Date de publication : 02/10/2025

Formation : Text Mining par la pratique

Formation pratique - 3j - 21h00 - Réf. MMD
Prix : 2010 € H.T.

Le data mining restreint aux données textuelles - le text mining - est de plus en plus utilisé dans les entreprises. Il permet, par exemple, de classer des produits à partir des commentaires des consommateurs. Vous mettrez en œuvre les algorithmes et les outils du text mining sur des exemples paradigmatiques.

Objectifs pédagogiques

À l'issue de la formation, le participant sera en mesure de :

- 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

Public concerné

Ingénieurs/chefs de projet IA, consultants IA et toute personne souhaitant découvrir le text mining pour le machine learning et le deep learning.

Prérequis

Bonnes connaissances en statistiques. Bonnes connaissances du machine learning et du deep learning. Expérience requise.

Vérifiez que vous avez les prérequis nécessaires pour profiter pleinement de cette formation en faisant ce test.

Modalités d'évaluation

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…

Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

Programme de la formation

PARTICIPANTS

Ingénieurs/chefs de projet IA, consultants IA et toute personne souhaitant découvrir le text mining pour le machine learning et le deep learning.

PRÉREQUIS

Bonnes connaissances en statistiques. Bonnes connaissances du machine learning et du deep learning. Expérience requise.

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.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

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.

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 formations pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque formation ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui 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 participant a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Pour toute question ou besoin relatif à l'accessibilité, vous pouvez joindre notre équipe PSH par e-mail à l'adresse psh-accueil@orsys.fr.

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.
- 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.

Dates et lieux

CLASSE À DISTANCE
2026 : 1 juin, 12 oct.

PARIS LA DÉFENSE
2026 : 1 juin, 12 oct.