Examen du cours Recherche d'information / web sémantique

Les documents sont autorisés pour les 2 parties.

Partie « recherche d'information »

(durée indicative : 1h – barême 10 pts)

Exercice 1 (2)

Répondre aux questions suivantes :

a) Un document ne comportant aucun terme de la requête ne peut pas être pertinent pour cette requête ? vrai/faux Justifier
b) La représentation en sac de mots permet de capturer (représenter) le sens des mots. Vrai/Faux
c) La représentation en Trigrammes permet de capturer la syntaxe. Vrai/Faux
d) La racinisation permet-elle d'améliorer le rappel ou la précision ou les deux ? justifier ?
e) A votre avis, quel modèle de RI (parmi ceux étudiés en cours) serait le plus approprié pour rechercher des *Tweets* ? Justifiez ? On suppose qu'un Tweet est formé d'un message où les mots se répètent rarement.

Exercice 2 (4)

Considérons la collection suivante de trois documents:

D1 = {2 t1, 3 t2, 4 t5, 1 t7}
D2 = {1 t2, 2 t3, 2 t5, 2 t6, 2 t8}
D3 = { 1 t1, 1 t3, 5 t4, 2 t8}

Et la requête : Q = {2 t1, 1 t2, 1 t4}

Un document (resp. requête) est représenté par une liste de termes pondérés ayant la forme suivante : Dj{tf(t_i) t_i}, i=1..8. Où, tf(t_i) désigne la fréquence du terme t_i dans D_j.

Questions

1- Donner graphiquement le fichier inversé (dictionnaire + Posting) représentant cette collection de 3 documents
2- Donner pour chacun des modèles ci-dessous le score de pertinence de la requête avec le document D1 (RSV(q, D1)) en réponse à la requête :

a) Le modèle Booléen simple en considérant une requête disjonctive
b) Le modèle vectoriel utilisant la pondération de type *qqq.ddd=nnn.ltn*
c) Le modèle probabiliste (*Probabilistic Ranking Principle*) (BIR Model)

Exercice 3 (3 pts)

Soit q = q1, …,qm une requête, d un document et P(qi|d) la probabilité du mot qi dans le modèle de langue de d. On suppose que nous disposons d'une collection de documents comportant au total 8 mots w1, ..., w8.

La Table ci-dessous liste, pour chaque mot, sa probabilité dans le modèle de langue de référence, Pml(w|REF), estimé sur la collection (colonne 2), la fréquence du terme c(w, d) dans un document (colonne3). Les colonnes 4 et 5 représentent les probabilités du terme dans le modèle langue du document d, estimé respectivement selon le maximum de vraisemblance et Dirichlet avec le paramètre μ.

Mots	Pml(w|REF)	c(w, d)	Pml(w|d)	Pdir(w| d)
w1	0.3	2		
w2	0.15	1		
w3	0.1	2		0.125
w4	0.1	4		
w5	0.05	1		
w6	0.1	0		
w7	0.1	0		
w8	0.1	0		

1. Compléter la colonne 4, (Pml(w|d)), le modèle de langue du document.
2. La colonne 5 représente la probabilité du terme calculée après un lissage de Dirichlet effectuée sur la collection. Seule la probabilité de w3 est donnée dans le tableau, déduire la valeur de m ? (posez l'équation puis déduire cette valeur)
3. Sans calculer les probabilités de la colonnes 5, indiquer pour chacun des mots de cette colonne si sa probabilité lissée (Pdirm(w|d)) est {>;=;<} à celle non lissée, calculée selon Pml(w|d), c'est-à-dire celle de la colonne 4.
4. Quelle condition doit satisfaire c(w, d) pour que la probabilité lissée du mot w soit toujours la même que la probabilité non lissée quel que soit le paramètre μ ? Ecrire c(w,d) en fonction de la probabilité non lissée).

Partie « Web sémantique »

Documents autorisés - durée indicative : 50 mn – **barême** 10 pts

1. Questions de cours (3 points). Réponses COURTES attendues

a. Quel est l'objectif général du web sémantique ? donner trois objectifs plus spécifiques

Enrichir le web de données et métadonnées permettant à des programmes d'accéder aux contenus du web (des pages mais aussi des BD et.) de la même façon que des humains. Ou encore : rendre explicites et manipulables par des programmes la « signification » des contenus des pages web et les données du web. Sous-objectifs :

- Ajouter des métadonnées standards et identifiables sur le web
- Pouvoir raisonner sur ces métadonnées
- Favoriser l'Interopérabilité des programmes et des données

b. En quoi le web des données est-il une évolution du web sémantique ?

