TP LO17 : Indexation sémantique

Objectifs

Découvrir les notions d'ontologie et de base de connaissance RDF.
Découvrir un langage de requêtes sur une base de connaissance RDF.
Associer indexation sémantique et indexation plein texte.

Documents (accessibles sur le site lo17)

cours/lo17-rdf.pdf :	support de cours
td/lo17-td-rdf.pdf :	énoncé du td
td/lo17-td-rdf.zip :	projet Eclipse

Documents (accessibles dans le projet Eclipse)

documents/lo17-ontology.n3 :	ontologie lo17 au format N3
documents/kblo17.owl :	base de connaissance documentaire lo17 au format OWL
src/main/tplo17main.java :	à compléter durant le tp. Contient toutes les méthodes utiles.

Librairies java nécessaires pour le tp. On utilise la version 2.7.4

jena-arq-2.9.4.jar ; jena-core-2.7.4.jar ; jena-iri-0.9.4.jar ;
jena-larq-1.0.0-incubating.jar; lucene-core-3.3.0.jar ;
log4j-1.2.16.jar ; slf4j-api-1.6.4.jar ; slf4j-log4j12-1.6.4.jar; xercesImpl-2.10.0.jar; xml-apis-1.4.01.jar
Ces librairies se trouvent dans le répertoire lib.

Aide

Cours LO17 :	cours/lo17-index.pdf
Jena Apache :	http://jena.apache.org/
Documentation :	http://jena.apache.org/documentation/
Langage SPARQL **:**	http://www.w3.org/TR/rdf-sparql-query/

Créer un répertoire servant de workspace pour Eclipse et importer le projet java (tdlo17.zip). Il possède un sous-répertoire documents, contenant l'ontologie et la base de connaissance, et un sous-répertoire lib contenant les librairies java nécessaires. Le fichier source main/tplo17main.java est à compléter. Il suffira d'écrire les requêtes SPARQL permettant de répondre aux questions et d'appeler les méthodes statiques nécessaires figurant dans le fichier source.

Question 1 (Ontologie)

Compléter la méthode q1() qui implémente le parsing de l'ontologie (fichier lo17-ontology.n3) pour exécuter une requête SPARQL SELECT à partir d'un fichier ou d'une chaîne (utiliser readFileQuery ou readStringQuery selon le cas). Ecrire trois requêtes pour retrouver :

a) Les concepts de l'ontologie (3)
b) Les relations de l'ontologie (1)
c) Les attributs de l'ontologie (11)

Question 2 (ne nécessite que SPARQL)

Compléter la méthode q2() qui implémente le parsing de la base de connaissance (fichier kblo17.owl). Ecrire une requête SPARQL pour :

1. Rechercher les titres des bulletins.
2. Rechercher les rubriques des bulletins. Montrer qu'il devrait y en avoir 10 au lieu de 18.
3. Savoir si des bulletins ont des titres contenant le sigle CNRS ?
4. Afficher les rubriques des bulletins de 2011. Quelles sont les deux rubriques qui ne figurent pas cette année là ?
5. Afficher les numéros des bulletins ayant au moins trois images et les légendes de ces images.

Question 3 (ne nécessite que SPARQL)

1. Combien de bulletins ont des contacts au CNRS ou au CEA ? (rep = 39 ; il suffira de compter les résultats retournés)
2. Quelles sont les trois rubriques les plus utilisées ? On devra utiliser les caractéristiques COUNT et GROUP BY de SPARQL.

Question 4 (Nécessite LARQ)

Créer un index LARQ des chaînes de la base de connaissance à l'aide de la méthode getWholeStringIndex.

1. En utilisant cet index, rechercher les bulletins dont le texte contient le terme « capteurs ». Etudier l'influence du facteur seuil de similarité (ranking) sur le nombre de résultats.
2. Rechercher les contacts des bulletins traitant de capteurs non infrarouges, et qui n'ont pas de lien avec l'Internet. Pour un ranking de 0.5 combien en reste-il ?
3. Rechercher les bulletins qui traitent de création. Est-ce que le mot création est nécessairement dans le titre ? Ecrire une requête unique qui permet de rechercher la présence de ce mot dans le titre ou le texte. On utilisera le seuil de similarité de 0.3. Afficher les titres pour vérifier. Quel serait le problème avec l'utilisation d'un index lié à la valeur d'une propriété ?
4. Rechercher les titres des bulletins traitant de « Création d'une chaire ».