Stage pratique de 2 jour(s)
Réf : GED

Participants

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

Pré-requis

Connaissances de base en gestion de contenu.

Dates des sessions

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 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.

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

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.

OBJECTIFS PEDAGOGIQUES

Comprendre les enjeux de l'exploitation des ressources textuelles non structurées
Identifier les composants et les étapes du cycle de traitement des contenus
Préparer les contenus en vue de leur exploitation par les moteurs de recherche
Classifier, catégoriser, marquer automatiquement les contenus

Méthodes pédagogiques

Exposé des concepts et des principes suivi de démonstrations et d'exemples pratiques.

1) 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.

2) 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.

3) 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.

4) 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.

5) 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.