EDM, optimize search and indexing of unstructured content

★★★★★

MR-95 **2 Days (14 Hours)**

Description

This course will expose you to the methods used to organize and optimize the use of unstructured textual resources. You will learn how to categorize them, automatically tag them or make them visible to search engines using tools like Apache Solr or Mahout.

Who is this training for ?

For whom

Project managers, GED administrators, developers, archivists, librarians.

Prerequisites

Aucune

Training objectives

- Understand the challenges of using unstructured textual resources
- Identify the components and stages of the content processing cycle
- Classify, categorize, automatically mark content

Training program

Les enjeux de l'exploitation des contenus non structurés

- Why the processing of textual resources is a strategic issue? The particularities of the processing of unstructured content.
- Exploiting textual resources: creating value from chaos.
- Presentation of the software platform used during the training.
- Practical work Search in an email given as an example and extract a particular paragraph.
- List all the words in the paragraph and display the names of the people cited.

Composants et étapes du cycle de traitement des contenus non structurés

- The basic grammatical categories.
- The morphological system: root, prefix, suffix.
- The identification of lexical units (tokenization).
- Detecting sentence boundaries.
- Practical work Extract the sentences from a newspaper article, list the words.
- Present each noun in singular form /plural.

Classifier, catégoriser, marquer automatiquement les contenus

- Group search results with Carrot2.
- Group document collections with Apache Mahout.
- Categorize documents with Apache Lucene.
- Search for semantic content using Falcons.
- Practical work Use the automatic classification of a corpus of documents to propose the classification plan for an EDM application.

Opérations avancées sur les contenus

- Access contents of different file formats.
- Extract contents of different file formats using Apache Tika.
- Analyze contexts to resolve ambiguities.
- Use graphs to model the syntactic and semantic information of unstructured content.
- Practical work From provided content, identify the units ambiguous.
- List the contexts in which the different ambiguous units appear.
- Propose a resolution strategy.

Préparer les ressources non structurées pour les moteurs de recherche

- The different search techniques.
- The concepts associated with search: indexing, interface, classification of results, presentation of results.
- Example of search by facets : Amazon.
- com.
- Example of using the Apache Solr search server.
- Practical work Extract and index the content of a journal article using Apache Solr.
- Establish a test set to evaluate the performance of the indexing system.

Introduction à Prism V6

- What is Prism? Main concepts.
- The problems that the framework solves and those that it does not solve.
- The Prism architecture .
- The different Prism modules.
- The objectives of each module.
- Quick Starts and Prism functionalities.
- L 'access to the Prism documentation.
- Demonstration Demonstrations of the different Quick Starts illustrating the functionalities of Prism.