Techniques Web
Elvis MBONING

Manuel methodologique du projet Web

(Projet personel)

Siyu WANG
21800525

Master TAL M2
INALCO

Présentation

Ce manuel est une instruction sur notre application *Scrape&See* qui permet à nos chers clients de découvrir les possibilités de conquérir de nouveaux domaines technologiques où ils pourraient investir et diversifier ainsi leurs activités.

Concrètement, d'une part, l'application vous montrera la possibilité d'un système d'aide au développement logiciel dont l'objectif sera la détection automatique de solution de débogage qui pourraient aider les développeurs Python à optimiser leur rendement professionnel. Et d'autre part, l'application vous montrera aussi la possibilité de soutenir la communauté des bénévoles travaillant sur les langues peu-dotées, dans l'objectif de créer un marché de la donnée autour de ces langues.

Installation

Ces deux possiblités, hébergées sur une seule application *Scrape&See*, lancées via streamlit, permettent aux clients visualliser les données (récupérées et traitées) sur un site html.

Attention : une connexion internet est obligatoire pour faire fonctionner notre application !

Tout d'abord, il faut créer un environement virtuel.

Pour installer le module pipenv :

```
MBP-de-Siyu:projet siyuwang$ pip3 install pipenv
```

Pour créer et activer l'environement virtuel pipenv :

```
MBP-de-Siyu:projet siyuwang$ pipenv --python 3 && pipenv shell
```

Pour installer les outils requis : vous avez simplement à exécuter le script la ligne de commande :

```
(scrapy) bash-3.2$ pip3 install -r requirements.txt
```

Pour lancer l'application, il faut utiliser le script bash launcher.sh.

```
(scrapy) bash-3.2$ bash launcher.sh
```

Quand le script est exécuté, on vous installera tous les outils et modules requis pour faire fonctionner notre application, ensuite, l'application va démarrer automatiquement.

Puis, vous pouvez visualiser notre projet via l'adresse : .

* Si vous avez déjà installé les outils et que vous voulez démarrer directement l'application, il suffit de utiliser cette ligne de commande :

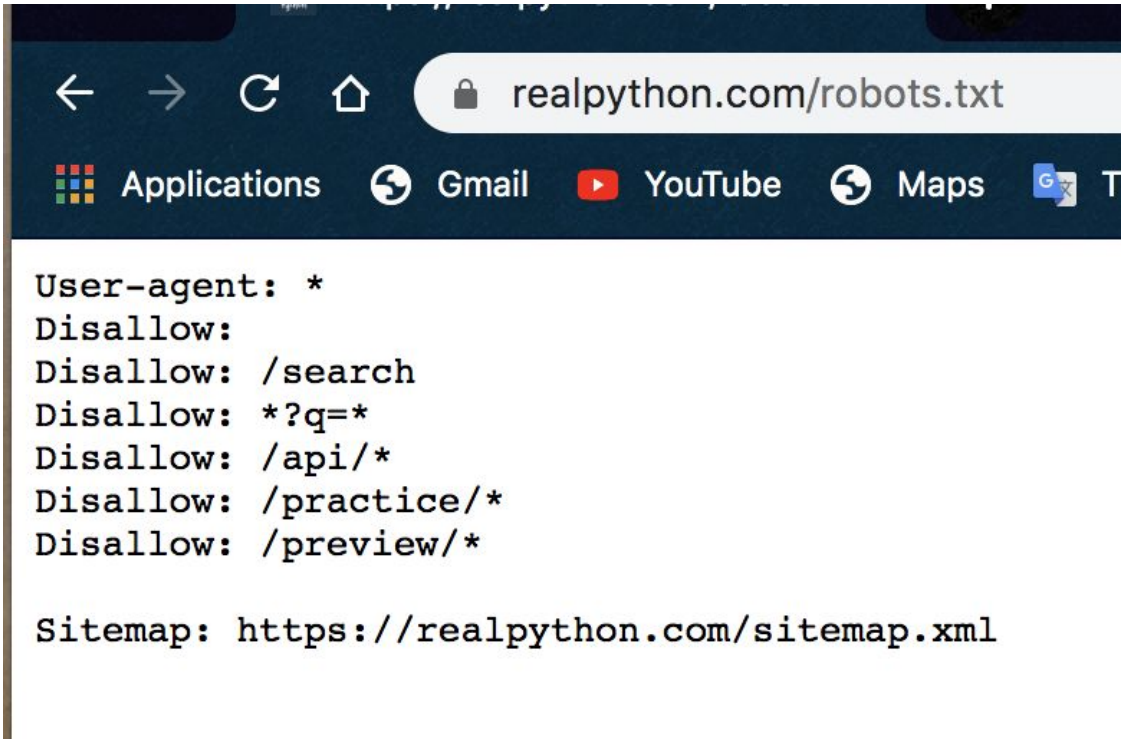
```
streamlit run projet_siyuwang.py
```

