**UNIVERSITE MAPON**

Faculté Polytechnique

Département de Génie Informatique



SUJET :

**« Extraction des compétences par l’approche du traitement du langage naturel pour l’adéquation des offres d’emploi »**

Travail présenté et défendu en vue de l’obtention du

diplôme d’ingénieur bachelier en Génie Informatique

**Présenté par** : MUKOMBOZI ZERO-FARAY Nickson

**Promotion** : Troisième Bachelier

**Présenté par** : MUKOMBOZI ZERO-FARAY Nickson

**Promotion** : Troisième Bachelier

**AVRIL** 2024

**Annexes**

1. **Description de la fonction de traduction des CV**

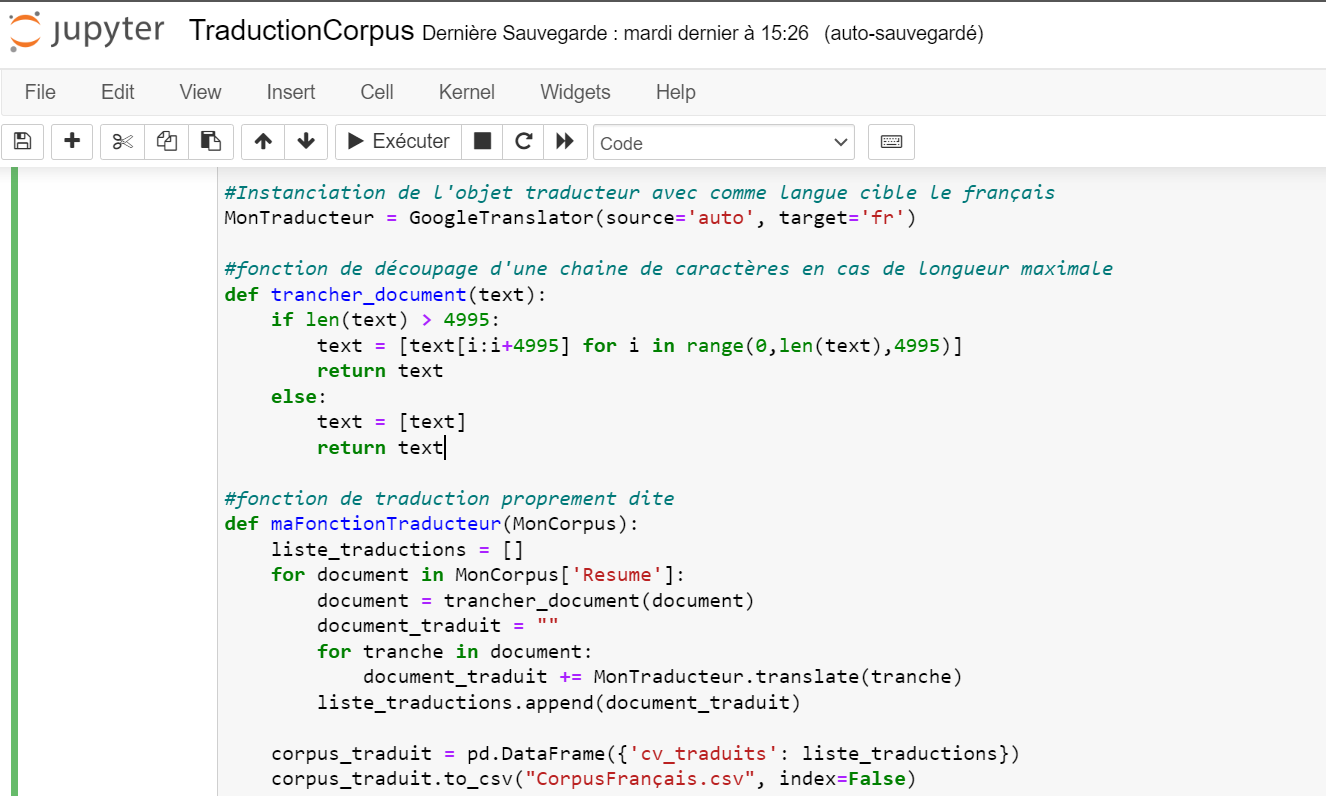


Figure 42: script python de traduction des corpus

*Le script commence par importer les bibliothèques nécessaires, notamment deep\_translator pour accéder à l'API GoogleTranslate et pandas pour manipuler les données. Ensuite, il instancie un objet traducteur GoogleTranslator avec le français comme langue cible. Deux fonctions sont définies : trancher\_document() pour découper les longs textes en tranches de longueur maximale autorisée par l'API, et maFonctionTraducteur() pour effectuer la traduction proprement dite.*

*Cette dernière fonction prend en entrée un corpus de données textuelles au format pandas DataFrame. Pour chaque texte du corpus, elle appelle trancher\_document() si nécessaire, puis traduit chaque tranche de texte de l'anglais vers le français à l'aide de GoogleTranslator.translate(). Les traductions sont concaténées pour reconstituer le texte complet en français. Toutes les traductions sont stockées dans une liste, puis converties en un nouveau DataFrame pandas. Enfin, le corpus traduit est exporté dans un fichier CSV nommé "CorpusFrançais.csv".*

1. **Description de la fonction de nettoyage de texte**

La fonction créée est dénommée *« cleanResume »,* elle permet de nettoyer les CV en supprimant les éléments indésirables suivants : *URL, retweets, mentions Twitter, caractères de ponctuation caractères non alphanumériques (sauf les caractères accentués français) et espaces blancs en double.* La fonction prend un texte de CV en entrée et renvoie le texte nettoyé. Le processus de nettoyage utilise des expressions régulières pour rechercher et remplacer les éléments indésirables dans le texte

Le processus de nettoyage se déroule en plusieurs étapes. Nous illustrons ces étapes à travers le fonctionnement même de la fonction python construite :

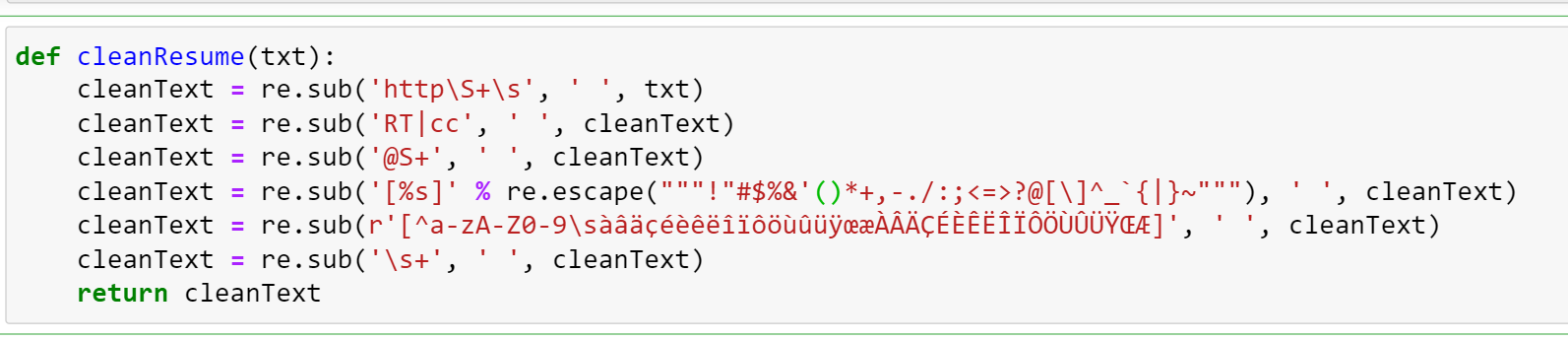


Figure 43: script python de la fonction de nettoyage

La fonction `cleanResume` nettoie et prétraite un texte (chaine de caractères) donné en entrée (« txt »). Elle effectue plusieurs opérations de nettoyage à l'aide d'expressions régulières (bibliothèque re) en Python. Les étapes de nettoyage se succèdent comme suit:

1. `re.sub('http\S+\s', ' ', txt)` : Remplace les URL commençant par "http" par un espace.
2. `re.sub('RT|cc', ' ', cleanText)` : Remplace les mentions "RT" (Retweet) et "cc" par un espace.
3. `re.sub('@S+', ' ', cleanText)` : Remplace les mentions d'utilisateurs commençant par "@" par un espace.
4. `re.sub('[%s]' % re.escape("""!"#$%&'()\*+,-./:;<=>?@[\]^\_`{|}~"""), ' ', cleanText)` : Remplace tous les caractères de ponctuation par un espace.
5. `re.sub(r'[^a-zA-Z0-9\sàâäçéèêëîïôöùûüÿœæÀÂÄÇÉÈÊËÎÏÔÖÙÛÜŸŒÆ]', ' ', cleanText)` : Remplace tous les caractères qui ne sont ni des lettres (majuscules ou minuscules), ni des chiffres, ni des espaces, ni des caractères accentués français, par un espace.
6. `re.sub('\s+', ' ', cleanText)` : Remplace les séquences d'espaces multiples par un seul espace.

Après avoir effectué toutes ces opérations de nettoyage, la fonction retourne le texte nettoyé `cleanText`.

