VESCRAPPER

RAPPORT DE PROJET MICRO-LANGAGE

https://github.com/vecolo-project/vescrapper

Swann HERRERA | Noé LARRIEU-LACOSTE

Projet Annuel

27 juillet 2021

TABLE DES MATIERES

I. INTRODUCTION

1. RAPPEL DU SUJET

Pour notre projet annuel, il nous a été demandé de concevoir un Micro-Langage.

Ce Micro-Langage doit nous permettre d'extraire du contenu sur internet en rapport avec notre sujet.

2. APPLICATION CHOISI

Nous avons décidé de développer un micro-langage pour remplacer le moteur de recherche Google à la façon d'un langage SQL.

Celui-ci pourra nous aider à récupérer des modèles de vélos, des images et les sites où ils sont vendus un exploitant la recherche avancée de Google.

https://www.astuces-aide-informatique.info/9691/commandes-recherche-avancees-google

II. FOCUS SUR L'APPLICATION

1. MENU

Lorsque nous lançons l'application (en ligne de commande), nous arrivons sur un menu nous détaillant ce qu'il est possible de faire sur l'application :

```
Welcome to Vescrapper! Enter the search you want
-----------------------------------------------
Enter exit(); to leave
Enter debugOn(); to show treeGraph (must have graphviz installed)
Enter debugOff(); to disable treeGraph showing (default)
-----------------------------------------------
Command you can use :
GET [text|image] OF <term> FROM <source> WITH [<conditional terms>]? LIMIT <number>?;

Examples:
- GET text,image OF rockrider FROM decathlon WITH ("ST 100" OR "ST 530 S") AND ("NOIR" OR "ROUGE") LIMIT 10;
- GET text OF rockrider FROM decathlon WITH "ST 100" LIMIT 10;
- GET image OF rockrider FROM go-sport LIMIT 5;

cmd >
```

Il suffit ensuite de taper la commande spécial, ou la requête désiré pour lancer une action.

2. MODE DEBOGAGE

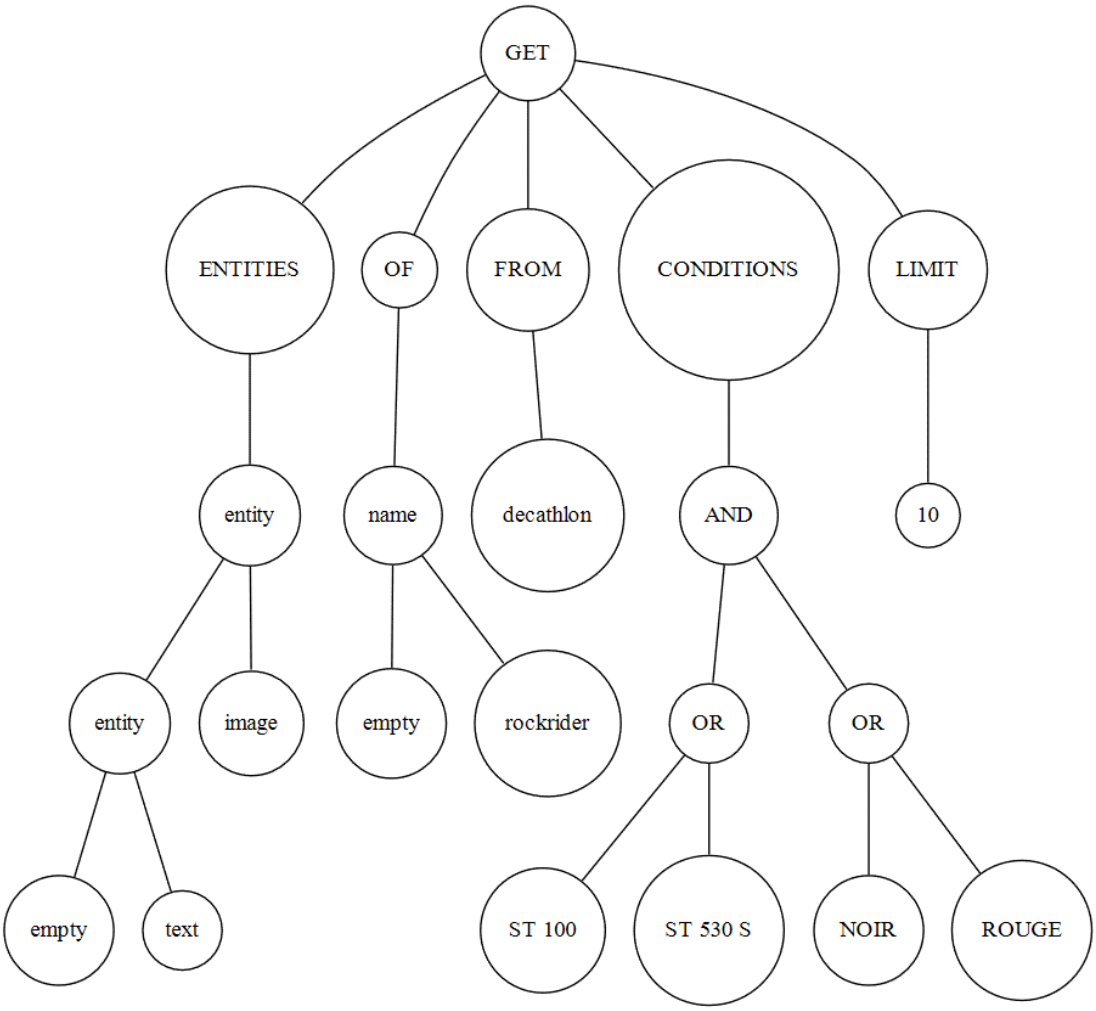
Le mode débogage permet de visualiser l'arbre de décision de notre requête grâce à la génération d'un fichier PDF.

