PROJET
De la recherche d'information au Web Sémantique

1. Présentation

Le projet consiste à mettre en place un moteur de recherche. Vous aurez au cours de l'UE plusieurs versions de ce moteur à livrer.

Nous vous recommandons de développer le cœur du système en Java et de stocker les index dans une base de données POstGreSQL. Vous pouvez également rechercher des modules existants et les intégrer dans la mesure où vous respectez les étapes détaillées dans la section suivante.

Le projet s'effectuera en binômes.

2. Etapes du projet

Le projet est décomposé en plusieurs étapes :
Partie 1 : implémentation du système de recherche d'information (SRI) classique
- sous étape 1-a : implémentation du module d'indexation
- sous étape 1-b : implémentation du module de recherche
- sous étape 1-c : évaluation de votre système

Partie 2 : utilisation d'une ontologie de domaine pour la reformulation de requêtes dans un SRI
Partie 3 : interrogation de triplets RDF à partir de SPARQL
Partie 4 : alignement de bases de connaissances pour l'interrogation du LOD

Les parties 2 à 4 seront détaillées ultérieurement.

Les sous-étapes de la partie 1 sont détaillées dans la suite du document. Elles s'étaleront sur les 2 premières périodes de l'année

3. Ressources à votre disposition

3. a Collection de documents

Il s'agit de documents en Français contenant des documents HTML en lien avec le cinéma extraits du Web pendant l'été 2014. Les documents n'ont subi aucun traitement et sont donc représentatifs des documents disponibles actuellement sur le Web..
La collection est composée de 138 documents et on trouvera un exemple de document en annexe 2.
La collection est téléchargeable : http://mass-cara2.univ-tlse2.fr/mathinfo/~hernandez/CORPUS.zip

3.b Collection de requêtes

Un ensemble de requêtes est à votre disposition pour évaluer votre système. Ces requêtes sont présentées dans l'annexe 2 et le fichier est téléchargeable : mass-cara2.univ-tlse2.fr/mathinfo/~hernandez/requetes.html.

Pour chacune d'entre-elles nous vous fournissons également les jugements de pertinence, ie les documents de la collection qui ont été jugés pertinents par des personnes pour chacune des requêtes. Les résultats sont mis dans un fichier appelé « qrelQXX.txt » où XX correspond au numéro de requêtes définies dans le fichier requetes.html. On trouvera un exemple de fichier qrelQ2.txt en annexe 3.
Les fichiers qrel sont disponibles à http://mass-cara2.univ-tlse2.fr/mathinfo/~hernandez/qrels.zip
Attention, ce n'est pas parce que les termes de la requête ne sont pas présents dans un élément que cet élément est non pertinent. Les jugements peuvent être subjectifs.

4. Détail des étapes de la partie 1

Lors de cette première partie vous allez réaliser un moteur purement syntaxique. Notez que nous vous demandons de réaliser une version de base pour chacune des sous-étapes suivantes. Vous devrez ensuite proposer des améliorations de cette version après avoir évaluer votre moteur (voir détail partie 1-c). Nous

attendons à ce que vous aillez une analyse critique de votre approche et à ce que vous nous livriez en fin d'UE le moteur le plus efficace possible.

Lors de l'étape 1-a, vous allez mettre en place les différents mécanismes relatifs à l'indexation. *Attention de bien réfléchir à la structure de vos index et à la façon dont vous allez lier l'information textuelle et l'information structurelle, ainsi qu'à l'algorithme de recherche que vous allez utiliser dans l'étape suivante.*

Il s'agit de :

- parcourir les documents HTML avec un parseur (nous vous recommandons l'utiliser le parseur JSOUP[1])
- récupération des informations nécessaires aux index
    - information textuelle
        - reconnaître les mots dans une séquence de lettres ou des symboles composant l'élément. On considère que les espaces et toutes les ponctuations constituent un séparateur de mots.
        - nettoyer les mots composant les passages à partir d'une stopliste dont on trouvera un exemple à l'adresse suivante : http://www.irit.fr/~Nathalie.Hernandez/M2ICE/stopliste.txt
        - implémenter les tables dans la base de données