1. **Description de la fonction de vectorisation**

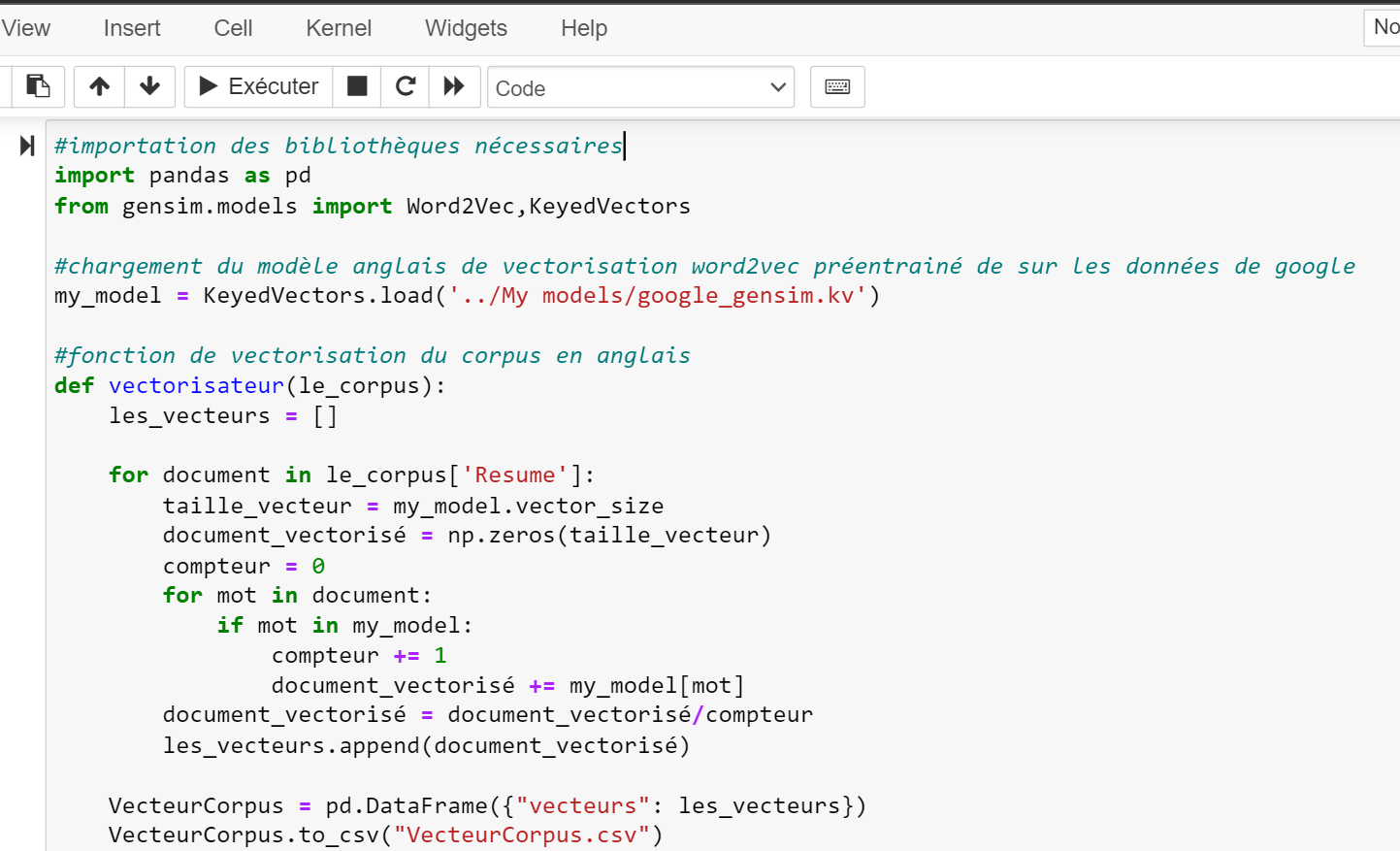


Figure 44: script python de vectorisation des CV en anglais

*Le script importe les bibliothèques pandas et gensim.models. Il charge ensuite un modèle Word2Vec pré-entraîné par Google sur de grandes données textuelles en anglais à l'aide de KeyedVectors.load(). La fonction vectorisateur () prend en entrée un DataFrame pandas le\_corpus avec une colonne 'Resume' contenant les documents textuels. Pour chaque document, elle initialise un vecteur nul de même dimension que les vecteurs Word2Vec. Puis, pour chaque mot du document présent dans le vocabulaire Word2Vec, elle ajoute son vecteur correspondant au vecteur du document. Après avoir parcouru tous les mots, le vecteur du document est normalisé par le nombre de mots pris en compte. Ces vecteurs de documents, représentant leur sémantique par la moyenne des vecteurs Word2Vec de leurs mots, sont stockés dans une liste. Celle-ci est finalement convertie en un DataFrame pandas VecteurCorpus, exporté dans un fichier CSV "VecteurCorpus.csv". La fonction de vectorisation du corpus en français est quasiment la même, la différence réside au niveau du modèle word2vec utilisé pour la vectorisation.*

1. **Vectorisation TFIDF**

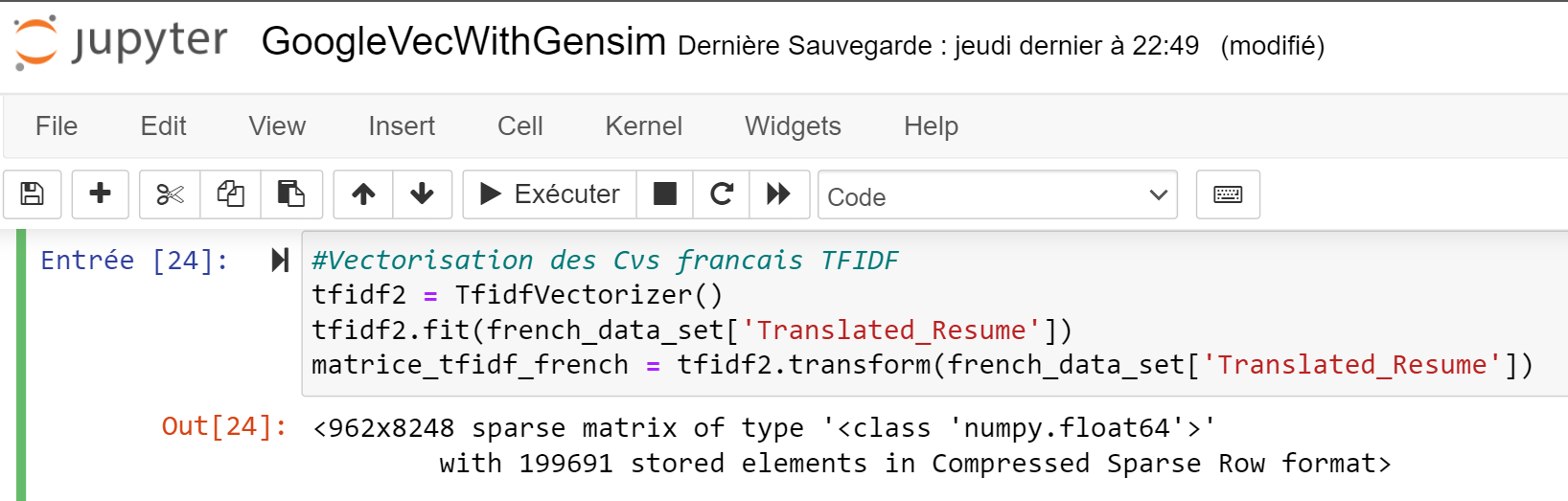


Figure 45: script python pour la vectorisation TFIDF

1. **Répertoire des modèles sérialisés**

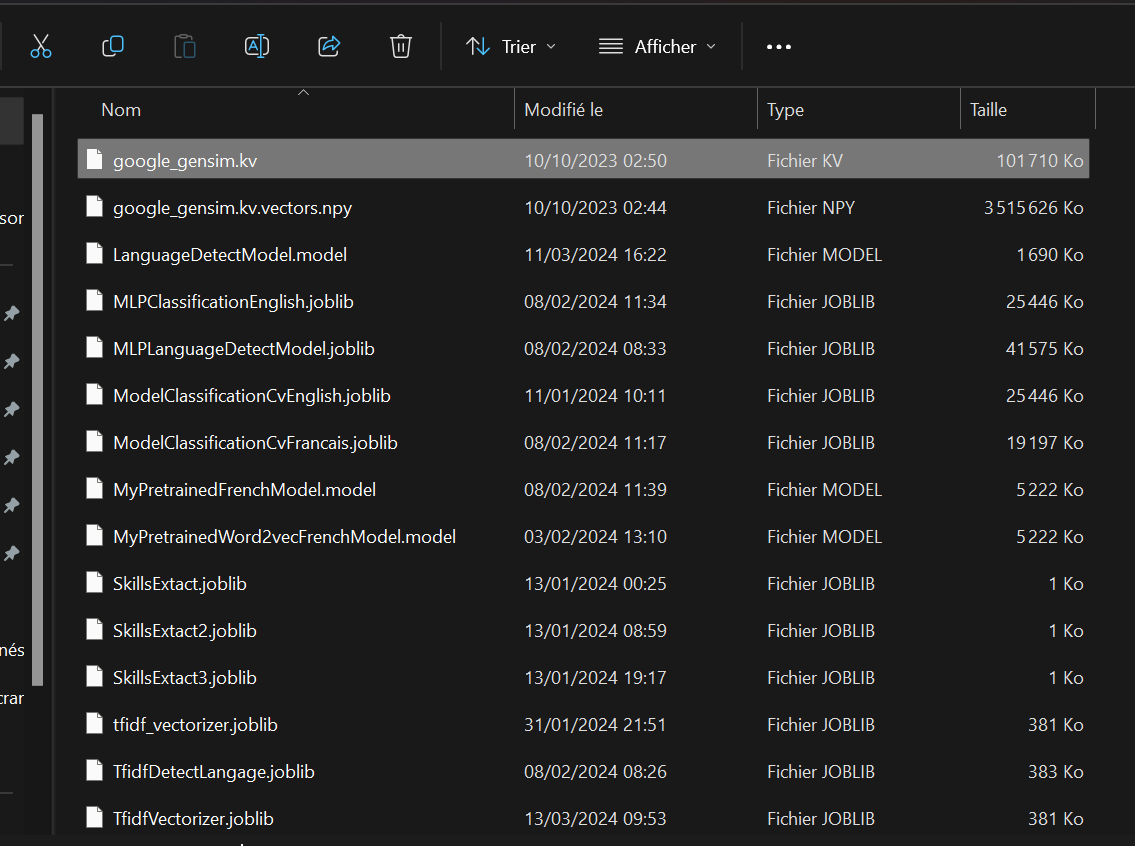


Figure 46:aperçu des modèles sérialisés

1. **Matrice de confusion du modèle KNN pour la classification des CV en anglais en utilisant la matrice tfidf**

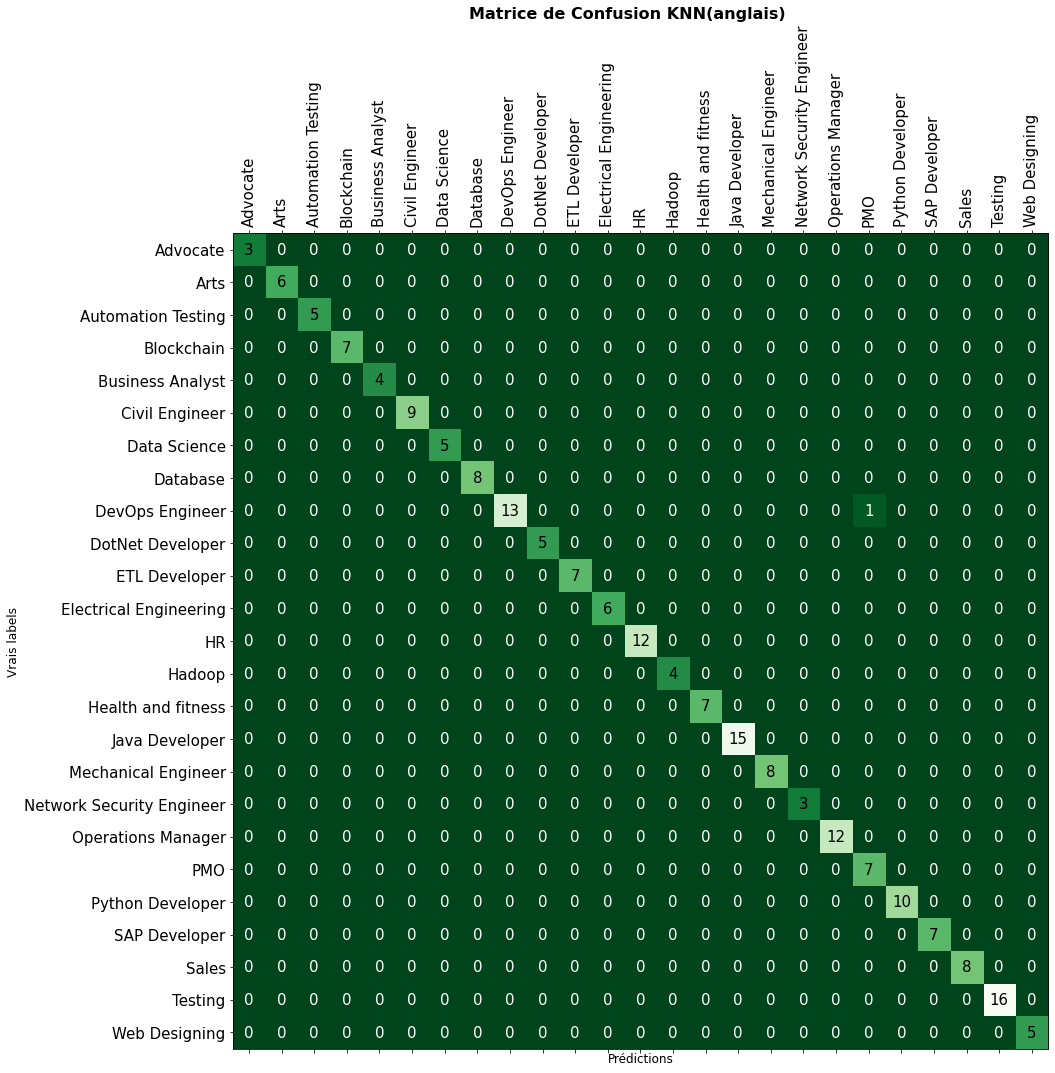


Figure 47:Matrice de confusion KNN pour la classification des CV en anglais(tfidf)

* Interprétation :

La matrice de confusion montre que le modèle a obtenu d'excellentes performances de classification pour chacune des 25 catégories professionnelles. En effet, on peut voir que presque tous les échantillons tests ont été bien classés car il y a moins de valeurs hors diagonale principale. Cela indique que le modèle a su prédire avec justesse la catégorie d'origine pour la quasi-totalité des exemples. On ne dénombre qu’une seule valeur hors-diagonale à savoir 1 exemple classé dans la catégorie "Database" au lieu de "Civil Engineering. Globalement, cela signifie que le modèle a appris de manière très efficace à bien distinguer les différentes catégories professionnelles les unes des autres.

