Titre :

Sujet de Travail d'Étude et de Recherche – Développement d'un Web service parseur de documents structurés en Java

Information :

Encadrants :	Clement Jonquet (LIRMM, UM) – jonquet@lirmm.fr
Spécialités :	Licence Info, Master DECOL, AIGLE, autres
Nombre d'étudiants :	2-3
Contexte:	Projet SIFR (www.lirmm.fr/sifr)
Ou:	LIRMM, Montpellier
Quand:	2nd semestre 2018-2019

Mots clés :

Application web, web service (REST), technologies web, parseur de documents structurés (PDF, DOC, XLS, HTML, XML) web sémantique, annotation sémantique.

Technologies :

Java/JEE, REST, ResTful web services, XML/JSON, Tomcat, pour le Web service. Éventuellement : RubyOnRails, web client technologies (HTML5, JavaScript, CSS), pour l'interface graphique Web.

Résumé :

Ce TER consiste à concevoir et implémenter un web service REST (en Java et hébergé dans un serveur d'application Tomcat) et éventuellement une petite application web (en RubyOnRails et hébergée au sein d'une plateforme Web existante) dont l'objectif est de parser des documents structurés (PDF, DOC, XLS, HTML, XML) pour en extraire certains champs texte qui seront envoyés à un service d'annotation existant.

Par exemple, pour un PDF donné, nous souhaitons extraire automatiquement des sections, (e.g., <titre> <abstract> <keywords>) les envoyer chacune séparément à un web service d'annotation de texte existant ; puis agréger les résultats pour l'ensemble du document. Il s'agira de développer cette fonctionnalité dans un web service « proxy » (en Java) qui traitera les documents structurés avant de les envoyer au web service d'origine. Une nouvelle vue web, modifiant celle déjà existante, sera proposée pour servir d'interface utilisateur.

Description détaillée :

Le service que nous souhaitons interroger est un service d'annotation sémantique avec des ontologies, le SIFR Annotator, accessible a :

Interface utilisateur : http://bioportal.lirmm.fr/annotator
API REST : http://data.bioportal.lirmm.fr/documentation#nav_annotator

Pour le moment, le SIFR Annotator n'accepte que du texte en entrée, et il renvoie l'ensemble des concepts d'ontologies biomédicales qui sont présents dans le texte. Nous voudrions que ce service traite des documents structurés en entrée et ainsi soit capable d'automatiquement extraire et annoter des parties différentes de documents et en faire la synthèse. En plus de permettre de traiter directement des documents structurés, cela permettrait également, lors de la synthèse des annotations, d'affecter plus ou moins d'importance a une annotation suivant sa section d'origine. Par exemple, une annotation faite à partir du titre d'un article scientifique en PDF pourrait avoir plus d'importance qu'une annotation faite dans le corps de ce même article.

Les documents que nous souhaitons traiter seront dans des formats hétérogènes tels que PDF, XLS, DOC, HTML ou XML. Et surtout, ils auront des structures/sections différentes. Il s'agira ainsi de

développer un mécanisme de configuration de template, qu'un utilisateur pourra définir, précisant le template à appliquer lors de l'annotation d'un document donné. Par exemple, l'utilisation du template « article en .pdf » indiquera au nouveau web service qu'il doit identifier et annoter les sections titre/abstract/keywords dans le PDF ; tandis que le template « données XYZ en .xls » indiquera qu'il s'agit d'annoter les colonnes X, Y et Z d'un tableau Excel.

Le nouveau web service sera développé en Java et hébergé dans un serveur Tomcat au sein d'une application « proxy » déjà disponible (https://github.com/agroportal/ncboproxy). Il s'agira de venir développer des nouvelles fonctionnalités de parsing dans cette application. De façon similaire, nous avons déjà une interface graphique (en RubyOnRails) qu'il s'agira, si le temps le permet, d'adapter aux appels du nouveau web service.

Des exemples de document structurés à traiter seront fournis.