème néanmoins, il faudrait d'abord comprendre le processus de recrutement et la sélection d'un candidat par rapport à un autre. En effet la touche humaine ne peut pas être négligée pour choisir un candidat, plusieurs plateformes échouent dans plusieurs essais pour plusieurs raisons comme le format ou l'encodage.

Malgré plusieurs avancées mais jusqu'à l'heure actuelle aucune plateforme n'a réussi à détecter les compétences générales qui veut dire un candidat intéressant mais qui aura besoin d'une petite formation cela présente beaucoup d'avantage pour l'entreprise au point de vue économique.

La conception d'une interface qui permet à la fois de trouver un curriculum vitae idéal en prenant en compte tous les aspects techniques mais aussi les aspects humains et donner une chance à tous les candidats d'une manière équitable. Vu le grand nombre de candidatures sur un poste dans une grande entreprise il sera impérative d'utiliser un algorithme qui permet de filtrer tous les documents reçus et les classées.

Notre dataset qu'on va utiliser dans ce papier serait fourni par Zoho¹, en premier lieu on va travailler sur ce dataset qui semble complet avec des fiches de postes et des curriculum vitae néanmoins ce logiciel présente la fonction-

1. Zoho Office Suite est une suite bureautique en ligne sur le Web contenant du traitement de texte, des feuilles de calcul, des présentations, des bases de données, la prise de notes, des wikis, des conférences Web, la gestion de la relation client, la gestion de projet, la facturation et d'autres applications.

nalités de correspondre une fiche de poste avec un résumé sauf que le but de l'entreprise est d'abandonner ce logiciel voir le remplacer au fils des années par un produit fait maison, aussi bien que dans le cas notre dataset ne sera pas suffisant la vision de faire du scrapping sur LinkedIn n'est pas négliger.

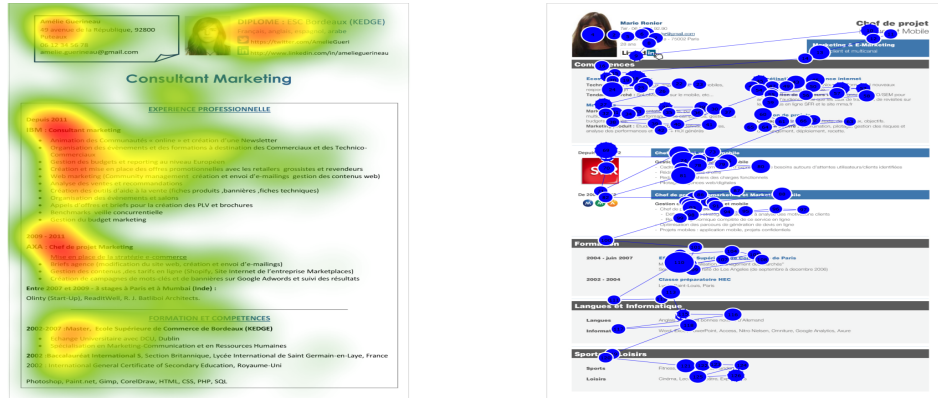


FIGURE 1.1 – Une lecture structurée d'un curriculum vitae (lecture en F)

Dans la figure 1.1, nous remarquons la fameuse lecture en F et cela nous permet en premier lieu de comprendre la structure d'un curriculum vitae. D'où on peut déduire que la lecture d'un résumé se base plutôt sur les mots clés, la plupart du temps cela mène à lire les postes occupés au par avant sans plonger dans les détails, la formation, et les informations du candidat. En effet, c'est tout à fait normal que les recruteurs vont opter pour une lecture en F vu qu'ils voient des centaines de curriculum vitae et ça devient automatique de comprendre rapidement un résumé juste en identifiant les mots clés.

Pour pouvoir identifier les compétences, les langues, les expériences... etc on a besoin tout d'abord de pouvoir trouver des mots clés sachant que le nom et prénom sont aussi des mots clés dans la figure 1.2 ci-dessous on résume le processus qu'on va adapter pour pouvoir comprendre et analyser un curriculum vitae.

Notre approche vise plutôt les documents *PDF* il faudrait tout d'abord pouvoir analyser tout type de documents qui nous donnera encore plus de donner pour tester notre modèle comme *Doc*, *Docx*, *HTML* ..etc

Ensuite classifier le texte à partir de l'identification des mots clés dans ce dernier et produire des données structurées comme le montre la figure 1.2.

Du même principe, on analyse la fiche de poste notre algorithme va essayer de faire correspondre les mots clés présents dans les deux documents. Du texte ICI

Du texte ICI

Du texte ICI

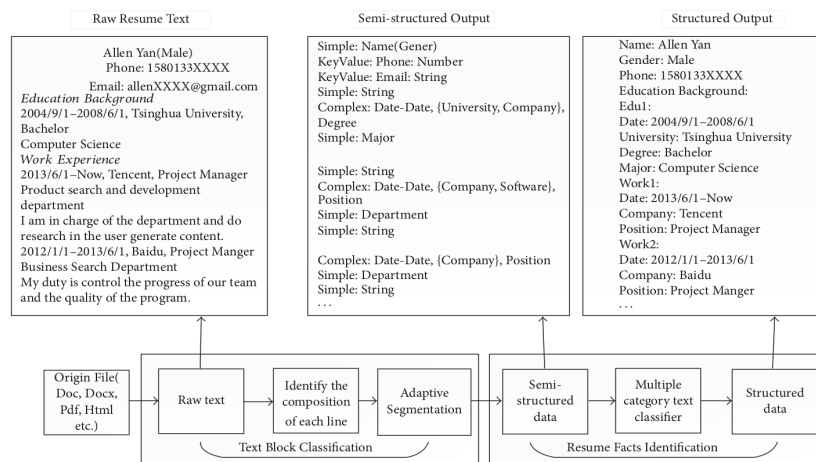


FIGURE 1: The pipeline of our algorithm and an example.

FIGURE 1.2 – U

[illegible]

Chapitre 2

État de l'art

Les avancées sur les processus ne cessent que d'améliorer et cela est liée au recherches réaliser dans le domaine de l'intelligence artificiel. On remarque plusieurs entreprises emplois massivement des interfaces qui permettent d'extraire des données à partir d'un curriculum vitæ.

Ce sujet est d'actualités dans plusieurs communauté comme on constater dans l'article suivant *A Two-Step Resume Information Extraction Algorithm* [1]

Chapitre 3

Conclusion

Bibliographie

- [1] George Greenwade. The Comprehensive Tex Archive Network (CTAN).
TUGBoat, 14(3) :342–351, 1993.