Projet TP DS51

Web sémantique, ontologies et extraction des connaissances

1. Informations générales

L'objectif de ce projet est de réaliser une petite application (le livrable sera le programme et un mini rapport de moins de 10 pages);

- Les projets sont réalisés par groupe. Une présentation de 10 minutes max aura lieu durant la dernière séance du TP.
- Votre livrable sera une archive (.zip) composée :
  - d'un répertoire "présentation" contenant la source et le PDF de votre présentation
  - d'un répertoire "rapport" contenant la source et le PDF de votre rapport
  - d'un répertoire "code" contenant ce que vous avez développé.
- Les archives (.zip) seront à déposer sur Moodle. Le dernier délai pour rendre le projet est le lundi 16 juin 2023 à 23h59.
- Le nom de l'archive et du répertoire obtenu après décompression seront la concaténation des **noms** des étudiants du groupe (en ordre alphabétique, sans espace), **séparés par des tirets** (-). Par exemple le livrable de Julie Le Cain et Jean Dupont se nommerait « **Dupont-Lecain.zip** ».

2. Contraintes

- 3 à 4 étudiants max par groupe.
- Le langage utilisé sera **Python/Java** en respectant les bonnes pratiques du langage; Introduction

3. Contexte du projet

L'apprentissage supervisé utilisé en Machine Learning permet aux ordinateurs entre autres d'apprendre à reconnaître, de manière automatique, des objets à partir de données labellisées. Les résultats des algorithmes les plus performants (tels que les réseaux de neurones profonds) sont impressionnants. Cependant, ils nécessitent de nombreux exemples pour effectuer la phase d'apprentissage. De plus, lorsqu'on les utilise pour labelliser une nouvelle donnée, ils ne sont pas capables de justifier leur décision. Ils peuvent toutefois remonter un score qui représente une probabilité que la décision soit une bonne décision ou montrer la partie de la donnée qui a été discriminante.

Pour ce projet, nous faisons l'hypothèse que les ontologies peuvent en partie lever ces deux verrous scientifiques.

4. Principe

Le système que nous voulons mettre en place est constitué d'une ontologie et d'algorithmes de classification supervisée.

a. L'ontologie

L'ontologie a pour objectif :

1. de décrire les objets du domaine d'un point de vue classification (Par exemple dans le domaine du vivant, la classification scientifique des espèces permettait d'identifier les classes (sous-espèce, espèce, genre, etc.) et les individus (Canis lupus familiaris, Canis lupus, Canis, etc.)) et d'un point de morphologique ;

2. de caractériser les images labellisées du corpus d'apprentissage. Les propriétés d'une image préciseront :
    a. où elle se trouve (URL) ;
    b. son format (taille, couleur, etc.) ;
    c. ce qu'elle contient.

L'utilisation d'un raisonneur permettra d'étendre, par inférence, ce que contient une image. Par exemple le fait qu'une image contienne un chien (Canis lupus familiaris) permet d'inférer que cette image

    a. contient aussi un Canis lupus, un Canis, un Canidae, etc.
    b. contient un museau, des pattes, etc.
    c. ne contient pas d'aile, de nageoire, etc.

b. La phase d'apprentissage

Le but de la phase d'apprentissage du système est :

1. d'associer à chaque Classe de l'ontologie (sous classe de Classification et Caractéristique Morphologique), un algorithme de classification supervisée binaire ;
2. d'activer la phase d'apprentissage de ces algorithmes à l'aide du corpus d'images, en précisant pour chaque image, en sortie de l'algorithme, si l'image contient, ou pas, un individu de la classe en question.

c. Utilisation d'une ontologie à l'issue de la phase d'exploitation

L'objectif de la phase d'exploitation est d'analyser une nouvelle image, d'y détecter, ou pas, des objets et de vérifier que ces détections sont cohérentes.

Pour cela, on commencera par créer un individu image *i* dans l'ontologie. L'image sera ensuite utilisée en entrée de chaque algorithme de classification. Les résultats des phases de décisions créeront des individus *a* et des relations entre *i* et *a* et entre les *a*.

