Text Mining par la pratique

Cours Pratique de 3 jours
Réf : MMD - Prix 2021 : 2 240€ HT

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 l'apprenant 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

LE PROGRAMME

dernière mise à jour : 06/2021

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.

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

Les apports théoriques et les panoramas des techniques et outils ne nécessitent pas d'avoir recours à une évaluation des acquis.

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.
• À l'issue de chaque stage 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 stagiaire a bien assisté à la totalité de la session.

- 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.
- 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.

LES DATES

PARIS
2021 : 19 juil., 18 oct.

CLASSES À DISTANCE
2021 : 19 juil., 18 oct.