1. **Matrice de confusion du modèle KNN pour la classification des CV en français en utilisant la matrice tfidf**

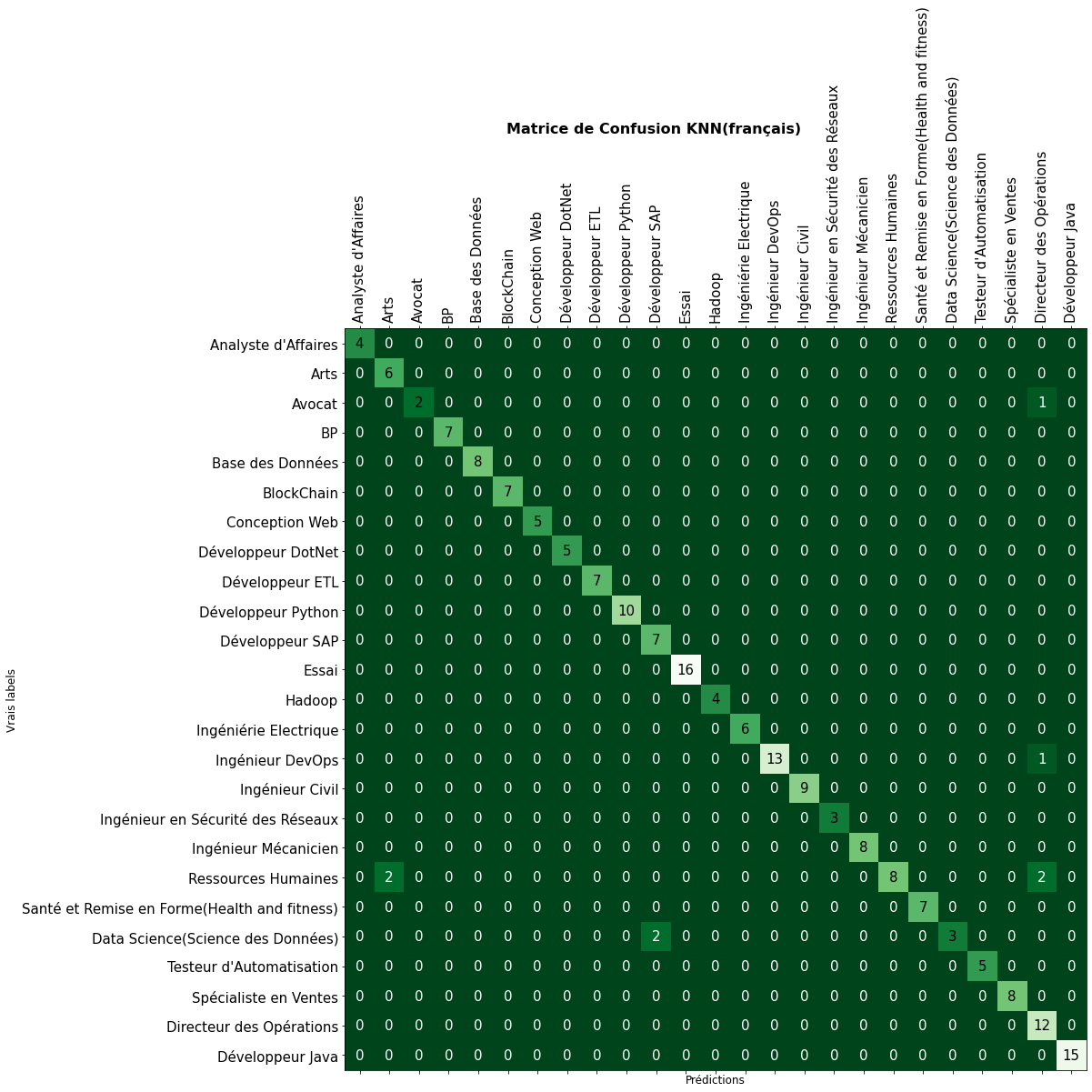


Figure 48:Matrice de confusion KNN pour la classification des CV en français (tfidf)

* Interprétation :

Le modèle de classification KNN a obtenu d'excellentes performances lors de son évaluation sur le corpus de CV en français, démontrant sa capacité à classifier correctement les données selon leur catégorie professionnelle. La précision atteinte est de 97,1% signifiant que moins de 3% des prédictions positives étaient erronées. Le rappel est légèrement inférieur mais reste très bon à 0,954, soit moins de 5% de faux négatifs. L'accuracy globale du modèle sur l'ensemble des tests s'élève à 0,959 et le score F1 moyen de 0,956 traduisant bien l'équilibre entre précision et rappel.

1. **Matrices de confusion du modèle SVM pour la classification des CV en anglais et en français (tfidf)**

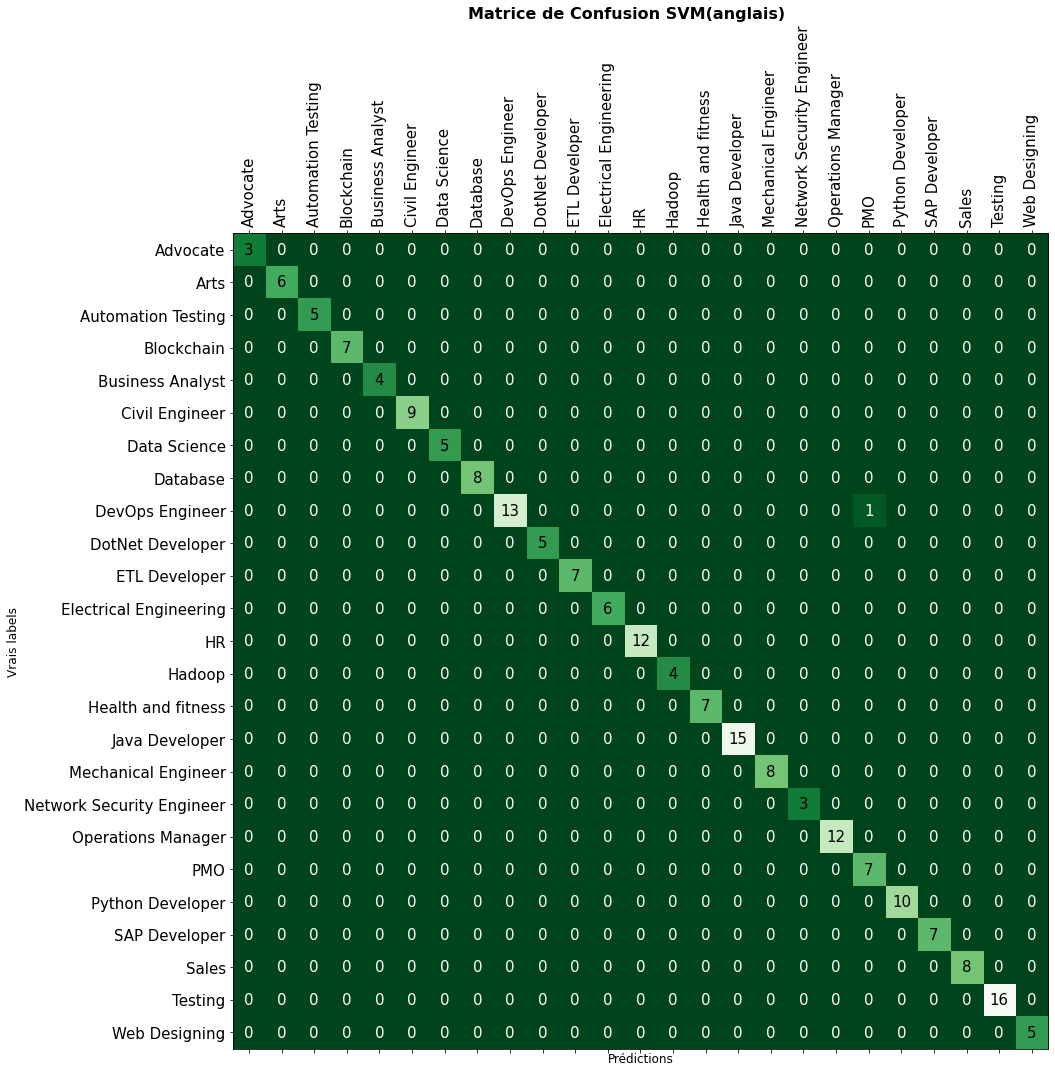


Figure 49:Matrice de confusion SVM pour la classification des CV en anglais (tfidf)

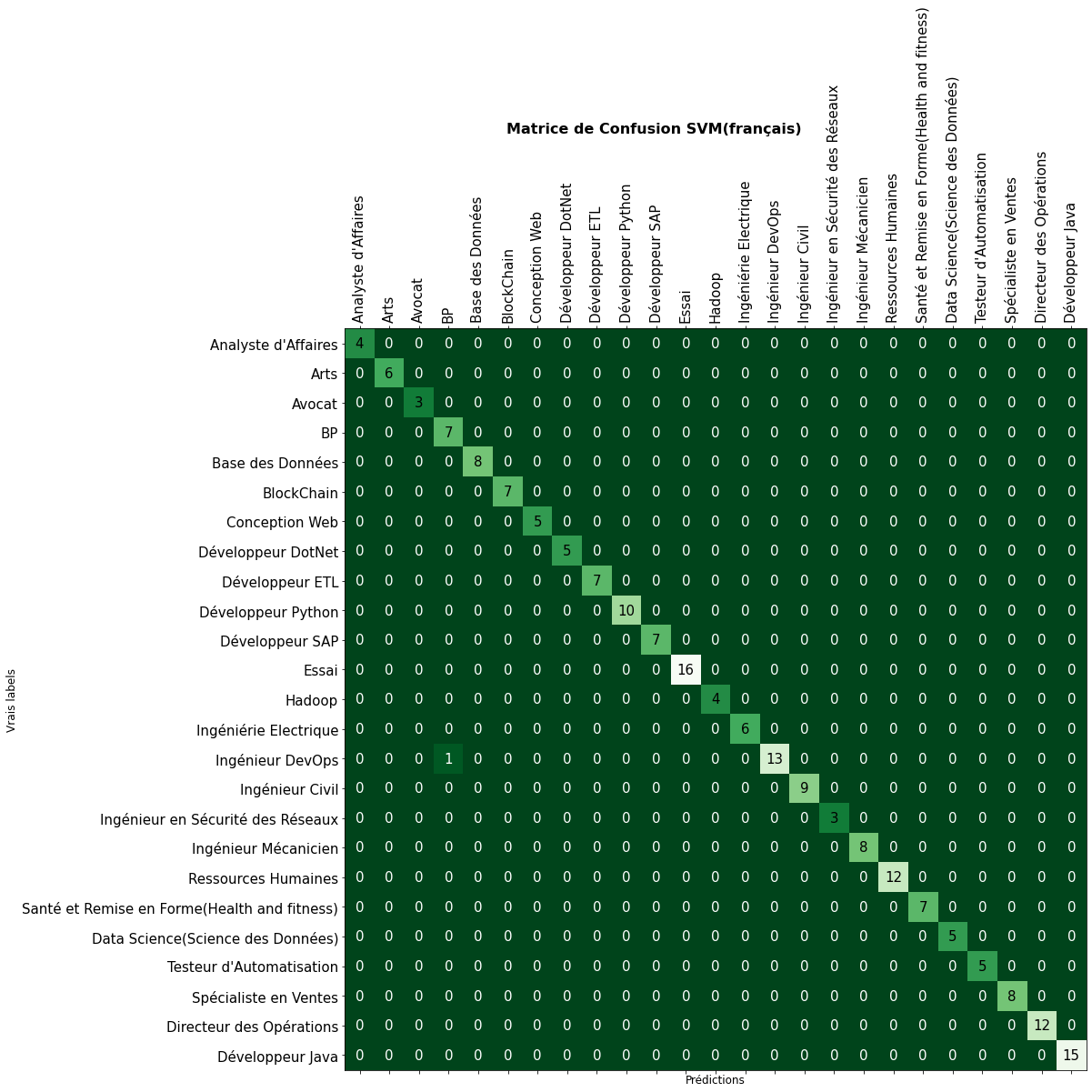


Figure 50:Matrice de confusion SVM pour la classification des CV en français(tfidf)

* Interprétation :

On constate globalement que le modèle SVM obtient d'excellents résultats de classification, avec de nombreuses catégories correctement identifiées.

