

Traitement automatique du langage par deep learning (Expert)

Durée

3 jours.

Participants

Analystes des données, ingénieurs ML.

Prérequis

Bonne connaissance et pratique du Machine Learning avec Python.

Description

Formation spécialisée dans le traitement automatique du langage naturel.

Objectifs pédagogiques

Acquérir les fondamentaux de l'état de l'art en traitement automatique du langage naturel.

Travaux pratiques

Multiples cas d'utilisation avec Keras.

Programme

Concepts fondamentaux d'un réseau de neurones

- Réseau de neurones : formalisme, fonctions d'activations ;
- Apprentissage d'un réseau de neurones : fonctions de coût, SGD, Adam ;
- Initialisation et régularisation : orthogonalité à l'initialisation, régularisations L1/L2, politiques de batchs, dropout.

Traitement automatique du langage naturel

- Représentation sac de mots, TFIDF ;
- Prétraitements usuels (étiquetage morphosyntaxique, lemmatisation, etc) ;
- Dictionnaire de mots et n-gram ;
- Représentation one-hot et word embeddings.
- Utilisation du texte dans le cadre de signaux hétérogènes.

Réseaux récurrents (RNNs)

- Présentation des RNNs : principes fondamentaux et applications ;
- Variantes modernes : LSTMs, GRUs, réseaux bi-directionnels ;
- Éléments techniques de gestion des RNNs : padding, taille des séquences ;
- Mécanisme de décodage pour produire une séquence.

Modèles à attention

- Attention simple intégrée à un RNN ;
- Modèles transformeurs ;
- Architectures encodeur-décodeur ou à base de masque ;
- Variantes modernes : Reformer, Performer, XLNet, etc.

Cas d'utilisation du traitement du langage moderne

- Classification ;
- Question/réponse, sélection de texte pertinent et compréhension de langage ;
- Génération conditionnelle ;
- Transfert & multitâche (SuperGLUE) ;
- Agents conversationnels avec DialogFlow ;
- Détection de thématiques avec BigARTM.

Questions éthiques

- Perpétuation des biais des données ;
- Représentation des sous-populations dans les données ;
- Intoxication des conversations publiques avec des données générées ;
- Utilisation des données utilisateurs (RGPD & CNIL).