Exercice : le moissonneur de l'espace

Alexandre Roulois (Université Paris Cité, LLF, CNRS)

L'objectif de ce court exercice est de mettre à profit vos premières compétences en *Web Scraping* pour analyser les titres d'une page Web.

Aurélien Barrau

3 langues

Article Discussion — Lire Modifier Modifier le code Voir l'historique

Pour les articles homonymes, voir Barrau.

Aurélien Barrau, né le 19 mai 1973 à Neuilly-sur-Seine, est un astrophysicien et philosophe français.

Spécialisé en relativité générale, physique des trous noirs et cosmologie, il est directeur du Centre de physique théorique Grenoble-Alpes et travaille au Laboratoire de physique subatomique et de cosmologie de Grenoble. Professeur à l'université Grenoble-Alpes, il travaille actuellement [Quand ?] sur la gravité quantique. Il est par ailleurs docteur en philosophie.

Militant écologiste, il est favorable à la décroissance. Il est lié aux thèses collapsologistes, bien qu'il récuse cette étiquette.

Biographie [modifier | modifier le code]

Études [modifier | modifier le code]

De 1990 à 1992, il étudie en classe préparatoire aux grandes écoles, Math-sup et Math-spé, au lycée Pasteur de Neuilly-sur-Seine. Il obtient son diplôme d'ingénieur à l'École nationale supérieure de physique de Grenoble (aujourd'hui fondue dans Grenoble INP-Phelma) en 1995, en étant major de promotion. Il obtient simultanément un diplôme d'études approfondies (DEA) en physique de la matière et du rayonnement (filière physique subatomique) de l'université Joseph-Fourier (Grenoble-I) en 1995 en étant à nouveau major de promotion. Il obtient son doctorat en astrophysique à l'université Joseph-Fourier[1] en 1998 avec la mention très honorable et les félicitations du jury, sur le sujet « Astrophysique gamma de très haute énergie, étude du noyau actif de galaxie mrk501 et implications cosmologiques », travail mené au LPNHE-Paris. Son habilitation à diriger des recherches (HDR) lui est délivrée en 2004 sur la thématique des trous noirs primordiaux.

Il obtient un autre doctorat, en philosophie, à l'université Paris-Sorbonne, soutenu en 2016 avec la mention très honorable et les félicitations du jury[2], portant sur « Anomies : une déconstruction de la dialectique de l'un et de l'ordre, entre Jacques Derrida et Nelson Goodman ». Le travail a été mené aux archives Husserl de l'École normale supérieure et dirigé par Marc Crépon.

Activité scientifique [modifier | modifier le code]

Aurélien Barrau

Aurélien Barrau en 2019.

Biographie

Naissance	19 mai 1973 (49 ans) Neuilly-sur-Seine
Nationalité	française
Formation	Lycée Pasteur de Neuilly-sur-Seine (classe préparatoire aux grandes écoles) (1990-1992) École nationale supérieure de physique de Grenoble (diplôme d'ingénieur) (1992-1995) Université Grenoble-I (diplôme d'études approfondies) (1992-1995) Université Grenoble-I (doctorat)

Figure 1: Page de Aurélien Barrau

Aperçu du travail

Rendez-vous sur la page *Wikipédia* de Aurélien Barrau. Sur cette page figurent plusieurs titres à des niveaux hiérarchiques différents :

- Niveau 1 : « Aurélien Barrau »
- Niveau 2 : « Biographie »
- Niveau 3 : « Études », « Activités scientifiques »
- …

Vous devrez récupérer tous ces titres et les lister.

Défi : essayez de parvenir au résultat sans consulter le reste du document.

Moissonner la page Web

Avant de commencer, préparez la structure minimale d'un script en Python dans un fichier nommé *sp_harvesting.py* que vous sauvegarderez sur votre machine.

Dans la partie *Modules*, importez *urllib.request*. Puis, dans la procédure principale, définissez une variable `url` à laquelle vous affecterez l'adresse de la page à moissonner.

Vous adapterez ensuite le code fourni dans le cours pour récupérer tout le contenu de la page HTML dans une variable `html` :

```
# additional headers
headers = { 'User-agent' : 'Titles extractor' }

# HTTP request
request = urllib.request.Request(url, headers=headers)

# load HTML document
with urllib.request.urlopen(request) as webpage:
    # get the html content
    html = webpage.read()
```

À fin de vérification, affichez le contenu de la variable `html` grâce à la fonction `print()` et lancez le script depuis un terminal. Pour mémoire, la syntaxe de la commande vaut :

```
python path/to/your_script.py
```

Récupérer le texte des titres

Pour parvenir au résultat, vous devrez recourir à la librairie *BeautifulSoup*. Importez-la puis créez une nouvelle instance grâce au constructeur `BeautifulSoup()` dans une variable `soup` :

```
# new instance of BeautifulSoup
soup = BeautifulSoup(html, 'html.parser')
```

Avec les outils de votre navigateur Web (souvent : clic droit puis *Inspecter*), repérez le codage HTML du titre de niveau 1 « Aurélien Barrau » :

```
<h1 id="firstHeading" class="firstHeading mw-first-heading">
    <span class="mw-page-title-main">Aurélien Barrau</span>
</h1>
```

Vous remarquez que le texte « Aurélien Barrau » se trouve dans une balise `<span>` elle-même imbriquée dans une balise `<h1>`. Comme le contenu de `<span>` est le seul de `<h1>`, il n'y a aucun risque à utiliser la méthode `.find_all()` avec l'argument `h1` :

```
# get textual content in h1 tags
h1 = soup.find_all('h1')
```

Répétez l'opération pour tous les autres titres et affichez-les à la fin avec la fonction `print()`. Que constatez-vous ?

Analyser le code HTML

Si vous avez exécuté le code ci-dessous pour les titres de niveau 2, vous avez remarquez qu'il a récupéré du bruit (les liens pour l'édition du contenu) :

```
h2 = soup.find_all('h2')
```

Comment ne conserver que le contenu qui nous intéresse ? Dans le code source de la page HTML, on observe que tous les titres du contenu éditorial sont dans une balise `<span>` à laquelle on a adjoint une classe CSS `mw-headline`.

Une meilleure façon de procéder serait alors de paramétrer un sélecteur CSS dans une méthode `.select()` :

```
# better solution
h2 = soup.select('h2 span.mw-headline')
```

Défi : pour finir, essayer de modifier votre procédure principale en définissant une fonction `main()`.