Cet arbre (sous la forme de tuples) sera également affiché dans la console.

Pour utiliser le mode débogage, il faut obligatoirement avoir installé graphviz sur sa machine.

```
cmd > GET text,image OF rockrider FROM decathlon WITH ("ST 100" OR "ST 530 S") AND ("NOIR" OR "ROUGE") LIMIT 10;
```

```
('GET', ('ENTITIES', ('entity', ('entity', 'empty', 'text'), 'image')), ('OF', ('name', 'empty', 'rockrider')), ('FROM', 'decathlon'), ('CONDITIONS', ('AND', ('OR', 'ST 100', 'ST 530 S'), ('OR', 'NOIR', 'ROUGE'))), ('LIMIT', 10))
```



3. EXEMPLES DE REQUETES

Il est possible de taper différents types de requêtes afin de trouver des liens ayant du contenu, des images, ou les 2.

On doit spécifier quel est le produit que l'on recherche et sur quel site on veut récupérer les résultats.

Il est possible également d'ajouter des mots-clés avec des opérateurs booléen.

Enfin, on peut spécifier la limite du nombre de résultats qui par défaut est à 10.

```
cmd > GET text,image OF rockrider FROM decathlon WITH ("ST 100" OR "ST 530 S") AND ("NOIR" OR "ROUGE") LIMIT 10;
Generated google search is  ("jpeg"|"jpg"|"png"|"gif"|"")&("("ST 100" OR "ST 530 S")" AND "('OR', 'NOIR', 'ROUGE')") site:decathlon.* intitle:rockrider
https://www.decathlon.fr/p/velo-vtt-st-530-s-noir-rouge-27-5/_/R-p-311716
https://www.decathlon.fr/p/velo-vtt-st-540-27-5/_/R-p-301097
https://www.decathlon.fr/p/velo-vtt-st-530-27-5/_/R-p-311274
https://www.decathlon.fr/browse/c0-tous-les-sports/c1-velo-tout-terrain-vtt/c3-velos-vtt-homme/_/N-v3jc22
https://support.decathlon.fr/vtt-rockrider-st-530-s-noir-rouge
https://www.decathlon.fr/browse/c0-tous-les-sports/c1-velo-tout-terrain-vtt/c2-velos-vtt/_/N-pjowga
https://www.decathlon.fr/p/velo-vtt-st-100-27-5/_/R-p-192872
https://www.decathlon.fr/p/velo-vtt-st-120-27-5/_/R-p-305496
```

```
cmd > GET text OF rockrider FROM decathlon WITH "ST 100" LIMIT 10;
Generated google search is  &"ST 100" site:decathlon.* intitle:rockrider
https://www.decathlon.fr/p/velo-vtt-st-100-27-5/_/R-p-192872
https://support.decathlon.fr/vtt-rockrider-st-100-jaune
https://www.decathlon.fr/p/vtt-enfant-rockrider-st-100-24-pouces-9-12-ans-blanc/_/R-p-187631
https://www.decathlon.fr/p/vtt-enfant-rockrider-st-100-20-pouces-6-9-ans-blanc/_/R-p-300759
https://www.decathlon.fr/p/velo-vtt-st-100-femme-blanc-rose-27-5/_/R-p-300809
https://support.decathlon.fr/vtt-rockrider-st-100-noir
https://support.decathlon.fr/vtt-rockrider-st-100-gris
https://www.decathlon.fr/p/casque-velo-vtt-st-100/_/R-p-323788
https://www.decathlon.fr/p/velo-vtt-electrique-e-st-100-bleu-27-5/_/R-p-309736
https://support.decathlon.fr/vtt-rockrider-st-100-rouge
```

```
cmd > GET image OF VTT FROM go-sport LIMIT 5;
Generated google search is  ("jpeg"|"jpg"|"png"|"gif") site:go-sport.* intitle:VTT
https://www.go-sport.com/nos-selections/selection-scrapper/vtt-scrapper/
https://www.go-sport.com/sports/cycle/velos/vtt/vtt-regulier-homme/