La classe "Web Designing" (classe 24) affiche un score parfait de 15/15, validant la très bonne prédiction de ces échantillons. Les classes "Data Science" (classe 6 avec 9 échantillons), "Database" (classe 7 avec 8 éléments), "DevOps Engineer" (classe 8 avec 8 CV) et "DotNet Developer" (classe 9 avec 13 prédictions justes) sont également bien classées, avec des scores compris entre 8 et 13 bonnes prédictions. D'autres professions sont également bien identifiées avec plus de 5 échantillons corrects, comme "HR" (classe 12 avec 6 prédictions), "Java Developer" (classe 15 avec 7 CV) et "SAP Developer" (classe 21 avec 7 individus). Certaines ambiguïtés demeurent néanmoins, à l'image de 4 CV de la classe "Health and fitness" (classe 14) prédits en classe "HR". La catégorie "Electrical Engineering" (classe 11) reste plus difficile à discerner, avec seulement 7 prédictions justes. La minime différence entre les deux matrices de confusions et que dans celle du modèle entrainé sur le corpus en anglais, le modèle SVM a mal classé un échantillon de la classe « Civil Engineer » en le désignant comme « Database», et dans la seconde, le modèle SVM a mal classé un échantillon de la classe «Engénieur DevOps» en le désignant comme «BP».

1. **Matrice de confusion du modèle Random Forest pour la classification des CV en anglais en utilisant (tfidf)**

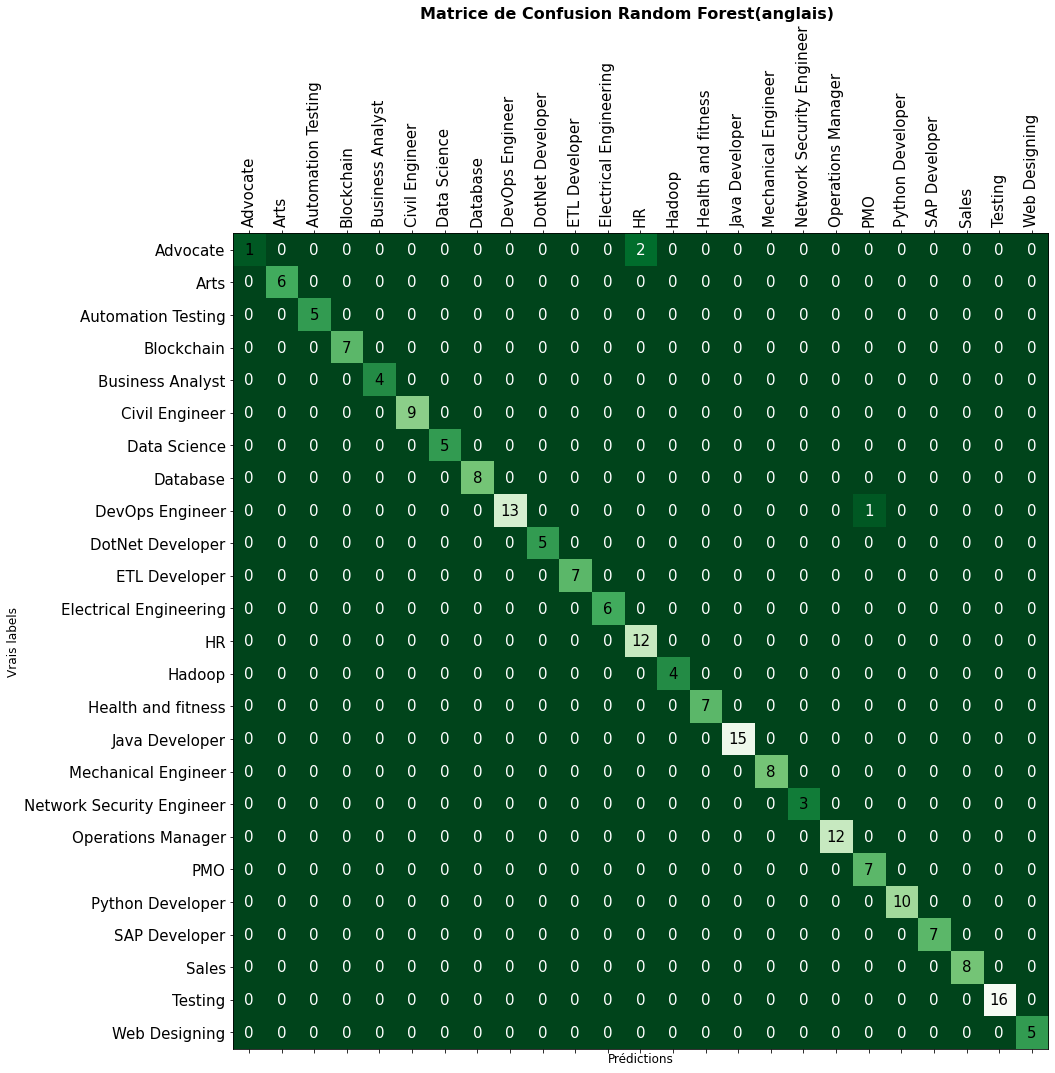


Figure 51:Matrice de confusion Random Forest pour la classification des CV en anglais(tfidf)

* Interprétation :

En effet, la majorité des valeurs sont situées sur la diagonale principale, indiquant que de nombreux échantillons tests ont été correctement affectés à leur classe d'origine, avec par exemple 4 CV classés en catégorie « Arts », 6 en catégorie « Automation Testing », 5 en catégorie « Blockchain », 7 en catégorie « Civil Engineer » et jusqu'à 13 échantillons bien classés en catégorie « DotNet Developer » et 15 documents en catégorie « Java Developer ». Quelques rares confusions subsistent notamment 1 CV attribué à la classe « PMO » au lieu de « DevOps Engineer », et 2 CV attribués à la classe « HR » au lieu de « Advocate ».

1. **Matrice de confusion du modèle Random Forest pour la classification des CV en français (tfidf)**

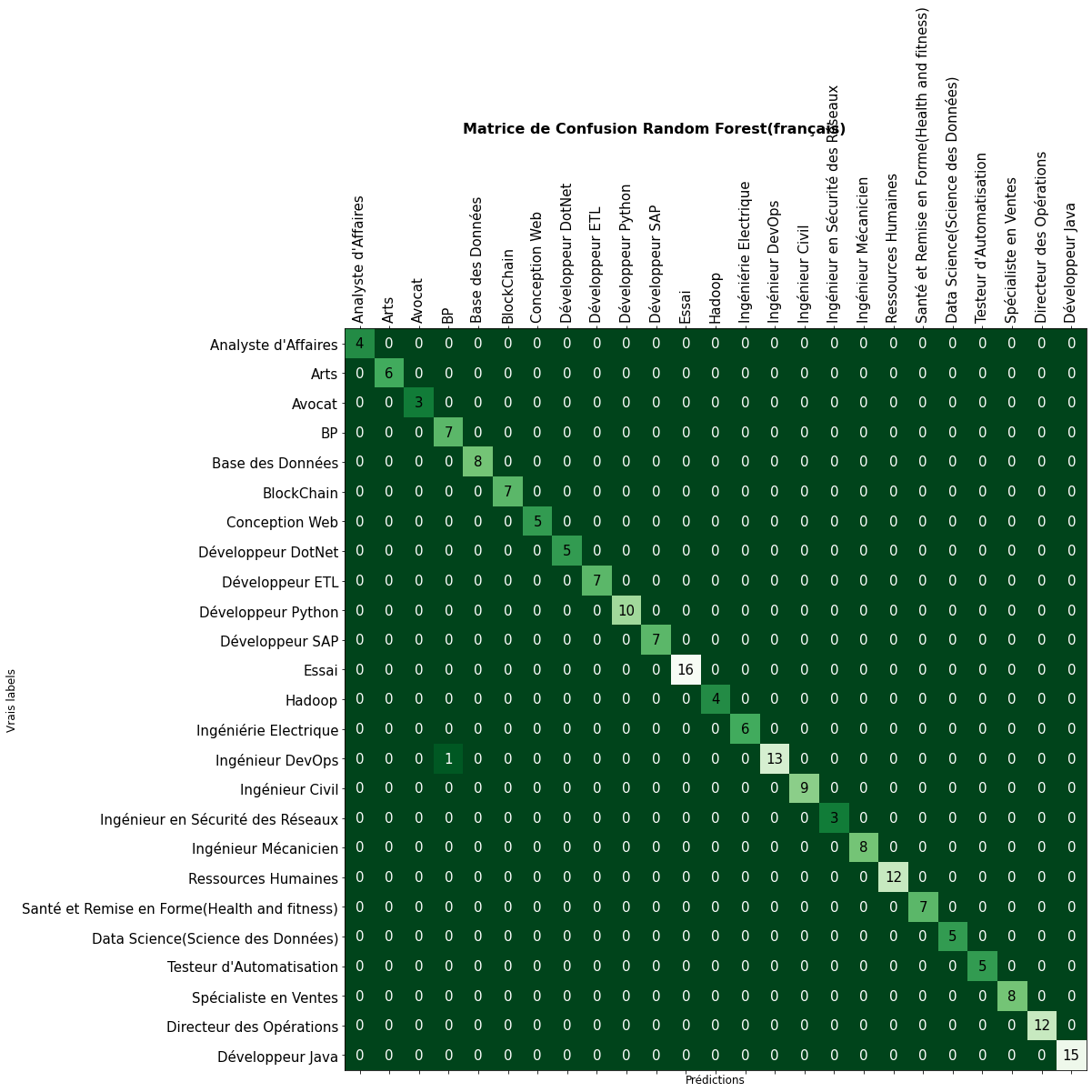


Figure 52:Matrice de confusion Random Forest pour la classification des CV en français(tfidf)

* Interprétation :

La matrice de confusion montre que presque toutes les catégories ont été correctement identifiées par le modèle Random Forest, avec quasiment tous les éléments non nuls situés sur la diagonale principale. Cela indique une très bonne performance du modèle, à l'exception d'un échantillon de la classe "Ingénieur DevOps" qui a été mal classé en "BP".

