Introduction au NLP (traitement automatique du langage) avec Python

5 j (35 heures) Ref : IAD003

Public

Développeurs, data scientists et professionnels du domaine de l'intelligence artificielle intéressés par l'apprentissage des bases du traitement automatique du langage naturel (NLP) avec Python

Pré-requis

Maitrise avancée de la programmation Python
Connaissance approfondie du machine learning
Connaissance de statistiques et d'algorithmique
Intérêt pour le traitement des données textuelles

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Nombreux exercices pratiques et mises en situation, échanges basés sur la pratique professionnelle des participants et du formateur, formation progressive en mode participatif. Vidéoprojecteur, support de cours fourni à chaque stagiaire

Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires
Questionnaire d'évaluation de la satisfaction en fin de stage
Auto-évaluation des acquis de la formation par les stagiaires
Attestation de fin de formation

Cette formation de 4 jours offre une introduction complète au traitement automatique du langage naturel (NLP) en utilisant le langage Python. Les participants apprendront les concepts fondamentaux du NLP ainsi que les principales techniques et outils pour analyser et traiter des données textuelles. À travers des exercices pratiques, ils découvriront comment nettoyer et préparer des corpus de textes, comment représenter les mots et documents dans un format exploitable par des algorithmes, et comment appliquer des modèles de NLP pour des tâches courantes : classification de documents, analyse de sentiments, extraction d'informations, etc. Les librairies Python spécialisées comme NLTK et SpaCy seront présentées, ainsi que des approches plus avancées basées sur le deep learning (word embeddings, réseaux de neurones récurrents). À l'issue de la formation, les participants seront capables de mettre en œuvre une chaîne complète de traitement automatique des langues sur des données réelles.

Objectifs

Programme détaillé

INTRODUCTION AU NLP ET PREPARATION DES DONNEES

PRESENTATION DU NLP ET DE SES APPLICATIONS

Définition et enjeux du NLP : Introduction aux concepts de base du traitement du langage naturel (NLP) et son importance dans diverses applications.

Exemples d'applications : chatbots, analyse de sentiments... : Exploration de cas concrets où le NLP est utilisé, comme les chatbots et l'analyse de sentiments.

PRISE EN MAIN DE PYTHON ET DE SES LIBRAIRIES POUR LE NLP

Installation de Python, NLTK, SpaCy : Guide étape par étape pour installer Python et les principales bibliothèques NLP.

Bases de la manipulation de texte en Python : Introduction aux manipulations de base de texte en utilisant Python.

PREPARATION DES DONNEES TEXTUELLES

Problématiques de format, d'encodage : Discussion sur les défis liés aux formats et à l'encodage des données textuelles.

Segmentation en phrases et en mots : Techniques pour segmenter du texte en phrases et en mots.

Nettoyage : suppression des stopwords, stemming, lemmatisation : Méthodes pour nettoyer le texte, y compris la suppression des stopwords, le stemming et la lemmatisation.

Projet pratique : nettoyage et préparation d'un corpus de textes : Exercice pratique de nettoyage et de préparation de données textuelles.

REPRESENTATION VECTORIELLE DU LANGAGE

LIMITES DES APPROCHES SYMBOLIQUES

Sacs de mots et pondération TF-IDF : Introduction aux sacs de mots et à la pondération TF-IDF pour représenter le texte.

Création de représentations vectorielles simples : Techniques pour créer des représentations vectorielles de base du texte.

Mesure de la similarité entre documents : Méthodes pour mesurer la similarité entre documents textuels.

REPRESENTATIONS DISTRIBUEES : WORD2VEC

Principe des plongements lexicaux (word embeddings) : Explication des plongements lexicaux et leur importance.

Entraînement de word2vec sur un corpus : Guide pratique pour entraîner le modèle word2vec.

Analogies et similarités entre mots : Exploration des analogies et des similarités générées par word2vec.
Projet pratique : analyse de similarité dans des descriptions de films : Exercice pratique d'analyse de similarité en utilisant des descriptions de films.

TACHES DE CLASSIFICATION DE TEXTES

CLASSIFICATION DE DOCUMENTS

Représentation par sacs de mots : Utilisation des sacs de mots pour représenter des documents.
Entraînement de classifieurs (Naive Bayes, régression logistique) : Entraînement de modèles de classification comme Naive Bayes et la régression logistique.
Évaluation des performances : Techniques pour évaluer les performances des modèles de classification.

ANALYSE DE SENTIMENTS

Approches lexicales et par apprentissage : Introduction aux approches lexicales et basées sur l'apprentissage pour l'analyse de sentiments.
Détection de la polarité et des émotions : Méthodes pour détecter la polarité et les émotions dans le texte.
Projet pratique : classification de critiques de films positives/négatives : Exercice pratique de classification des critiques de films en positives ou négatives.

EXTRACTION D'INFORMATIONS

RECONNAISSANCE D'ENTITES NOMMEES

Détection de personnes, lieux, organisations... : Techniques pour détecter les entités nommées dans le texte.
Utilisation des librairies NLTK et SpaCy : Utilisation des bibliothèques NLTK et SpaCy pour la reconnaissance des entités nommées.

EXTRACTION DE RELATIONS ENTRE ENTITES

Patrons syntaxiques simples : Utilisation de patrons syntaxiques pour extraire des relations entre entités.
Visualisation de graphes de relations : Techniques pour visualiser les relations entre entités sous forme de graphes.
Projet pratique : extraction d'informations à partir de descriptions de produits : Exercice pratique d'extraction d'informations à partir de descriptions de produits.

PROJET FINAL ET INTRODUCTION AUX MODELES AVANCES

PROJET FIL ROUGE : ANALYSE DE SENTIMENTS SUR DES TWEETS

Collecte de données sur Twitter : Méthodes pour collecter des données sur Twitter.

Préparation des données et extraction de features : Préparation des données et extraction des caractéristiques nécessaires.
Entraînement de différents classifieurs : Entraînement de divers modèles de classification.
Interprétation des résultats : Techniques pour interpréter les résultats obtenus.

PRESENTATION DES MODELES AVANCES

Réseaux de neurones récurrents (RNN, LSTM) : Introduction aux réseaux de neurones récurrents et aux LSTM.
Transformers et modèles pré-entraînés (BERT) : Présentation des transformers et des modèles pré-entraînés comme BERT.
Application au résumé automatique et au question-answering : Exploration des applications avancées telles que le résumé automatique et le question-answering.

CONCLUSION ET PERSPECTIVES

Synthèse des concepts clés : Récapitulatif des principaux concepts couverts pendant la formation.
Ressources pour aller plus loin : Suggestions de ressources pour approfondir les connaissances en NLP.