Introduction au data mining

4 j (28 heures) Ref : BDMI

Public

Développeurs, statisticiens et Business Analysts

Pré-requis

Connaissance d'un langage de script (Python ou R)
Connaissances de bases en statistiques et/ou méthodes numériques
Expérience du shell sous Linux

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois
Un poste par stagiaire, 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

Objectifs

- Appréhender les différentes facettes du métier de Data Scientist
- Appréhender la collecte de données, l'identification de mesures aberrantes
- Apréhender l'analyse de texte, les modèles prédictifs...
- Mener des analyses exploratoires pour identifier des opportunités de service
- Choisir des supports visuels à fort impact pour communiquer vos résultats
- Connecter les différences sources de données à un entrepôt de données
- Croiser les différences sources de données avec des sources externes
- Tester les algorithmes sur des sous-ensembles de données
- Exploiter Hadoop et les plateformes de calcul distribué
- Représenter les résultats de façon graphique et concise

Programme détaillé

LE DATA SCIENTIST ET SON ROLE DANS L'ENTREPRISE

Fiche d'identité et profils chez LinkedIn, Amazon, Facebook...

Les compétences recherchées

LA BOITE A OUTILS DU DATA SCIENTIST

Langages de script : R, Python

Langages compilés: C/C++, Java/Clojure

Plateformes et frameworks: Hadoop, Mahout, Weka, Orange

TYPOLOGIE DE DONNEES

Données structurées et non structurées

Documents texte, emails, logs

Séries temporelles, données spatiales

Transactions (e-commerce, banque)

Télécoms et données d'appel

TP : collecte de données web publiques

L'ANALYSE EXPLORATOIRE

Qualifier les données

Détecter les tendances, patterns récurrents et anomalies

Gérer les outliers

Versionner son code

Comment organiser une chaîne de traitement : Make, Camel

TP : mise en place d'une chaîne évolutive de traitement de données

ALGORITHMES

Clustering

Corrélation et Frequent Itemset

Classification et prédiction

Analyse de séquences, filtrage et modèles de Markov

Méthodes d'ensemble

PLATEFORMES ET ENVIRONNEMENTS POUR LA FOUILLE DE DONNEES

Pig/Hive et Mahout

Introduction à Weka

Python et Numpy, Scipy

R

APPLICATIONS

Moteurs de recommandation

Optimisation d'allocation de ressources

Identification d'anomalies

TP : conception d'un moteur de recommandation d'articles web

MISE EN OEUVRE

Validation d'un modèle – jeux d'apprentissage, test et courbes ROC

Déploiement : l'atout "DevOps"

Passage à l'échelle: l'avantage MapReduce

Intégration à Hadoop

Visualisation de jeux de données massives

Publier via une IHM web: D3.js

PENSER LES IMPACTS SOCIAUX

Effets indirects d'une approche orientée données

La CNIL, devoirs d'éthique et le respect de la vie privée