Le raisonneur de l'ontologie devrait alors nous dire que ce qui a été reconnu dans l'image et si ces reconnaissances sont cohérentes.

5. Travail à réaliser

On se propose dans le cadre de ce projet de tester cette hypothèse de recherche sur la reconnaissance d'animaux dans des images. Nous utiliserons pour cela une partie de la base de données ImageNet[1] proposée par l'un des concours Kaggle[2]. ImageNet est un ensemble d'images annotées à l'aide du thésaurus WordNet[3].

Concrètement vous disposerez entre autres :

- d'un fichier texte associant un identifiant à des mots décrivant ce que contiennent les images associées à cet identifiant. Par exemple : n01614925 bald eagle, American eagle, Haliaeetus leucocephalus
- des répertoires (un par identifiant), contenant environ un millier de fichiers images au format JPEG. Par exemple :
  $ ls n01614925

[1] https://www.image-net.org/
[2] https://www.kaggle.com/c/imagenet-object-localization-challenge
[3] https://wordnet.princeton.edu/

n01614925_10000.JPEG	n01614925_2263.JPEG	n01614925_46966.JPEG
n01614925_1001.JPEG	n01614925_226.JPEG	n01614925_469.JPEG
n01614925_10038.JPEG	n01614925_2278.JPEG	n01614925_47025.JPEG
...		

La chaine de traitements de notre système est la suivante :

a. Phase d'apprentissage :

a. Identifier les identifiants ImageNet qui correspondent à des animaux. Vous utiliserez **Wikidata**[4]. Cela vous permettra de désigner dans le Web des données (URI Wikidata) ces animaux.
b. À partir d'un ensemble d'URI Wikidata d'animaux, construire automatiquement la partie Classification de l'ontologie, sous classes de la classe racine **Animalia**[5] avec optionnellement une **hauteur maximale**[6]. Vous n'oublierez pas de préciser les exclusions multiples (un chien n'est pas un poisson !) ;
c. Compléter « à la main » l'ontologie avec :
    - un arbre d'héritage de classes décrivant des caractéristiques morphologiques d'animaux (museau, patte, aile, bec, queue, poil, plume, etc.). Sa racine sera la classe Caractéristique Morphologique ;
    - une propriété possède entre Animalia et Caractéristique Morphologique ;
    - des contraintes sur la propriété possède pour les sous classes de Animalia ;
d. Peupler automatiquement l'ontologie avec des individus de types Images, Animalia et Caractéristique Morphologique, à partir des 99% des images d'ImageNet présentant les animaux retenus à l'étape 2 (le 1% restant sera utilisé pour la validation de l'expérience). Ces individus seront reliés entre eux à l'aide des relations contient (entre Image et sous classes de Animalia et Caractéristique Morphologique) et possède ;
e. Après inférence, associer à chaque individu de l'ontologie relié à des individus Image, un algorithme de classification supervisée, pour ensuite lancer la phase d'apprentissage de ces algorithmes.

b. Phase d'exploitation :

f. Valider l'approche avec les 1% d'images restantes. Pour cela, vous ajouterez à l'ontologie précédente un individu image avec ses les relations « positives » ou « négatives » issues des algorithmes d'apprentissage, puis vous lancerez un raisonneur pour valider la cohérence ou l'incohérence de l'ensemble.

Les identifiants ImageNet à retenir pour cette expérimentation sont (à partir de l'étape 2) :

- n02114367 (loup)
- n01484850 (requin)
- n01614925 (aigle)
- n02133161 (ours)
- n01537544 (passerin indigo)
- n01443537 (poisson rouge

[4] https://www.wikidata.org
[5] URI : https://www.wikidata.org/entity/Q729
[6] Par exemple si la hauteur est de 2 on aura pour la classe Canis Lupus familiaris l'héritage suivant : Animalia ← Canis lupus ← Canis Lupus familiaris