GED, optimiser la recherche et l'indexation des contenus non structurés

★★★★★

MR-95 **2 Jours (14 Heures)**

Description

Ce cours vous exposera les méthodes utilisées pour organiser et optimiser l'exploitation de ressources textuelles non structurées. Vous apprendrez à les catégoriser, à les marquer automatiquement ou à les rendre visibles des moteurs de recherche en utilisant des outils comme Apache Solr ou Mahout.

À qui s'adresse cette formation ?

Pour qui

Chefs de projet, administrateurs GED, développeurs, archivistes, documentalistes.

Prérequis

Aucune

Les objectifs de la formation

- Comprendre les enjeux de l'exploitation des ressources textuelles non structurées
- Identifier les composants et les étapes du cycle de traitement des contenus
- Classifier, catégoriser, marquer automatiquement les contenus

Programme de la formation

Les enjeux de l'exploitation des contenus non structurés

- Pourquoi le traitement des ressources textuelles est un enjeu stratégique ? Les particularités du traitement des contenus non structurés.
- Exploiter les ressources textuelles : créer de la valeur à partir du chaos.
- Présentation de la plateforme logicielle utilisée pendant la formation.
- Travaux pratiques Faire une recherche dans un courriel donné en exemple et en extraire un paragraphe particulier.
- Lister tous les mots du paragraphe et afficher les noms des personnes citées.

Composants et étapes du cycle de traitement des contenus non structurés

- Les catégories grammaticales de base.
- Le système morphologique : racine, préfixe, suffixe.
- L'identification des unités lexicales (tokenization).
- La détection des limites de phrase.
- Travaux pratiques Extraire les phrases d'un article de journal, en lister les mots.
- Présenter chaque nom sous forme singulier/pluriel.

Classifier, catégoriser, marquer automatiquement les contenus

- Regrouper les résultats de recherche avec Carrot2.
- Regrouper des collections de documents avec Apache Mahout.
- Catégoriser des documents avec Apache Lucene.
- Rechercher des contenus sémantiques à l'aide de Falcons.
- Travaux pratiques Utiliser la classification automatique d'un corpus de documents pour proposer le plan de classement d'une application de GED.

Opérations avancées sur les contenus

- Accéder aux contenus des différents formats de fichier.
- Extraire du contenu de différents formats de fichier à l'aide d'Apache Tika.
- Analyser les contextes pour résoudre des ambiguïtés.
- Utiliser les graphes pour modéliser l'information syntaxique et sémantique des contenus non structurés.
- Travaux pratiques A partir d'un contenu fourni, identifier les unités ambiguës.
- Lister les contextes d'apparition des différentes unités ambiguës.
- Proposer une stratégie de résolution.

Préparer les ressources non structurées pour les moteurs de recherche

- Les différentes techniques de recherche.
- Les concepts associés à la recherche : indexation, interface, classement des résultats, présentation des résultats.
- Exemple de recherche par facettes : Amazon.
- com.
- Exemple d'utilisation du serveur de recherche Apache Solr.
- Travaux pratiques Extraire et indexer le contenu d'un article de journal à l'aide d'Apache Solr.
- Etablir un jeu de test pour évaluer la performance du système d'indexation.

Introduction à Prism V6

- Qu'est-ce que Prism ? Principaux concepts.
- Les problématiques que résout le framework et celles qu'il ne résout pas.
- L'architecture Prism.
- Les différents modules Prism.
- Les objectifs de chaque module.
- Les Quick Starts et les fonctionnalités de Prism.
- L'accès à la documentation Prism.
- Démonstration Démonstrations des différents Quick Starts illustrant les fonctionnalités de Prism.