Dans le cadre d'amélioration possible de cette partie, vous pourrez :

- prendre en considération des éléments de la structure (title, balises h1, h2, em, b, ..., meta)
- prendre en compte la distance entre les mots composant un terme (gestion des expressions à partir des mots consécutifs ou proches)

Lors de l'étape 1-b, vous allez devoir implémenter le module permettant de retrouver pour une requête mots clés donnée la liste ordonnée des documents pertinents. Pour simplifier l'évaluation, nous vous recommandons de prendre en entrée du module la liste des requêtes fournies et de générer en sortie pour chaque requête la liste des documents pertinents dans un format qui pourra être comparable au qrel fourni pour la requête (voir partie 1-c).

Vous devrez :

- choisir un modèle de pondération des paragraphes. Pour la première version de ce module vous pouvez ne considérer que *tf*. Par la suite, vous pourrez apporter des évolutions en considérants *idf* et éventuellement l'information portée par les autres éléments des documents.
- réaliser le processus de recherche en vous fondant sur les résultats d'indexation que vous avez produits à l'etape 1-a, le modèle de pondération choisi, le modèle de RI choisi (nous vous recommandons pour la première version de choisir le modèle vectoriel), les mots clés de la requête.

L'étape 1-c consiste à évaluer les performances de votre système en le comparant pour chaque requête à ce que des humains ont jugé pertinent. L'évaluation sera effectuée sur toutes les requêtes fournies. Vous devrez pour cela programmer un 'évaluateur' qui prendra en entrée un fichier de type *qrel* et un fichier lisant les résultats de votre système. L'évaluateur proposera en sortie les **précisions à 5, 10 et 25 éléments pour chaque requête**, ainsi que les **précisions moyenne à 5, 10 et 25 éléments pour toutes les requêtes**. Il pourra également proposer les courbes interpolées rappel/précision.

Afin que vous puissiez comparer vos résultats à ceux obtenus par les autres groupes, nous vous proposons de partager dans un document les précisions que vous obtenez. Ce document est disponible `https://docs.google.com/spreadsheets/d/1ZMSVyDAHl84OC836xPSSRa9Fu7_k04Xq-4khCfkVTpE/edit?usp=sharing`.

[1] http://jsoup.org/

Annexe 1 : Documentation

Documentation parseurs HTML : *http://jsoup.org/*

Documentation JDBC :
http://www.dil.univ-mrs.fr/~massat/ens/java/jdbc.html
http://jguillard.developpez.com/JDBC/

Documentation PostgreSQL : http://www.postgresql.org/

Documentation JDBC PostgreSQL *: http://jdbc.postgresql.org/*

Et le site du W3C pour en savoir plus….

Annexe 2 : Exemple de document HTML de la collection

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="fr">
<head>
<title>Charlotte Gainsbourg - Actualités, potins et informations - Actu-C4.fr</title>
<meta name="description" content="Actualités, films, potins de stars sur Charlotte Gainsbourg,
sur Actu-C4.fr actualités des célébrités." />
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link rel="icon" href="style/bombe.png" type="image/png" />
<link rel="stylesheet" type="text/css" href="style/actu.css" />
<link rel="alternate" type="application/atom+xml" title="Actualités des stars - Actu-C4.fr"
href="rss/"/>
<script type="text/javascript" src="javascript/actu.js" defer="defer"></script>
<link rel="alternate stylesheet" type="text/css" href="style/actu-mobile.css" media="screen"
title="Mobile" id="mobile" />
<style type="text/css"> p.autres { background:#F7F1F8; } p.autres a { color: #718; } p.autres
strong.titre { background-color: #c9c } </style>
</head>