Méthodologies scrapy

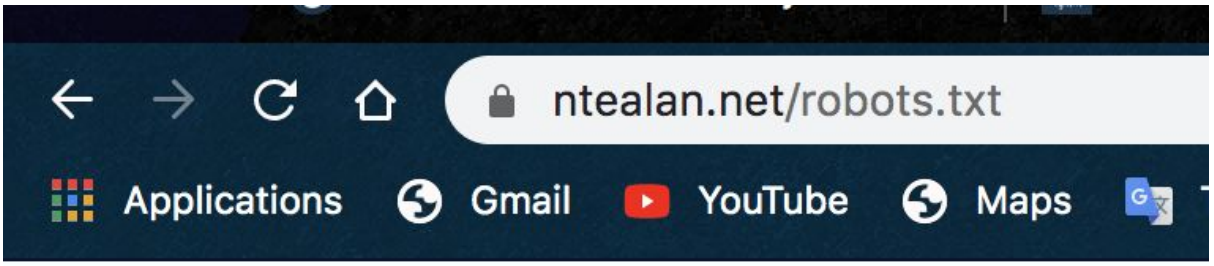
1. Accès aux données

Certains utilisateurs peuvent se voir l'accès refusé quand ils veulent extraire des informations d'un site, c'est pourquoi il faut savoir quels robots sont autorisés à se rendre sur un site. Et pour faire cela, il suffit de lire son fichier robots.txt placé à la racine de chaque serveur.

Ici, sur le site de realpython.com, le fichier est ceci :



Et puis, sur le site https://ntealan.net/, tous robots sont autorisés.



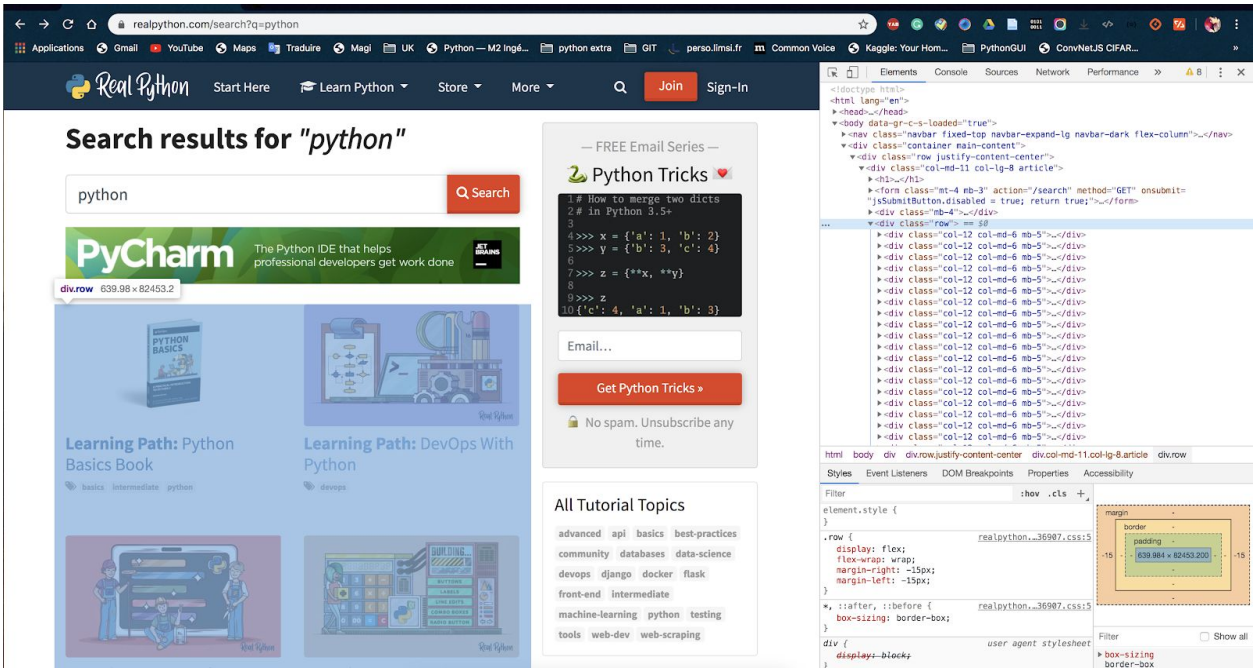

```
User-agent: CuteStat
Disallow: /
```