1. **Matrice de confusion du modèle MLP pour la classification des CV en anglais en utilisant la matrice tfidf**

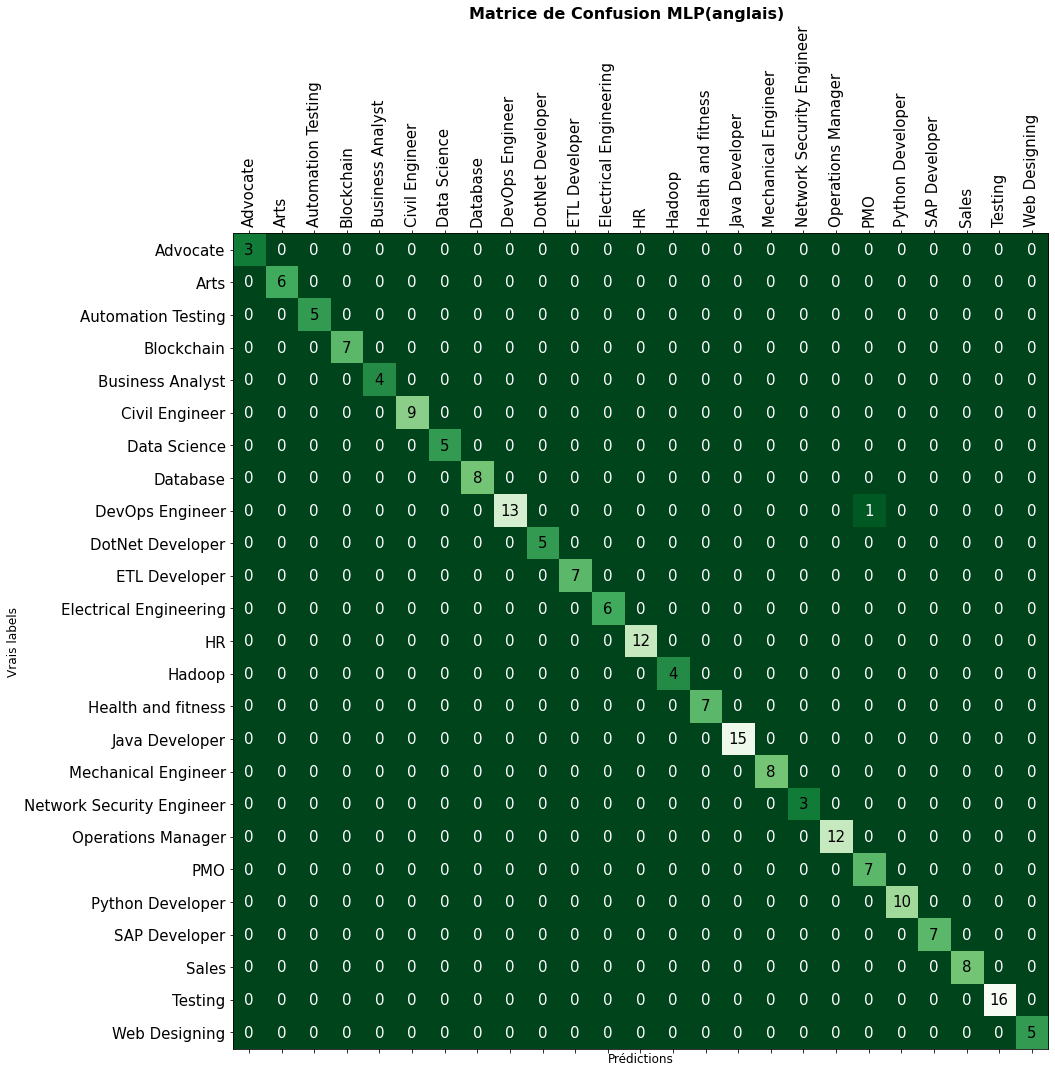


Figure 53:Matrice de confusion MLP pour la classification des CV en anglais (tfidf)

* Interprétation :

L'analyse de cette matrice de confusion révèle que le modèle MLP a réalisé une classification très performante, avec la majorité des éléments non nuls concentrés sur la diagonale principale. Cela démontre que pratiquement toutes les catégories ont été correctement identifiées par le modèle, témoignant d'une excellente capacité de prédiction. Cependant, on note une exception pour un échantillon de la classe "DevOps Engineer" qui a été erronément classé dans la catégorie "PMO", indiquant une légère confusion entre ces deux classes.

1. **Matrice de confusion du modèle MLP pour la classification des CV en français en utilisant la matrice tfidf**

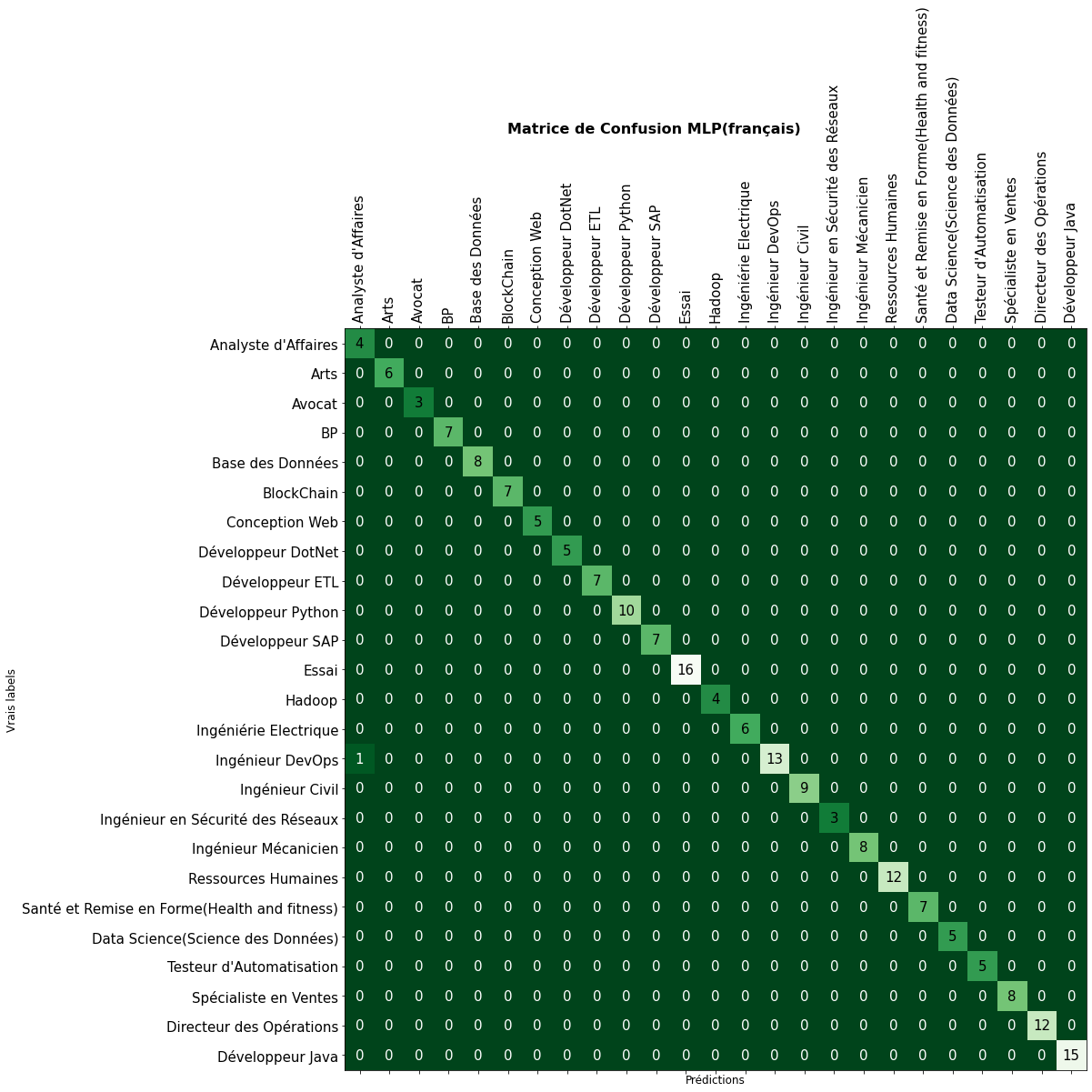


Figure 54:Matrice de confusion MLP pour la classification des CV en français (tfidf)

* Interprétation :

L'analyse de la matrice de confusion du modèle Random Forest montre que la grande majorité des catégories ont été très bien identifiées, avec presque tous les éléments non nuls situés sur la diagonale principale. Cela indique une performance globale remarquable du modèle, avec 100% des observations correctement classées pour 21 des 25 catégories.

Toutefois, on note une exception pour la classe "Ingénieur DevOps", dont un échantillon a été mal classé dans la catégorie "Analyste d’affaires", indiquant une légère confusion entre ces deux classes.

1. **Matrice de confusion du modèle KNN pour la classification des CV en anglais en utilisant les vecteurs word2vec**

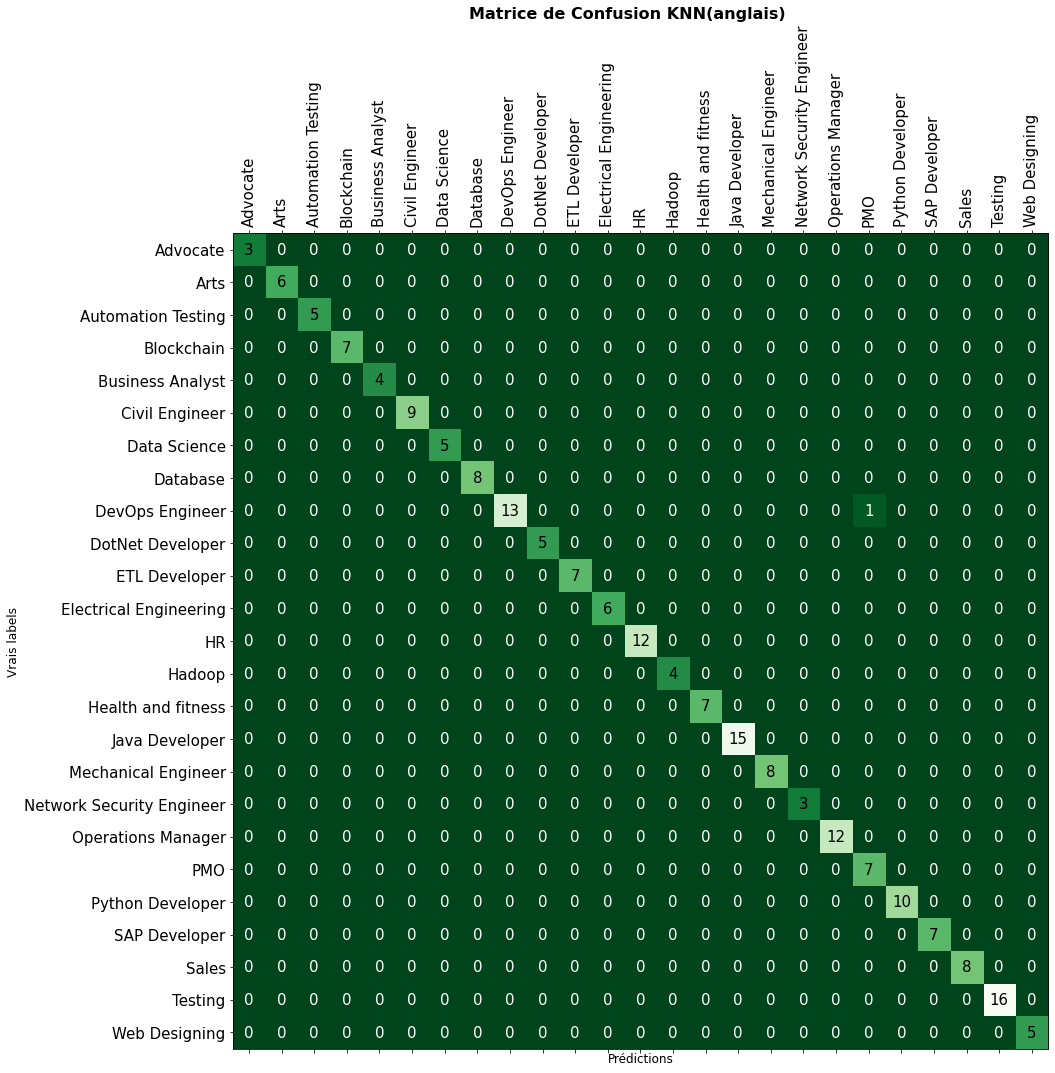


Figure 55:Matrice de confusion KNN pour la classification des CV en anglais (wv)

* Interprétation :

L'analyse de cette matrice de confusion de ce modèle classificatoire révèle des résultats remarquables. La majorité des éléments sur la diagonale principale indiquent une précision de classification de 100% pour la plupart des catégories, témoignant d'une excellente capacité du modèle à identifier correctement les différentes classes. Seule la catégorie "Devops Engineer" présente un taux de classification légèrement inférieur, avec un exemple mal classé, attribué à la classe « PMO ».

