Traitement de l'information

Ophélie Fraisier

2018 – 2019

[Pour information] Installation des bibliothèques Python manquantes

Si vous rencontrez une erreur lors du lancement de vos scripts, de la forme `ImportError: no module named X`, cela signifie que la bibliothèque X n'est pas installée. Pour corriger cela vous pouvez suivre les étapes suivantes :

1. Installation de pip, l'utilitaire permettant d'installer, mettre à jour et supprimer des paquets Python
   - Entrez la commande `pip3 help` dans un terminal.
     Si l'aide de pip3 s'affiche, vous présentant les différentes commandes possibles, cela signifie que pip3 est déjà installé et que vous pouvez passez directement à l'étape 2.
     Sinon, continuez l'installation de pip3.
   - Téléchargez le script à l'adresse suivante : `https://bootstrap.pypa.io/get-pip.py`
   - Dans un terminal : `python3 get-pip.py --user`
     Cette commande installe pip dans ∼/.local/bin
   - Ajoutez ∼/.local/bin au path : `echo "PATH=$PATH:∼/.local/bin" >> ∼/.bashrc`
     Pour cette étape, **ne faites pas de copier-coller depuis le PDF**. Certains caractères invisibles pourraient faire échouer la commande.
2. Installation des bibliothèques manquantes
   - Ouvrez un nouveau terminal
   - `pip3 install --user nom_de_la_bibliothèque_manquante`

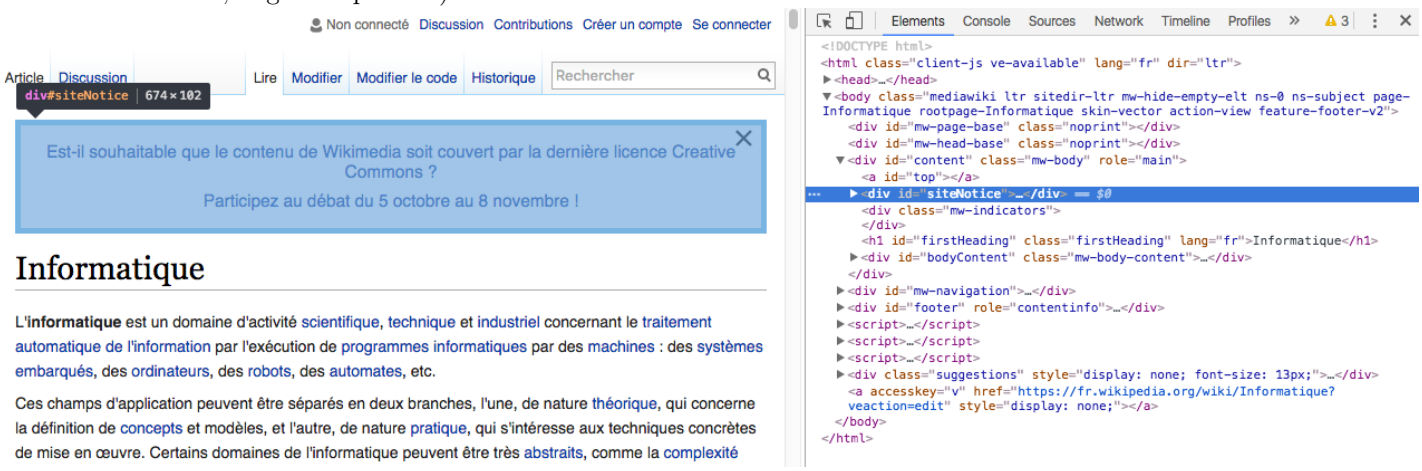
1 Scrapping

Allez sur la page Wikipedia consacrée à l'informatique : `https://fr.wikipedia.org/wiki/Informatique`.

1.1 Observez le code HTML et trouvez les balises permettant d'identifier les éléments suivants :

1. Les liens des images présentes dans l'article ;
2. Les citations placées au début de la section « Définitions ».

Vous pouvez vous aider des outils de développement de votre navigateur (Ctrl+Maj+I sous Chrome et Firefox, onglet Inspecteur).



1.2 Écrivez un script Python permettant d'extraire de la page HTML ces éléments grâce à BeautifulSoup.