Donc nous pouvons utiliser le module scrapy de python sans problèmes.

2. Structure des pages HTML

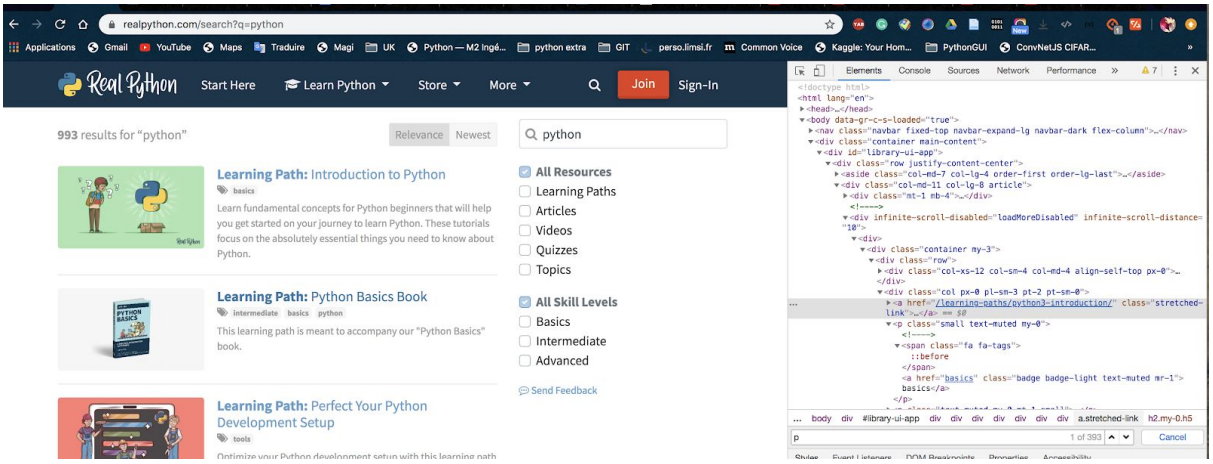
Il faut toujours bien connaitre la structure de la page HTML avant d'extraire les informations, cela facilitera la récupération des informations pertinentes.

Pour le site https://realpython.com, la structure des balises est la suivante (l'ancienne version du site):