1. **Matrice de confusion du modèle KNN pour la classification des CV en (word2vec)**

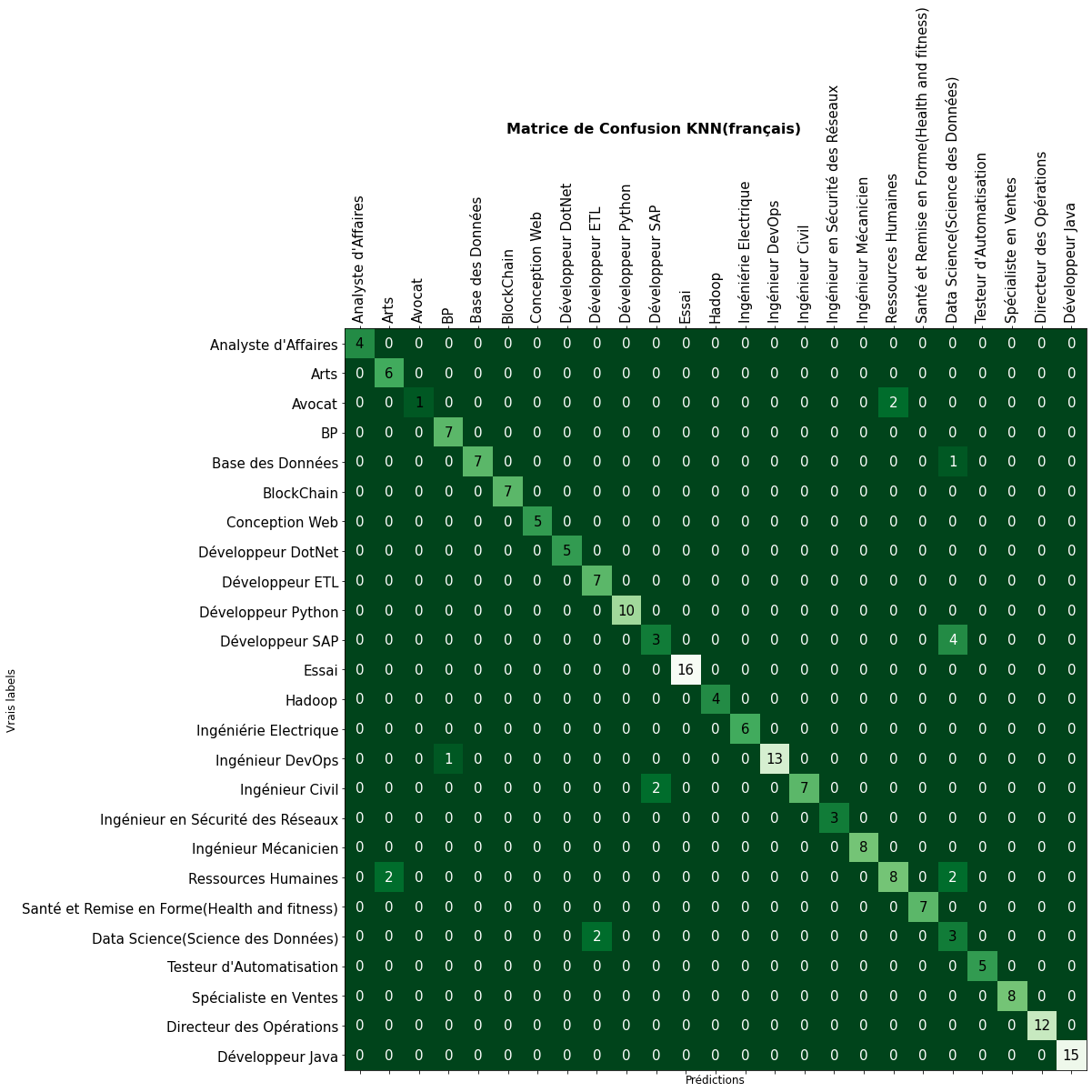


Figure 56:Matrice de confusion KNN pour la classification des CV en français (word2vec)

* Interprétation :

D'après la matrice de confusion, on constate que certaines catégories sont bien classées avec une précision élevée, comme les catégories "Essai" avec 16 vrais positifs sur 16, "Développeur Python" avec 10 vrais positifs sur 10, "Développeur Java" avec 15 vrais positifs sur 15 et "Développeur DotNet" avec 5 vrais positifs sur 5. Cependant, d'autres catégories semblent poser plus de difficultés avec des confusions fréquentes, notamment les catégories "Développeur SAP" qui est confondue 4 fois avec "Science des données", "Avocat" qui est confondue 2 fois avec "Ressources Humaines", "Science des données" qui est confondue 2 fois avec "Développeur ETL", « Ressources humaines» qui est confondue 2 fois avec "Arts" et également 2 fois avec "Science des données", "Ingénieur civil" qui est 2 confondue avec "Développeur SAP", "Ingénieur DevOps" qui est 1 confondue avec "BP", et "Base des données" qui est 1 fois confondue avec "Science des Données".

1. **Matrice de confusion du modèle SVM pour la classification des CV en anglais en utilisant les vecteurs word2vec**

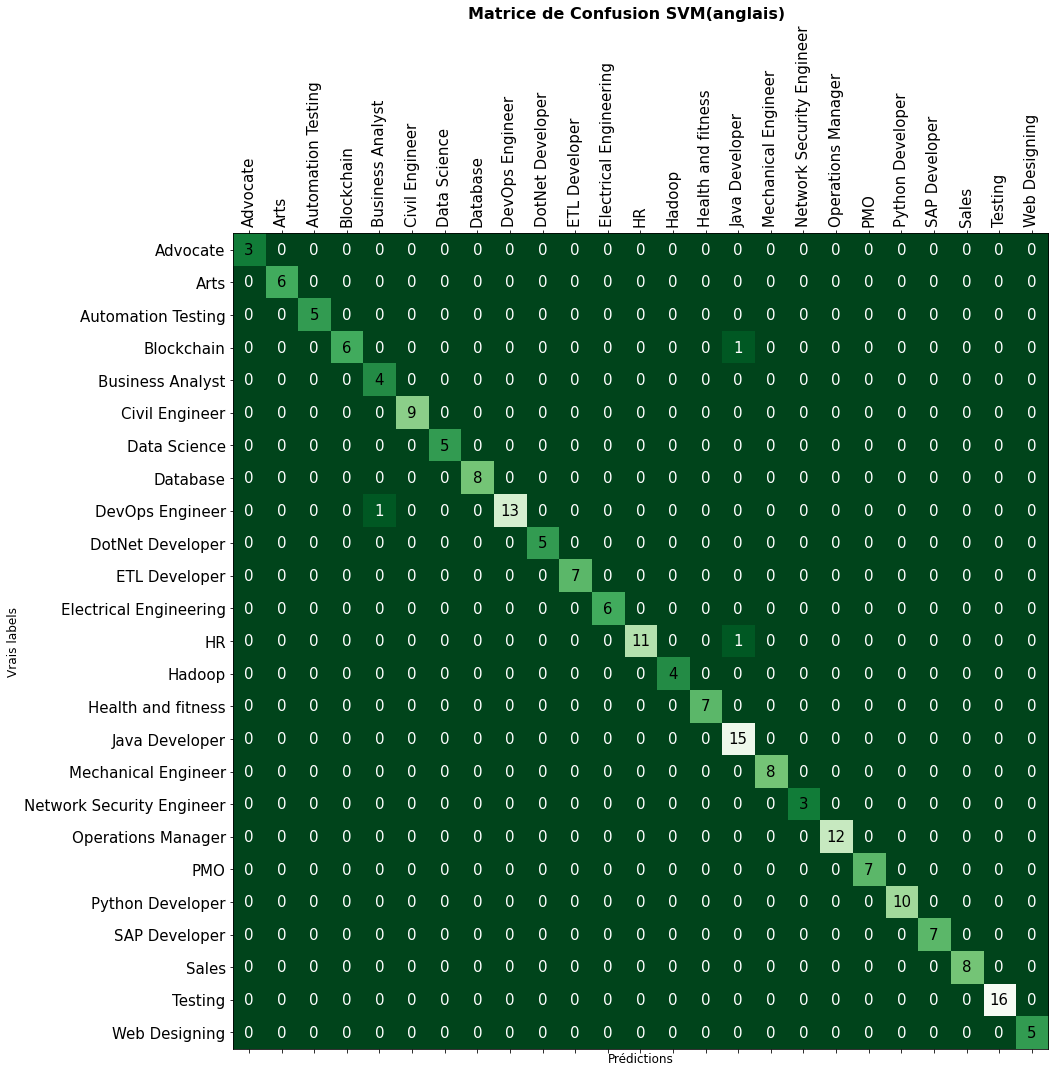


Figure 57:Matrice de confusion SVM pour la classification des CV en anglais(wv)

* Interprétation :

Cette matrice de confusion montre que le modèle de classification a globalement de bonnes performances, avec la majorité des instances correctement classées (par exemple, 5 instances correctement classées dans la catégorie 24 "Web Designing", 15 dans la catégorie 15 "Java Developer", 12 dans la catégorie 18 "Operation Manager"). Cependant, on peut noter quelques erreurs de classification, comme 1 instance de "Business Analyst" (catégorie 4) classée dans la catégorie "Java Developer", 1 instance de "DevOps Engineer" classée dans la catégorie "Business Analyst" et 1 instance de "Blockchain" classée dans la catégorie "Java Developer", 1 instance de "HR" classée dans la catégorie "Java Developer".

1. **Matrice de confusion du modèle SVM pour la classification des CV en français en utilisant les vecteurs word2vec**

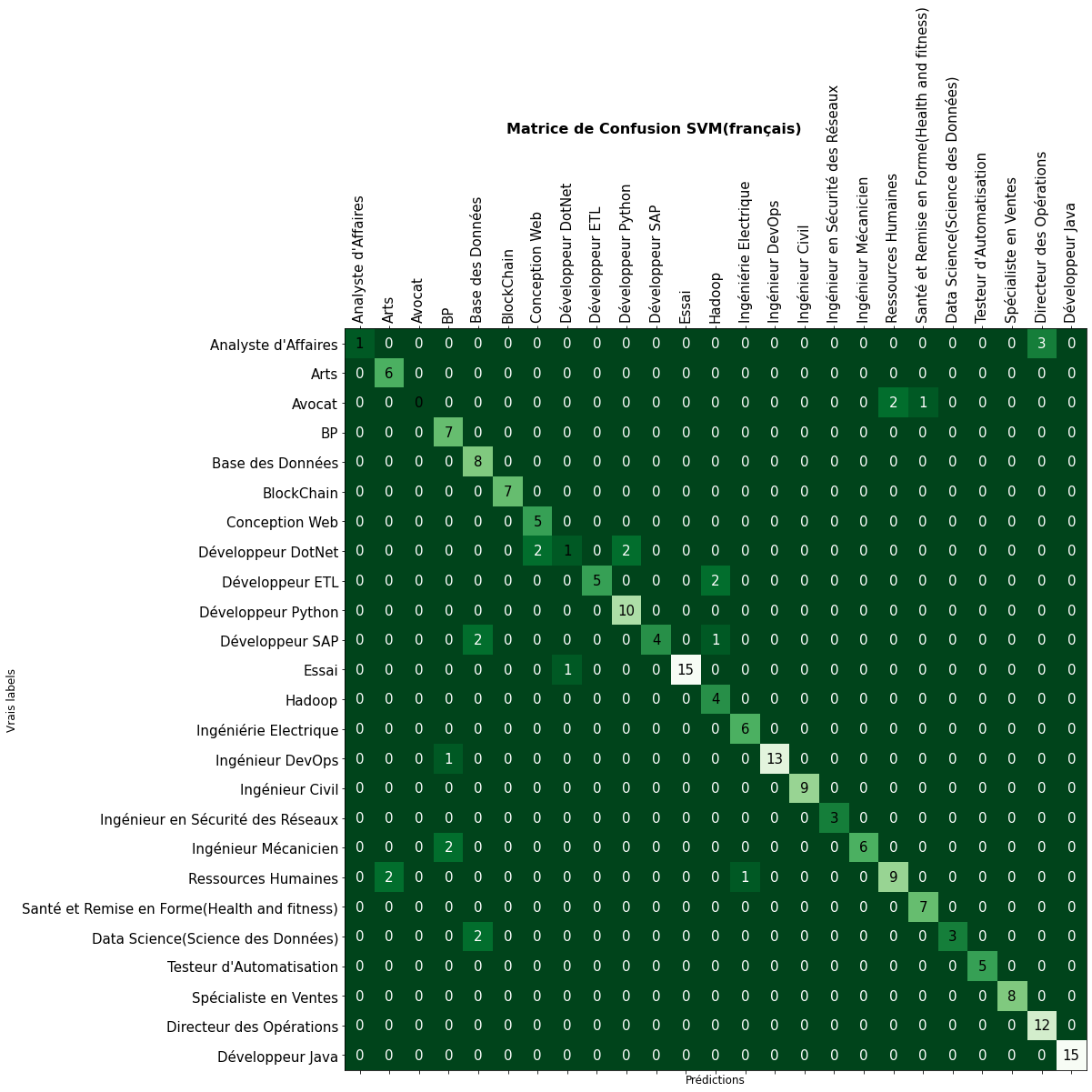


Figure 58:Matrice de confusion SVM pour la classification des CV en français (word2vec)

