serda formation

Informations :

Durée: 1.00 jour(s)
Prix: 700€
Code Formation: FA126
Eligible à un parcours de certification:

Prochaines dates :

- 24/09/2025

Public concerné :

Documentalistes, archivistes, bibliothécaires, gestionnaires de l'information

Pré-requis :

Avoir une base de connaissance en gestion documentaire et gestion des données. Être à l'aise avec les outils numériques

Méthodes pédagogiques :

- Apports théoriques et méthodologiques : 50%
- Exercices en ateliers : 30%
- Echanges, débats et retours d'expérience : 20%

Préparer ses données et ses documents pour entrainer les IA

Domaine: Formations Intelligence Artificielle
Niveau formation: Intermédiaire

Objectifs :

Former les documentalistes, archivistes et bibliothécaires aux méthodes et outils nécessaires pour préparer, structurer et enrichir les données textuelles et documents non structurés, afin de les rendre exploitables par des modèles d'intelligence artificielle (Machine Learning et IA générative).

Contenu du stage :

Comprendre les bases de l'intelligence artificielle et son application aux documents et données non structurées

- Définitions et concepts clés : IA, Machine Learning, NLP, IA générative
- Rôles des documentalistes, archivistes et bibliothécaires dans l'écosystème IA
- Exemples d'application de l'IA à la gestion documentaire et archivistique (OCR, extraction d'entités, recherche intelligente)
- Enjeux éthiques et biais dans l'IA
- Comment fonctionne l'IA générative ? De quelles données a-t-elle besoin ? Quels sont les modules qui interviennent ?
- Guide méthodologique : s'appuyer sur ISO 42001

Caractérisation et structuration des données documentaires : comment les rendre exploitables

- Typologie des données : structurées, semi-structurées, non structurées
- Métadonnées et standards documentaires : Dublin Core, MARC, METS, PREMIS, TEI
- Structuration et annotation des données textuelles pour l'IA
- Prétraitement des documents : OCR, reconnaissance de la mise en page, tokenisation
- Analyse et extraction automatique de métadonnées
- Le rôle majeur des bases de données vectorielles

Nettoyage et normalisation des données textuelles : préparer des données textuelles pour l'entraînement de modèles d'IA

- Techniques de nettoyage et prétraitement des textes :
  - Suppression du bruit (caractères spéciaux, formats)
  - Normalisation des textes (casse, ponctuation, stopwords, lemmatisation, stemming)
  - Détection et correction des erreurs (fautes typographiques, OCR)
- Standardisation des formats de données pour une meilleure interopérabilité : lien avec les ontologies
- Gestion des jeux de données déséquilibrés

Annotation et enrichissement des données

- Annotation manuelle vs automatique : outils et stratégies
- Reconnaissance d'entités nommées (NER) : personnes, organisations,

lieux, dates
- Techniques d'enrichissement : ontologies, lexiques, bases de connaissances (Wikidata, DBpedia)
- Introduction aux Corpus d'apprentissage supervisé
- Annotation manuelle avec un outil comme Prodigy / Brat / Label Studio
- Enrichissement d'un corpus avec des métadonnées externes

Apprendre à traiter des documents contenant texte, images et multimédia

- Rappel des contraintes légales et réglementaires sur les données, les modèles et les usages (AI Act)
- Extraction et traitement d'images à partir de documents
- Technologies de reconnaissance de texte dans les images (OCR avancé : Tesseract, Transkribus)
- Audio et transcription automatique avec Whisper
- Structuration des documents multimodaux pour l'IA générative

Constitution de datasets pour l'IA et l'Open Data : constituer des jeux de données adaptés aux modèles d'apprentissage

- Aligner les pratiques avec la gouvernance de l'intelligence artificielle et celle des données
- Critères de qualité des datasets : diversité, échantillonnage, équilibre
- Open Data et exploitation de jeux de données publics (data.gouv.fr, Europeana, Kaggle)
- Introduction aux FAIR Data Principles (Findable, Accessible, Interoperable, Reusable)
- Formatage et étiquetage des jeux de données pour entraînement IA

Parlons d'argent et d'autres coûts…

- Estimer le coût de préparation des données d'entrainement
- Estimer le coût d'entrainement des modèles
- Valoriser les coûts sous différentes dimensions (monétaire, consommation énergétique, impact climatique et impact social)

Évaluation et mise en production des modèles

- Critères d'évaluation d'un modèle NLP : précision, rappel, F1-score
- Détection des biais algorithmiques dans les modèles entraînés
- Déploiement d'un modèle en environnement documentaire
- Bonnes pratiques pour l'intégration de l'IA dans les systèmes d'information documentaire

Compétences cibles :

- Comprendre les fondamentaux de l'intelligence artificielle et ses applications en gestion documentaire
- Structurer et préparer les données documentaires pour leur exploitation par l'IA
- Nettoyer, normaliser et annoter les données textuelles pour l'entraînement des modèles d'IA
- Construire et exploiter des jeux de données pour l'intelligence artificielle
- Déployer et évaluer des modèles d'IA en environnement documentaire