1. Les liens des images présentes dans l'article ;
2. Les citations placées au début de la section « Définitions ».

Par sécurité, téléchargez la page HTML dans votre répertoire de travail (Ctrl+S sur la plupart des navigateurs) et cherchez les informations dans le fichier local plutôt que d'interroger directement la page Wikipedia en ligne.
La seule différence dans votre script sera la source du code HTML :

```
# version avec interrogation en ligne
r = requests.get("https://fr.wikipedia.org/wiki/Informatique")
if r.status_code == 200:
    text = r.text
# version avec fichier local
f1 = open(source_file, "r")
text = f1.read()
f1.close()
```

Les contenus seront récupérés grâce à des requêtes de la forme :

```
results = [e.get("title") for e in soup.select("h3.yt-lockup-title a")]
```

Si vous rencontrez l'erreur `bs4.FeatureNotFound: Couldn't find a tree builder with the features you requested: lxml.`, cela signifie que lxml n'est pas installé sur votre poste. Pour corriger cela :

```
pip3 install --user lxml
```

2 À l'aide de l'API Open Movie Database, trouvez la date de sortie de la série Friends.

Vous trouverez ci-dessous un extrait de la documentation traduite en français pour vous aider, la documentation complète étant disponible à l'adresse `http://www.omdbapi.com`.

Paramètre	Optionnel	Options valides	Valeur par défaut	Description
`t`	Non			Titre à rechercher.
`type`	Oui	*movie*, *series*, *episode*		Type d'élément.
`y`	Oui			Année de sortie.
`plot`	Oui	*short*, *full*	*short*	Scénario résumé ou complet.
`r`	Oui	*json*, *xml*	*json*	Format des données renvoyées.

3 API Twitter

3.1 Créez les tokens d'authentification nécessaires pour interroger l'API Twitter.

3.2 Écrivez un script permettant de récupérer les derniers tweets contenant le mot-clé `#Toulouse` (retweets compris).

3.3 Écrivez un script permettant de récupérer les derniers tweets du profil `IutToulouse3` (`https://twitter.com/IutToulouse3`) (retweets compris).

3.4 Exportez dans un fichier CSV le nombre de retweets de chaque tweet du profil `IutToulouse3`.

Pour cette question, considérez uniquement les tweets originaux de `@IutToulouse3` et non les retweets qu'il a réalisé. Plusieurs solutions pour cela :

- Vérifiez que le tweet ne débute pas par **"RT @"**
- Vérifiez la présence d'un objet `retweeted_status` dans l'objet JSON du tweet : s'il s'agit d'un tweet original, il n'existe pas, sinon il contient les informations du tweet original ayant été retweeté par le profil `@IutToulouse3`.

Pour chaque objet JSON représentant un tweet, le nombre de retweets de celui-ci est stocké dans le champ `retweet_count`.

Une solution simple lorsque vous souhaitez rapidement stocker des informations dans un fichier est de rediriger la sortie de votre script à l'aide de l'opérateur `>` :

```
python3 mon_script.py > my_file.txt
```

Cette commande indique au système de stocker la sortie du script `mon_script.py` dans le fichier `my_file.txt` au lieu de l'afficher dans le terminal. Attention, si le fichier `my_file.txt` existe déjà, cette commande écrasera son contenu.
Si vous voulez ajouter du contenu à la fin d'un fichier existant, utilisez plutôt l'opérateur `>>` :

```
python3 mon_script.py >> my_file.txt
```

4 Suppléments

4.1 Téléchargez le dossier WIKI à l'adresse `http://irit.fr/~Ophelie.Fraisier/data/WIKI.zip` et modifiez le script fait dans la question 1 pour extraire les informations de tous les fichiers HTML présents dans ce dossier.

4.2 Écrivez un script Python interrogeant l'API Omdb et présentant les titres et dates de sortie des films correspondants à un mot-clé fourni en paramètre.

4.3 Interrogez l'API Twitter pour répondre aux questions suivantes :

4.3.1 Affichez la liste des différents lieux nommés « Toulouse ».

4.3.2 Déterminez combien de tweets ont été postés chaque jour de la semaine (lundi, mardi, ...) sur les 500 derniers tweets parlant de `#Toulouse`.

4.3.3 Même question avec les heures de la journée.