Le web des données est une version simplifiée du web sémantique qui se focalise sur les données présentes dans les BD accessibles depuis le web. Le but est de donner une sémantique « universelle » à ces données en leur associant un type représenté selon un format standard et permettant de produire des inférences. Le web sémantique visait initialement plutôt le texte des pages html du web.

c. Quelle est la différence entre une ontologie et une base de connaissance ?

1. Une ontologie est une représentation des catégories d'un domaine de connaissances, de manière structurée, de façon à respecter des principes de distinction et de bonne définition des concepts de ce domaine et des relations qu'ils entretiennent. L'ontologie décrit des classes ou concepts en les situant dans une hiérarchie correspondant à une relation d'inclusion entre ensembles, et en associant à ces classes des relations qui en définissent les propriétés et qui indiquent les contraintes respectées par les entités de ces ensembles.
2. Une base de connaissances (BC) décrit non seulement les classes mais aussi des individus d'un domaine qui appartiennent à ces classes. Elle peut aussi comporter des règles et contraintes portant sur ces individus. On y représente aussi les relations entre individus. La base de connaissances peut former un graphe.

2. Exercices

Exercice 2.1 (2,5 points)

Figure 1

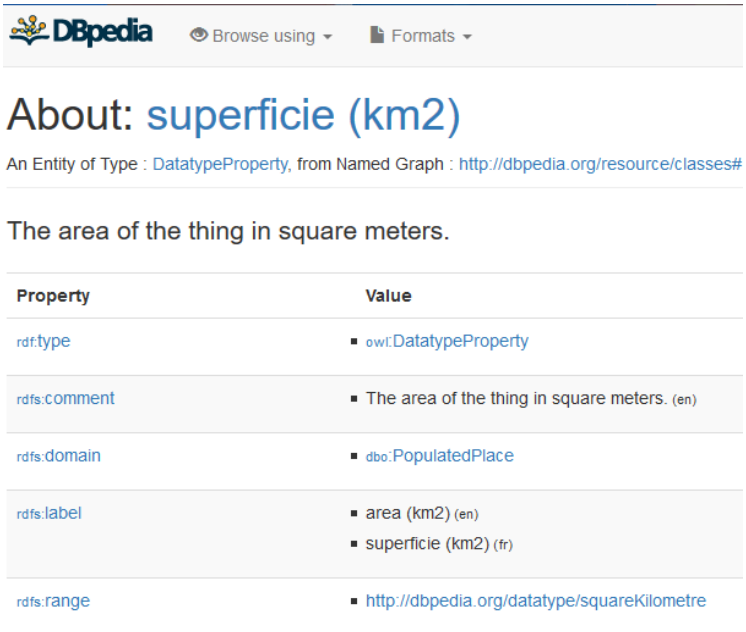

Figure 2

Les deux copies d'écran ci-dessus présentent les entités DBpedia suivantes
Figure 1 : http://dbpedia.org/page/Toulouse
Figure 2 : http://dbpedia.org/ontology/PopulatedPlace/area

a. Quels sont les espaces de noms utilisés pour décrire les propriétés visibles dans chacune des figures ?
   dbo http://dbpedia.org/ontology/
   rdf
   rdfs owl
   http://dbpedia.org/resources/classes http://dbpedia.org/resources/classdatatype

b. Quel est le Type principal de chacune de ces entités ? Ecrire en Turtle cette information.
   area a DataTypeProperty
   Toulouse a Lieu

c. A partir de la figure 2, quelle information est affichée après About: sur ces figures ?
   Le label en français

d. Etant données les propriétés sur la figure 2, et des informations de la figure 1, si on lance un raisonneur, quel type serait associé à http://dbpedia.org/page/Toulouse ? et que représente « 118,3 » ? expliquer la sémantique du vocabulaire rdfs que vous utilisez pour affirmer cela.
   dbo:PopulatedPlace car c'est le co-domaine (ou range) de la DataTypeProperty area et que Toulouse area 118 (km²) d'après la ligne « area »

e. Que diriez-vous du rdfs:comment de la figure 2 ?
   Il donne une définition qui n'est pas cohérente avec celle donnée dans le Label : le label indique que la mesure est en km² alors que le comment dit que la mesure sera

en m². Pour Toulouse, le chiffre est en km² et se conforme donc à la définition du label et non du champ « comment ».