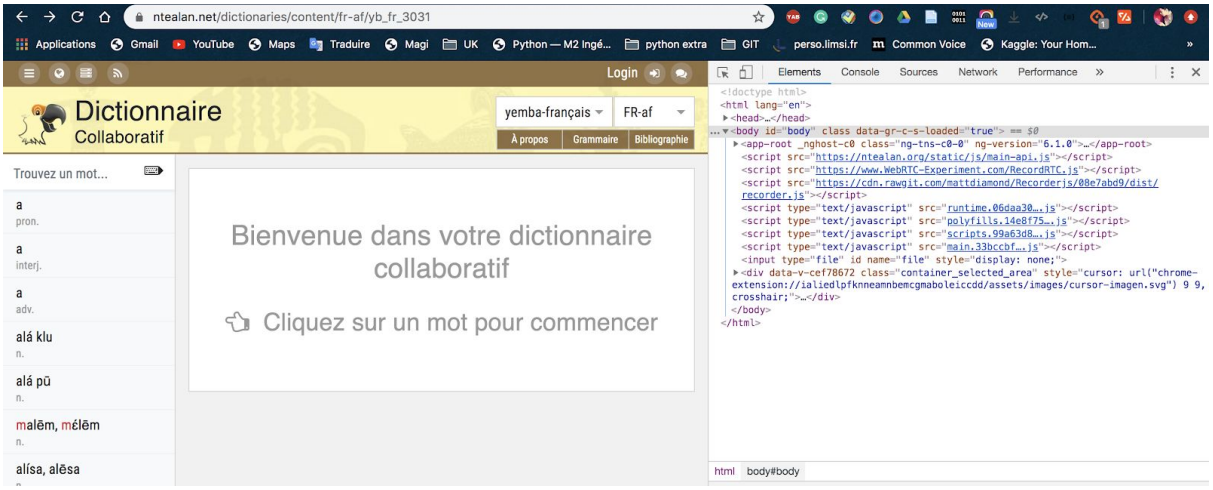


Nous avons besoin donc le div class="row" pour extraire les informations pertinentes, c'est à dire chaque block de div classe="col-12 col-md-6 mb-5".

Puis, la structure du site a été changé au milieu du projet, et la structure devienne celle ci :



Et puis, le site https://ntealan.net/dictionaries/content/fr-af/yb_fr_3031 a une autre structure puisqu'il s'agit un site SPA.



3. Sélection des données

- **Python**

Pour mieux présenter la possibilité de la détection automatique de solution de débogage qui pourraient aider les développeurs Python, j'ai décidé de récupérer les titres, les liens, les étiquettes de chaques block des résultats de recherche et les catégories sur le site https://realpython.com.

Puis, j'ai pensé au site https://stackoverflow.com/ qui est aussi une site très utile pour les développeurs. Ce sera très interessant que les utilisateurs puisse

rechercher directement les mots clés des message d'érreur sur le site de Stack Overflow dans notre application.

- **Ntalan**

Pour mieux soutenir la communauté des bénévoles travaillant sur les langues peu-dotées, l'idée d'une proposition de traduction des mots est très intéressante.

J'ai donc décidé d'ajouter la fonctionnalité de faire saisir un mot en français par l'utilisateur et puis proposer les traductions possibles en Yemba (des propositions de mots avec leurs partis de discours).

4. Extraction

- **Real Python**

Tout d'abord, j'ai récupéré le nombre total des articles trouvés.

Puis, j'ai récupéré tous les catégories à droite de la page et le nombre d'articles correspondant à chaque catégorie en simulant le mouvement de click avec le module selenium de python.

Ensuite, il y a des étiquêttes liés à chaque titre d'article, je les ai donc aussi récupérées et stockées dans un dictionnaire avec le titre comme clé et les étiquêttes correspondantes comme valeur (si il y en a). Un autre dictionnaire avec des étiquêttes comme clé et les articles correspondants à l'étiquêttes est aussi créé.

Enfin, j'ai récupéré le lien de chaque article trouvé pour que l'utilisateur puisse consulter les articles pertinantes. Un échantillon de 50 pages d'articles sont aussi téléchargés en locale lors de l'éxecution du programmes pour future analyse.

- **Stack Overflow**

Le site de Stack Overflow consiste principalement à debugger quand on a les message d'erreur.

J'ai donc d'abord récupéré le nombre total des résultats trouvées avec les mots clés saisis par l'utilisateur, puis récupéré les 10 questions les plus pertinentes dans un dictionnaire avec le titre comme clé et leur lien comme valeur.

Un échantillon de 15 pages de résultats sont aussi téléchargés en locale lors de l'éxecution du programmes pour future analyse.

- **Ntalan**

Pour faire la fonctionnalité de proposition de traduction, il faut faire un recherche de mot saisi par l'utilisateur sur le site de ntealan, puis récupérer les entrées de dictionnaires trouvées et leurs partie de discours.

J'ai donc utilisé le module Selenium de python pour controler la zone d'input sur la page pour entrer automatiquement le mot à rechercher, puis stocké les données dans un dictionnaire avec le mot en yemba comme clé et son partie de discours comme valeur.

Visualisations

La visualisation s'agit d'une application présentée avec Streamlit.

Veuillez faire tourner le programme et puis consulter l'url pour visualiser notre application : http://localhost:8501

Attention : l'application va ouvrir et faire tourner votre navigateur Chrome de façon automatique, veuillez attendre jusqu'à la fermeture automatique du Chrome pour chaque requête pour bien visualiser les résultats.

- Ntealan