<body>
<h1>Charlotte Gainsbourg - Actualités et films</h1>
<div id="article" style="border-top-color:#c9c">
<div id="b1" style="text-align:center">
<script type="text/javascript">//<![CDATA[
google_ad_client = "ca-pub-7664170618954789"; google_ad_slot = "1901195353";
google_ad_width = 728; google_ad_height = 90;
google_color_link = "771188";
//]]>
</script>
<script type="text/javascript"
src="//pagead2.googlesyndication.com/pagead/show_ads.js"></script>
</div>
<img src="http://cine.kiao.net/images/portraits/1418.jpeg" alt="Charlotte Gainsbourg"
class="photo" />
<p>Charlotte Gainsbourg, née le 21 juillet 1971 à Londres, est une actrice et
chanteuse française.</p>
<p>Fille de Serge Gainsbourg et de Jane Birkin, elle commença sa carrière cinématographique
très tôt, poussée par sa mère. Elle est mariée avec Yvan Attal, lui-même comédien et
réalisateur. Ils ont deux enfants, Ben et Alice Jane. Charlotte Gainsbourg a également deux
demi-sœurs, Kate Barry et Lou Doillon du côté de sa mère, un demi-frère, Paul, une autre demi-
sœur, Natacha, ainsi qu'un petit demi-frère Lucien, mieux connu sous le diminutif
« Lulu », fils de Bambou, du côté de son père.</p>
...
<p class="i" style="border-top-color:#c9b">
<a href="a84822-omar-sy-va-jouer-au-cinema-avec" rel="bookmark"
style="color:#813"><strong>Omar Sy va jouer au cinéma avec Charlotte
Gainsbourg</strong></a><br />
…
<p>Source : <a href="http://actu-c4.fr/p1418-charlotte-gainsbourg"> http://actu-c4.fr/p1418-
charlotte-gainsbourg</a> téléchargé le 14/09/2014</p>

</body>
</html>
```

Annexe 2 : liste des requêtes

Q1

mots clés : personnes, Intouchables

description : Quelles sont les personnes impliquées dans le film Intouchables?

Q2

mots clés : lieu, naissance, Omar Sy

description : Quel est le lieu de naissance d'Omar Sy?

Q3

mots clés :personnes, récompensées, Intouchables

description : Qui a até récompensé pour Intouchables?

Q4

mots clés : palmarès, Globes de Cristal, 2012

description : Quel est le palmarès des Globes de Cristal 2012?

Q5

mots clés : membre, jury, Globes de Cristal, 2012

description : Quels sont les membres du jury du Globes de Cristal 2012?

Q6

mots clés : prix, Omar Sy, Globes de Cristal, 2012

description : Quels prix ont été décernés à Omar Sy aux Globes de Cristal 2012?

Q7

mots clés : lieu, Globes Cristal, 2012

description : Où a eu lieu les Globes de Cristal 2012?

Q8

mots clés :prix, Omar Sy

description :Quels prix ont été décernés à Omar Sy?

Q9

mots clés : acteurs, joué avec, Omar Sy

description : Quels acteurs ont joué avec Omar Sy?

Annexe 3 : Exemple de fichier QREL

D1.html 0
D2.html 0
D3.html 1
D4.html 1
D5.html 0
D6.html 1
D7.html 0
D8.html 0
D9.html 0
D10.html 0
D11.html 0
D12.html 0
D13.html 0
D14.html 0
D15.html 0
D16.html 0
D17.html 0
D18.html 0
D19.html 0
D20.html 0
D21.html 0
D22.html 0
D23.html 0
D24.html 0
D25.html 0
D26.html 0
D27.html 0
...
D91.html 0
D92.html 0
D93.html 0
D94.html 0
D95.html 0
D96.html 0
D97.html 0
D98.html 0
D114.html 0
D115.html 0
D116.html 1
D117.html 0
D118.html 1
D119.html 1
D137.html 0
D138.html 0

La deuxième colonne indique si l'élément est pertinent (1), moyennement pertinent (0,5) ou non (0) pour la requête.