* Interprétation :

La matrice de confusion révèle que le modèle SVM a connu certaines difficultés à classer correctement certaines catégories. Par exemple, 3 échantillons de la catégorie "Analyste d'Affaires" ont été classés à tort dans la catégorie "Directeur des Opérations". De même, 2 exemples de la catégorie "Avocat" ont été mal classés dans la catégorie "Ressources Humaines". Le modèle a également confondu quelques exemples de "BP" avec "Ingénieur DevOps", de "Développeur DotNet" avec "Conception Web" et "Développeur python", de "Développeur ETL" avec "Hadoop", de "Développeur SAP" avec "Base des Données" et "Hadoop", et de "Ingénieur Mécanicien" avec "BP".

1. **Matrice de confusion du modèle Random Forest pour la classification des CV en anglais (word2vec)**

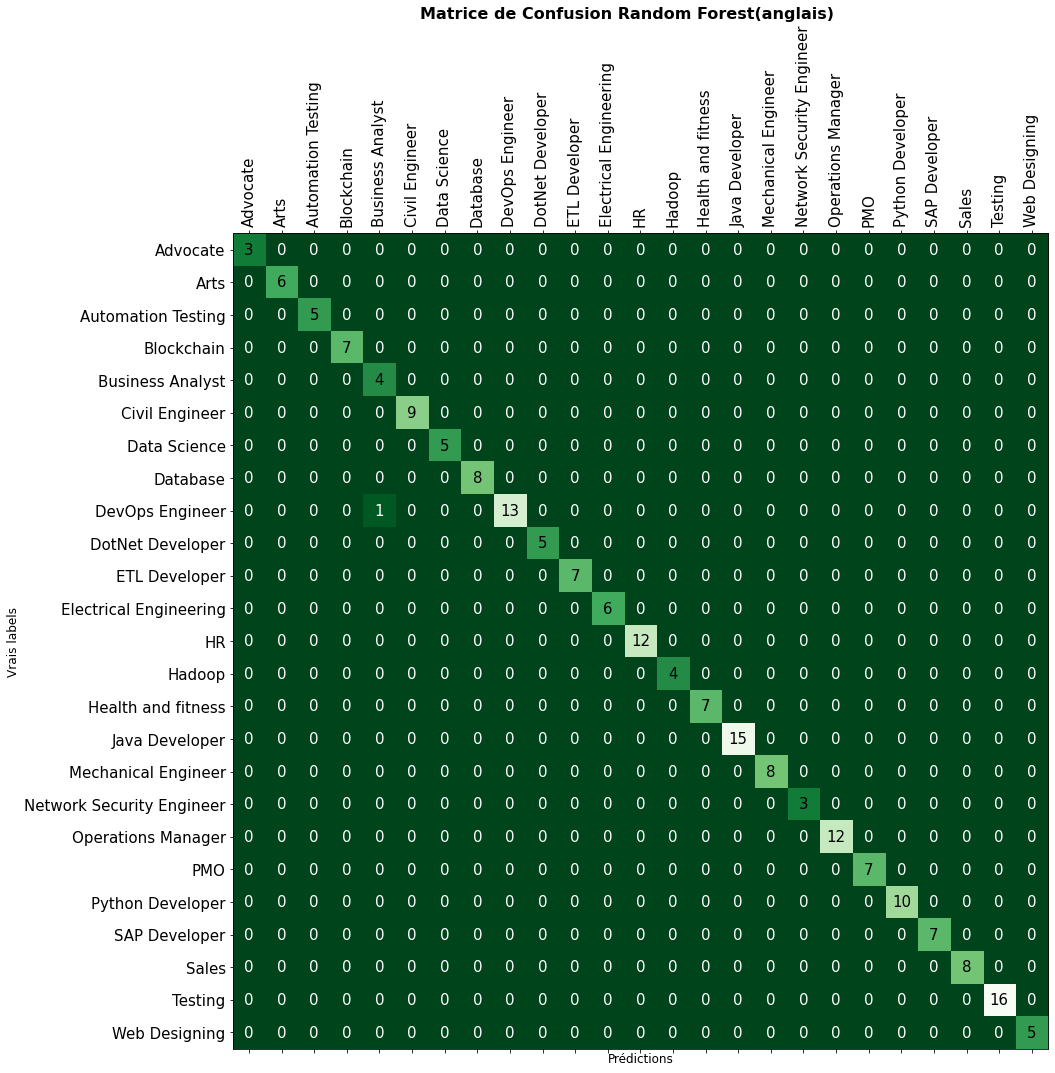


Figure 59:Matrice de confusion Random Forest pour la classification des CV en anglais (wv)

* Interprétation :

La matrice de confusion illustre que la grande majorité des catégories sont classées correctement. Notamment, "Testing" (16/16), "Java Developer" (15/15), "Python Developer" (10/10), "HR" (12/12), "Operations Manager" (12/12) et "Database" (8/8) sont parfaitement classées. La seule véritable confusion concerne 1 seul exemple de "DevOps Engineer" mal classé comme "Business Analyst".

1. **Matrice de confusion du modèle Random Forest pour la classification des CV en français en utilisant les vecteurs word2vec**

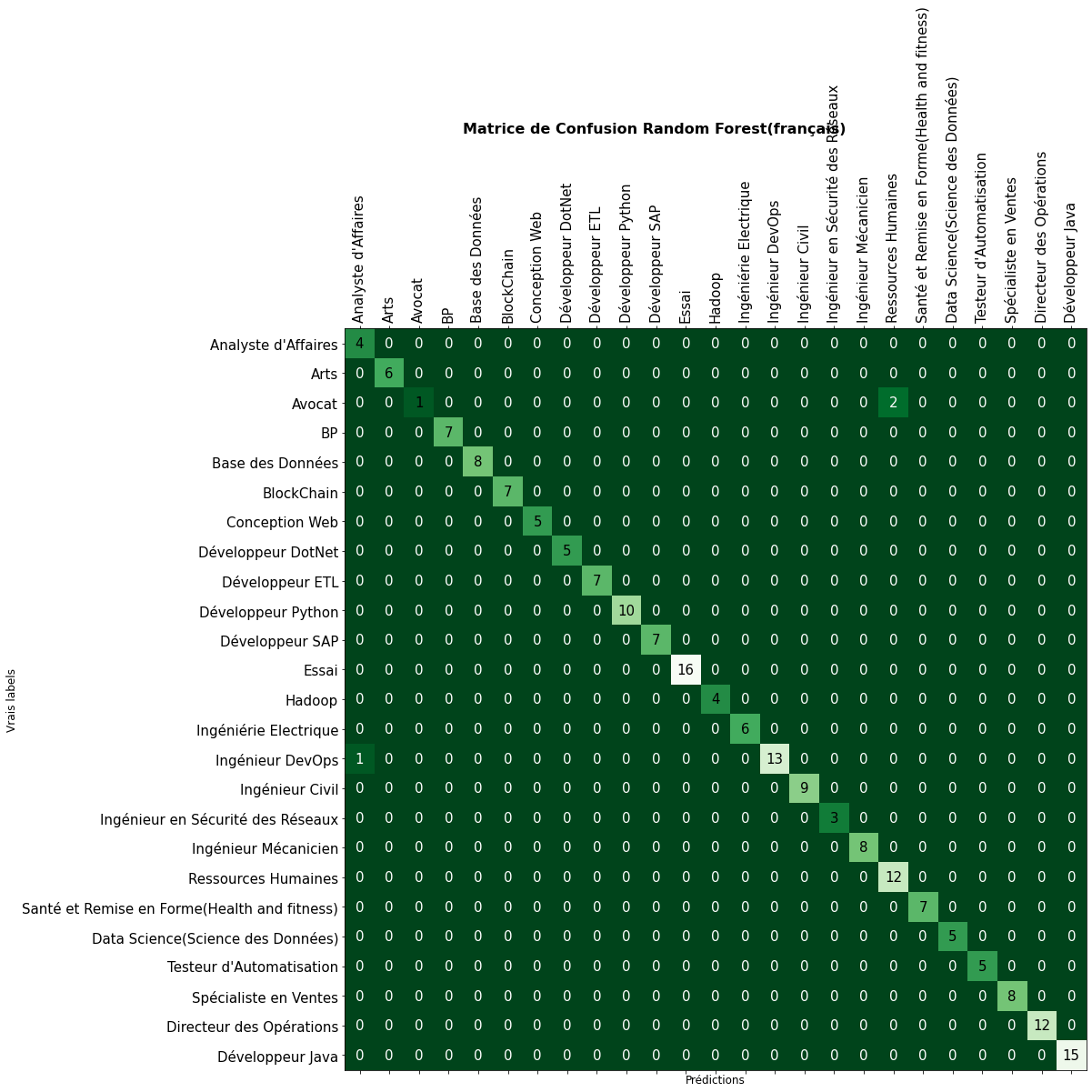


Figure 60:Matrice de confusion (Random Forest) pour la classification des CV en français (wv)

* Interprétation :

La matrice de confusion illustre des nombreuses catégories classées parfaitement comme "Testeur d'Automatisation" (5/5), "Développeur Java" (15/15), "Développeur Python" (10/10), "Essai" (16/16) et "Directeur des Opérations" (12/12). Cependant, quelques confusions mineures sont à noter, avec 1 exemple d'"Ingénieur DevOps" confondu avec "Analyste d’Affaires" et 2 exemples d'"Avocat" classés dans "Ressources Humaines".

