Stage pratique de 3 jour(s)
Réf : MMD

Participants

Ingénieurs/chefs de projet IA, consultants IA et toute personne souhaitant découvrir le Text Mining pour le Machine Learning et le Deep Learning.

Pré-requis

Bonnes connaissances en statistiques. Bonnes connaissances du Machine Learning et du Deep Learning. Expérience requise.

Prix 2020 : 2040€ HT

Dates des sessions

PARIS

02 juin 2020, 14 sep. 2020
16 nov. 2020

Modalités d'évaluation

L'évaluation des acquis se fait tout au long de la session au travers des multiples exercices à réaliser (50 à 70% du temps).

Compétences du formateur

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

Moyens pédagogiques et techniques

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.

• A l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui

Text Mining par la pratique

Machine Learning et Deep Learning pour les données textuelles s'inscrivent dans le cadre du traitement statistique et de la valorisation des données dans tout projet Big Data. Ce cours pratique vous présentera toute la chaîne de conception appliquée au Machine Learning dans un contexte Big Data batch et streaming.

OBJECTIFS PEDAGOGIQUES

Comprendre les méthodes de la statistique textuelle
Mettre en œuvre l'extraction des caractéristiques de données textuelles
Créer des sélections et des classements dans de grands volumes de données textuelles
Choisir un algorithme de classification
Évaluer les performances prédictives d'un algorithme

1) Les approches traditionnelles en Text Mining
2) Feature Engineering pour la représentation de texte
3) La similarité des textes et classification non supervisée
4) La classification supervisée du texte
5) Natural Language Processing et Deep Learning

1) Les approches traditionnelles en Text Mining

- Les API pour récupérer des données textuelles.
- La préparation des données textuelles en fonction de la problématique.
- La récupération et l'exploration du corpus de textes.
- La suppression des caractères accentués et spéciaux.
- Stemming, Lemmatization et suppression des mots de liaison.
- Tout rassembler pour nettoyer et normaliser les données.

Travaux pratiques
La recherche des documents, la préparation, la transformation et la vectorisation des données en DataFrame.

2) Feature Engineering pour la représentation de texte

- Comprendre la syntaxe et la structure du texte.
- Le modèle Bag of Words et Bag of N-Grams.
- Le modèle TF-IDF, Transformer et Vectorizer.
- Le modèle Word2Vec et l'implémentation avec Gensim.
- Le modèle GloVe.
- Le modèle FastText.

Travaux pratiques
Mise en place des opérations d'extraction des caractéristiques de données textuelles afin d'effectuer des classifications.

3) La similarité des textes et classification non supervisée

- Les concepts essentiels de similarité.
- Analyse de la similarité des termes : distances Hamming, Manhattan, Euclidienne et Levenshtein.
- Analyse de la similarité des documents.
- Okapi BM25 et le palmarès de classement.
- Les algorithmes de classification non supervisée.

Travaux pratiques
Construire un système de recommandation des produits similaires sur la base de la description et du contenu des produits que vous avez choisi.

4) La classification supervisée du texte

- Prétraitement et normalisation des données.
- Modèles de classification.
- Multinomial Naïve Bayes.
- Régression logistique. Support Vector Machines.
- Random Forest. Gradient Boosting Machines.
- Évaluation des modèles de classification.

Travaux pratiques
Mise en œuvre des classifications supervisées sur plusieurs jeux de données.

5) Natural Language Processing et Deep Learning

- Les librairies NLP : NLTK, TextBlob, SpaCy, Gensim, Pattern, Stanford CoreNLP.
- Les librairies Deep Learning : Theano, TensorFlow, Keras.
- Natural Language Processing et Recurrent Neural Networks.

est ensuite analysé par nos équipes pédagogiques.

• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

- RNN et Long Short-Term Memory. Les modèles bidirectionnels RNN.
- Les modèles Sequence-to-Sequence.
- Questions et réponses avec les modèles RNN.

Travaux pratiques
Construire un RNN pour générer un nouveau texte.