https://www.go-sport.com/nos-marques/edito/scrapper/
https://www.go-sport.com/sports/cycle/velos/vtt
https://www.go-sport.com/sports/cycle/composants-velo/chambre-air/chambre-a-air-vtt/caa-vtt-26p-std-43600.html
```

III. CHOIX D'IMPLEMENTATIONS

1. LANGAGE CHOISI

Pour ce projet, nous avons choisi d'utiliser le langage python car c'est selon nous celui qui répondait le mieux à notre besoin. Il possède beaucoup de librairies permettant de parser et traiter des arbres de décision ce qui nous a permis d'aller relativement vite.

2. TECHNOLOGIES ET LIBRAIRIES PRINCIPALES UTILISEES

Afin de passer correctement nos requêtes pour le langage, il a fallu utiliser différentes librairies dont 2 principales.

Ply

Pli est un outil d'analyse écrit uniquement en python il s'agit d'une ré implémentation de Lex Yacc à l'origine en langage c.

Lex : Générateur d'analyseur lexical.

- Prends en entrée la définition des unités lexicales
- Produit un automate fini minimal permettant de reconnaître les unités lexicales

Yacc : Générateur d'analyseur syntaxique.

- Prends en entrée la définition d'un schéma de traduction (produit par Lex)
- Produit un analyseur syntaxique pour le schéma de traduction.

Graphviz

Graphviz est un logiciel de visualisation graphique open source. La visualisation de graphes est un moyen de représenter des informations structurelles sous forme de diagrammes de graphes abstraits et de réseaux.

Il a des applications importantes dans les réseaux, la bio-informatique, le génie logiciel, la conception de bases de données et de sites Web, l'apprentissage automatique et les interfaces visuelles pour d'autres domaines techniques.

Dans notre cas, il est utilisé dans le mode débogage pour visualiser notre arbre de décision générée par Ply.

IV. BILAN DU PROJET

1. PROBLEMES RENCONTRES

Construction de la requête Google

N'ayant pas trouvé de documentation officielle de Google sur toutes les requêtes avancées possibles, il a fallu chercher sur des forums et essayer nous-mêmes différentes combinaisons avant de trouver celle qui fonctionnait le mieux.

Il fallait donc adapter le langage aux combinaisons possibles pour que cela fonctionne.

Théorie d'un langage de programmation

Même si nous avons eu des cours cette année, la théorie des langages reste un domaine que nous ne maîtrisons difficilement dans le groupe et il a fallu redoubler d'efforts pour arriver à créer un nouveau langage.

2. CONCLUSION

Pour conclure, cette application a été difficile à réaliser de par la complexité de ce qu'est un langage de programmation que par le fait de trouver une idée concrète qui puisse venir enrichir le projet de base. L'application finale nous a tout de même été très utile au moment du remplissage de la base de données car elle nous a permis de trouver très rapidement des modèles de vélo électrique ainsi que des images associés