1. **Matrice de confusion du modèle MLP pour la classification des CV en anglais en utilisant les vecteurs word2vec**

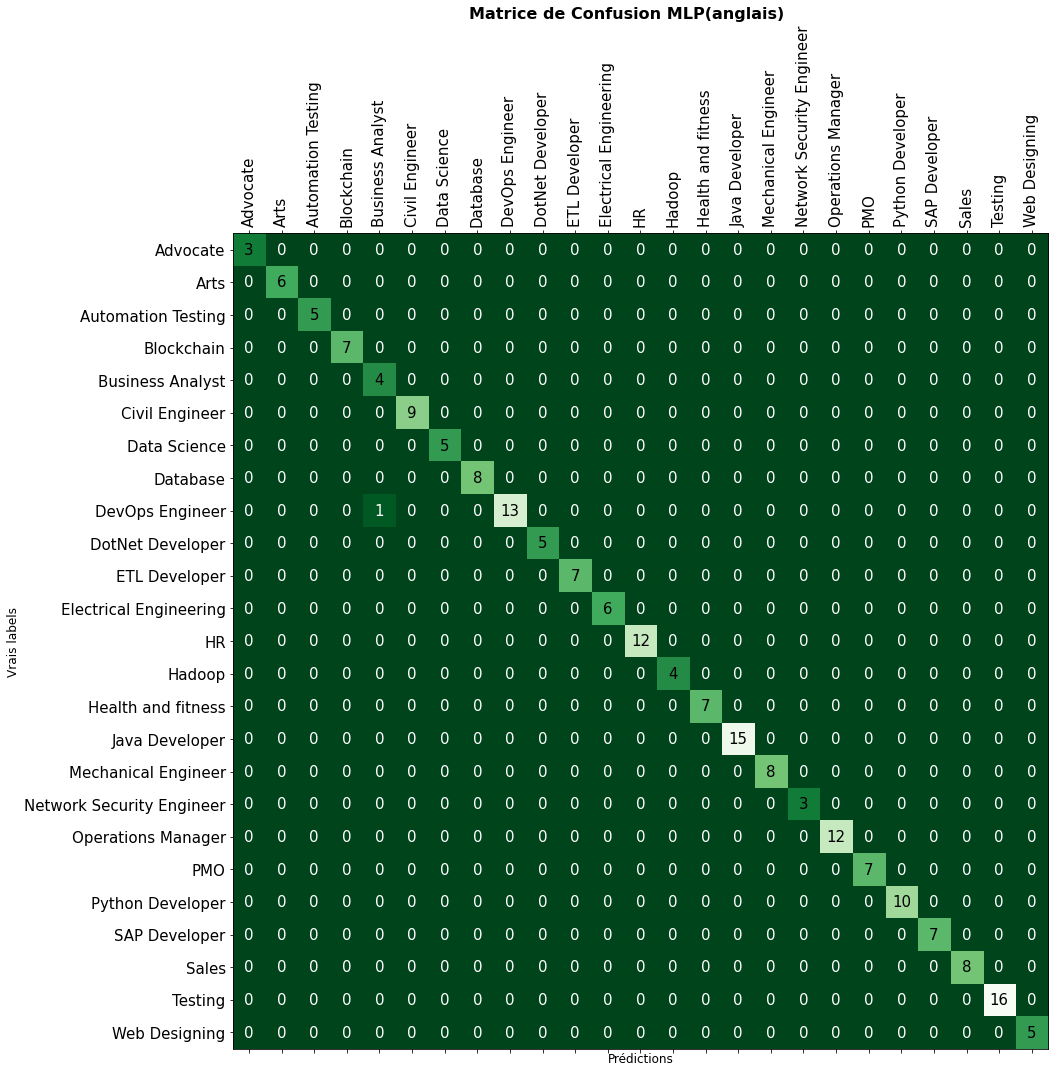


Figure 61:Matrice de confusion MLP pour la classification des CV en anglais (wv)

* Interprétation :

La matrice de confusion montre que le modèle a correctement classé quasiment tous les échantillons car la quasi-totalité des valeurs non nulles de la matrice sont concentrées sur la diagonale principale à l’exception d’un seul échantillon. Il s’agit d’un échantillon de la classe « DevOps Engineer » classé à tort dans la classe « Business Analyst».

1. **Matrice de confusion du modèle MLP pour la classification des CV en français (word2vec)**

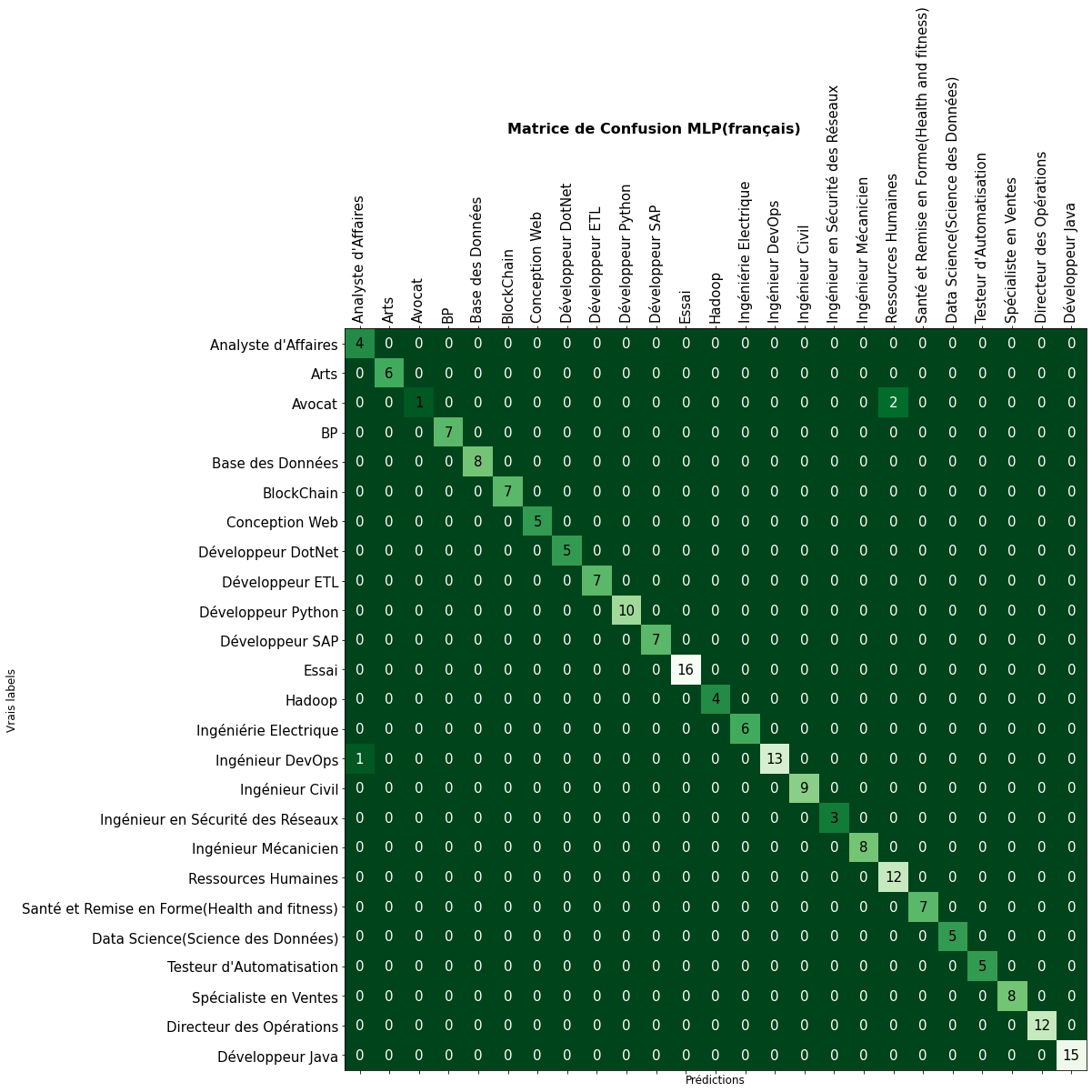


Figure 62:Matrice de confusion MLP pour la classification des CV en français (wv)

* Interprétation :

La matrice de confusion du modèle MLP pour ce cas révèle que certaines catégories comme « Développeur Python », « Essai », « Développeur ETL », « Ingénieur Civil » et « Directeur des Opérations » ont été bien classées, tandis que d'autres catégories comme « Avocat » et « Ingénieur DevOps » ont connu des confusions, notamment avec « Analyste d'Affaires » pour « Ingénieur DevOps » (1 « échantillon mal classé) et « Ressources Humaines » pour « Avocat » ( 2 échantillons mal classés).

1. **Fonction permettant le web scraping d’offres d’emploi**

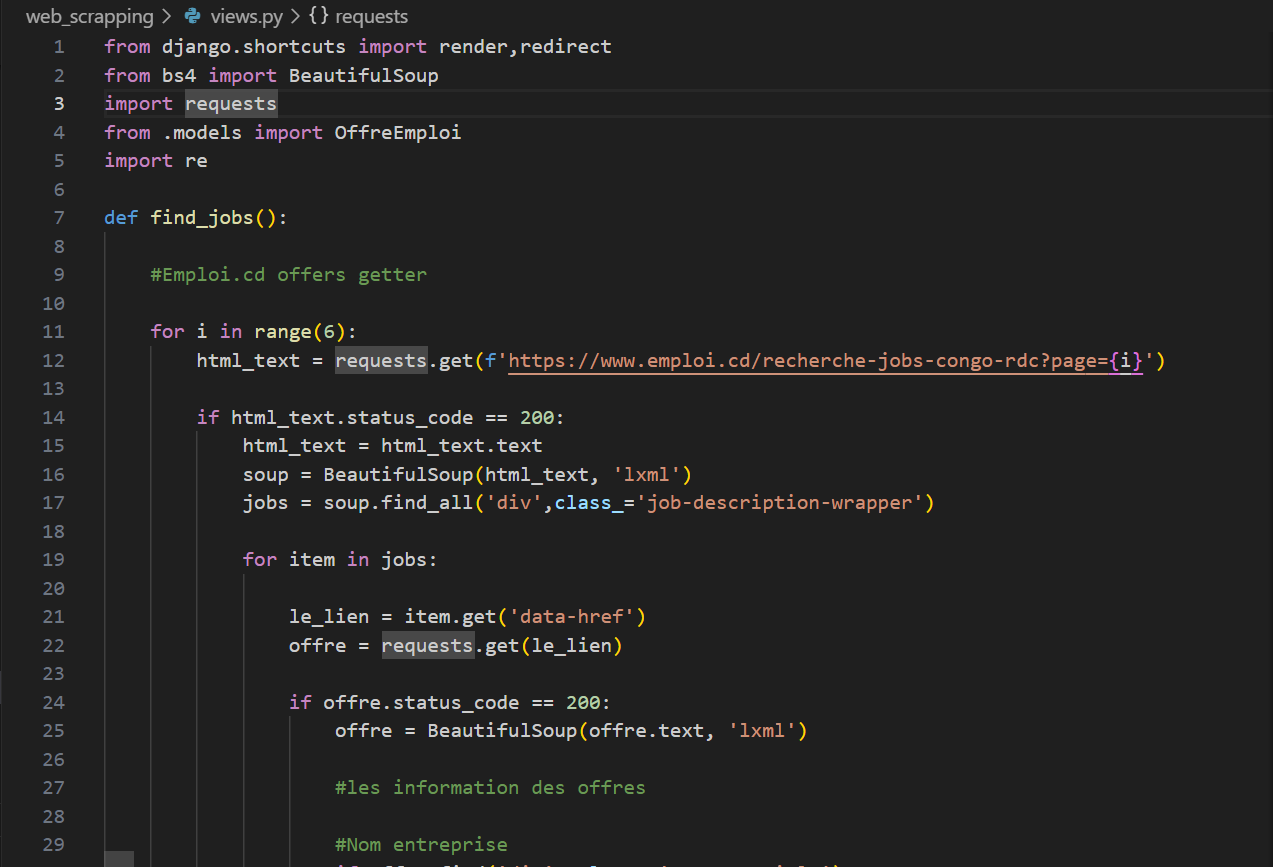


Figure 63:extrait de la fonction de web scraping

La fonction « find\_jobs () » implémente le web scraping du site d'emplois "emploi.cd" afin d'extraire et afficher les détails des offres d'emploi disponibles. Elle utilise les bibliothèques Python requests pour récupérer le contenu HTML des pages web, et BeautifulSoup pour analyser (parser en anglais) ce contenu et en extraire les informations pertinentes telles que le nom de l'entreprise, les compétences requises, la localisation, le type de contrat, etc. La fonction parcourt les 6 premières pages de résultats et effectue une analyse approfondie de chaque offre d'emploi trouvée en utilisant des expressions régulières pour cibler les données spécifiques à extraire. Les informations sont ensuite enregistrées dans la base des données pour être recommandées plus tard aux chercheurs d’emploi.