Module 10 – NLP et OCR
Synthèse et points clés

01 **Le NLP (Natural Language Processing ou traitement du langage naturel) est un domaine de l'intelligence artificielle qui se concentre sur la compréhension, l'analyse et la génération de texte ou de langage humain par des machines.**

02 **Le NLP simplifie le langage pour le rendre compréhensible par un algorithme** (réduction d'une phrase en tokens optimisés, puis transformation des tokens en valeurs numériques).

03 **Le NLP nécessite un processus de nettoyage des données** pour éliminer les erreurs de saisie, les caractères indésirables et les données non pertinentes, **garantissant ainsi la qualité des informations.**

04 Les ordinateurs ne comprennent pas les mots de la même manière que les humains, les mots doivent être transformés en formes numériques pour être traités par les algorithmes.

05 **Le TF-IDF (Term Frequency-Inverse Document Frequency) est une méthode de pondération des mots,** en fonction de leur fréquence dans un document et de leur rareté dans l'ensemble du corpus. **Cela permet de donner plus de poids aux mots rares et significatifs.**

06 **Les embeddings de mots sont des représentations distinctes de mots qui capturent les relations sémantiques et contextuelles entre les mots, permettant de mieux comprendre le sens des phrases.**

07 **Le NLP est utilisé notamment pour :**

- Automatiser les réponses aux e-mails,
- Générer automatiquement du contenu textuel,
- Effectuer la reconnaissance automatique de la parole,
- Analyser les médias sociaux pour déterminer les sentiments,
- Traduire des textes d'une langue à une autre.

08 **L'OCR (Optical Character Recognition ou reconnaissance optique de caractères) est une technologie qui permet de convertir des textes imprimés ou manuscrits en texte numérique éditable. Il a évolué de l'OCR classique à l'IDP (Intelligent Document Processing), grâce à l'intelligence artificielle.**

09 **L'OCR comprend deux étapes :** la détection des zones de texte dans un document, suivie de la reconnaissance des caractères pour traduire le texte manuscrit en format numérique.

10 **L'OCR est notamment utilisé pour :**

- Numériser des documents imprimés,
- Extraire des informations de factures, cartes de visite, documents historiques,
- Indexer des documents numérisés,
- Simplifier la gestion des documents dans divers secteurs.