Exercice 2.2 (2,5 points)

a. Corriger les 3 erreurs de syntaxe (vérifier la « ponctuation », les préfixes et les propriétés ou relations utilisés) et ajouter les informations manquantes (signalées par **????**) dans la définition de propriété suivante (pour répondre, indiquer le numéro de la ligne et pour chacune, ce qui doit remplacer **????)** :

```
1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
  @prefix ????        <http://dbpedia.org/ontology/> .
  @prefix ????        <http://www.w3.org/2002/07/owl#> .
  @prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
5 @prefix wikidata:   <http://www.wikidata.org/entity/> .
  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

   dbo:inseeCode     rdf:type          rdf:Property .
   dbo:inseeCode     ????              owl:DataProperty .
10 dbo:inseeCode     rdf:subPropertyOf     dbo:codeSettlement .
   dbo:inseeCode     owl:equivalentProperty      wikidata:P374 ;
       rdfs:label    "INSEE code"@en ,
                        "INSEE-code"@nl ????
       ????:domain   dbo:Settlement ;
15     rdfs:range    xsd:string ;
       rdfs:comment "numerical indexing code used by the French National Institute for
Statistics and Economic Studies (INSEE) to identify various entities"@en .
```

```
1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
  @prefix dbo:        <http://dbpedia.org/ontology/> .
  @prefix owl:        <http://www.w3.org/2002/07/owl#> .
  @prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
5 @prefix wikidata:   <http://www.wikidata.org/entity/> .
  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

   dbo:inseeCode     rdf:type          rdf:Property .
   dbo:inseeCode     rdf:type                owl:DataProperty .
10 dbo:inseeCode     owl:subPropertyOf     dbo:codeSettlement . erreur 1
   dbo:inseeCode     owl:equivalentProperty      wikidata:P374 ;
              rdfs:label    "INSEE code"@en ,
                               "INSEE-code"@nl ;
              rdfs:domain   dbo:Settlement ;
15            rdfs:range    xsd:String ; erreur 2
              rdfs:comment "numerical indexing code used by the French National Institute for
Statistics and Economic Studies (INSEE) to identify various entities"@en .
```

b. En utilisant la même syntaxe,
   - Définir que le range de cette propriété dbo:inseeCode est une chaîne de caractères

   dbo:inseeCode rdfs:range xsd:String

- définir que cette propriété est de type owl:FunctionalProperty.
  dbo:inseeCode rdf:type owl:FunctionalProperty .

c. Que signifie que cette propriété soit « functional » ? après avoir lancé un raisonneur, quel nouveau type sera associé à « Toulouse » ?

Si la propriété est « functional » alors elle doit être obligatoirement définie pour toute entité de type ville. Le raisonneur va conclure que Toulouse a dbo:area

Exercice 2.3 (2 points)

Soit un extrait de la représentation de Toulouse dont le début est présenté figure 1.

```
@prefix dbr:   <http://dbpedia.org/resource/> .
dbr:Toulouse  rdf:type        dbo:Location , dbo:Place , yago:YagoPermanentlyLocatedEntity .
dbr:Toulouse  dbp:population  461190 ;
              dbo:mayor       dbr:Jean-Luc_Moudenc ;
              dbp:website     <http://www.toulouse.fr/> ;
              dbo:country     dbr:France .
```

a. Ecrire une requête SPARQL qui retourne toute entité située en France et dont la population est supérieure à 10 000 (habitants). (Utiliser les propriétés présentes dans la définition de Toulouse). Cette requête retourne-t-elle dbr:Toulouse ? pourquoi ?

```
SELECT ?city ?pop
WHERE {   ?city dbp:population ?pop .
          ?city dbo:country dbr:France .
}
FILTER (?pop > 10000)
```

Oui car Toulouse dbp :population 461190 > 10000

b. Ecrire une requête qui cherche toutes les lieux d'habitation (dbo:Settlement) vérifiant ces mêmes critères. Cette requête permet-elle de retourner dbr:Toulouse ? Pourquoi ? Si vous répondez non, pourquoi et que faut-il faire pour que la réponse soit Oui ?

```
SELECT ?city ?pop
WHERE {       ?city dbp:population ?pop .
        ?city dbo:country dbr:France .
        ?city a dbo:Settlement .
}
FILTER (?pop > 10000)
```

Non car Toulouse n'est pas de type dbo:Settlement

Il faut lancer le raisonneur. A cause de la relation dbo:population, Toulouse va devenir de type dbo:Settlement et donc la requête retournera aussi Toulouse.