Big Data & BI

Collecte de Données

1. **Installation Python v3.x:**

Visiter le site web et suivez les étapes d’installation pour votre système (en occurrence CentOS): <https://realpython.com/installing-python/>

1. **Installation de Scrapy:**

C:\Users\Adil\AppData\Local\Microsoft\Windows\INetCache\Content.Word\scrapylogo.pngAprès l’installation du python lancer la commande suivante pour installer Scrapy :

* pip install scrapy

Scrapy est un framework open-source permettant la création de robots d'indexation (web crawler ou collecteur web)

1. **Install Docker :**

Installer Docker : <https://docs.docker.com/install/>

Docker jouera le rôle de Conteneur pour les services du web crawler

1. **Install Splash for Scrapy :**

Pour collecter et pouvoir scraper les sites qui utilisent JavaScripts il faut installer la bibliothèque scrapy-splash

* pip install scrapy-splash

Assurer que le service est bien démarré et installé

* sudo systemctl start docker

Pour lancer le service Splash et pouvoir l’utiliser, lancer la commande suivante :

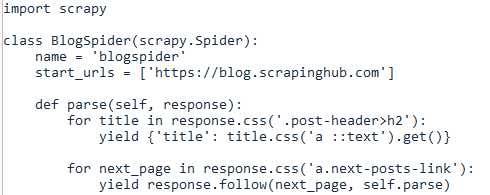
* docker run -p 8050:8050 scrapinghub/splash

1. **Scraper les données (Google Scholar, IEEE, USPTO, LinkedIn …):**

* **Example d’un blog HTML :**

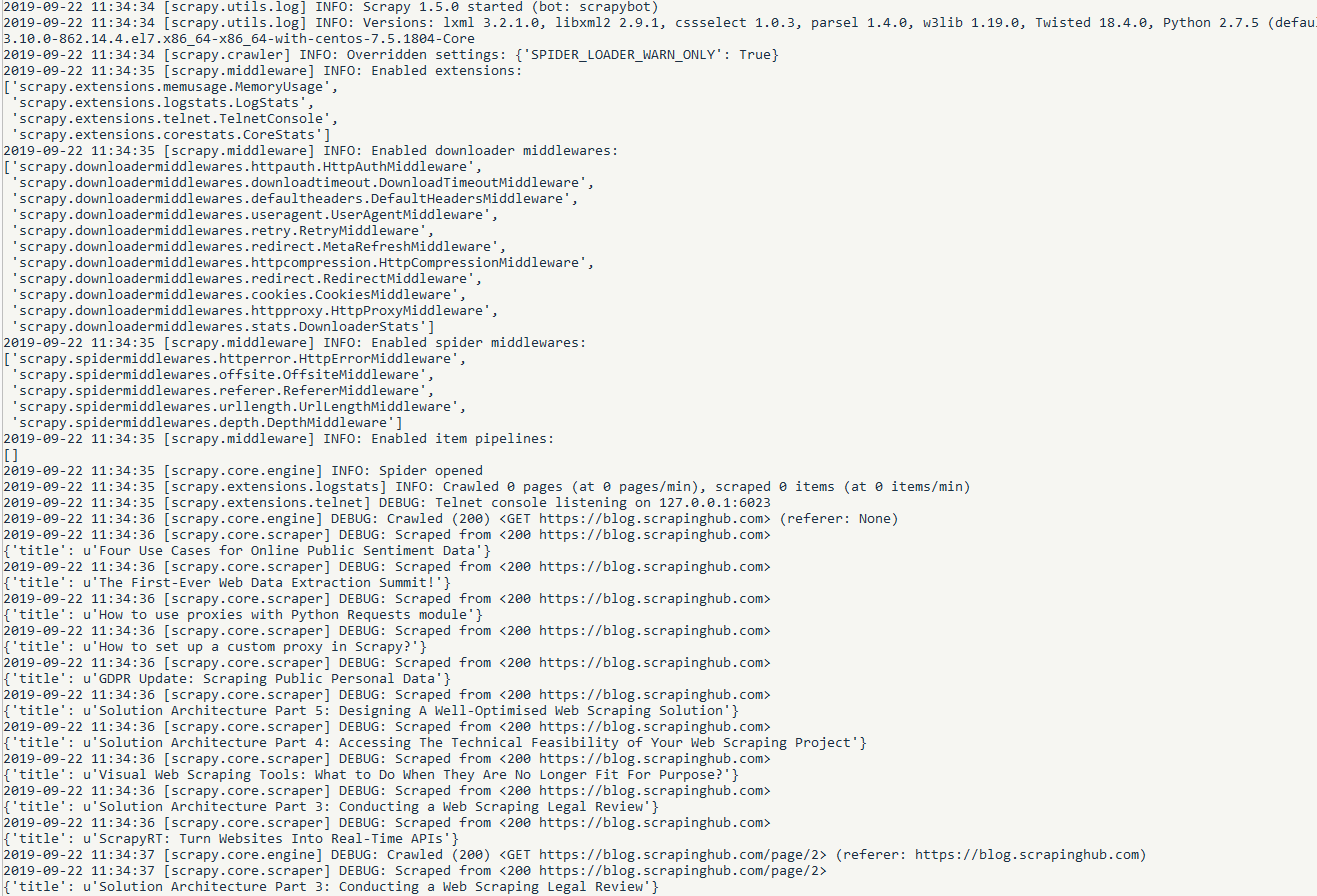
Créer un fichier et saisir le code suivant :

* touch myspider.py
* nano myspider.py



Pour lancer le script de Scraping, tapez la commande suivante :

* scrapy runspider myspider.py



Le résultat obtenu contient les titres de tous les postes publiés sur le blog, dans la suite on va scraper un site avancé qui utilise JavaScript et les appels AJAX qui demandent un traitement différent de ce qu’on a vu dans cet exemple.

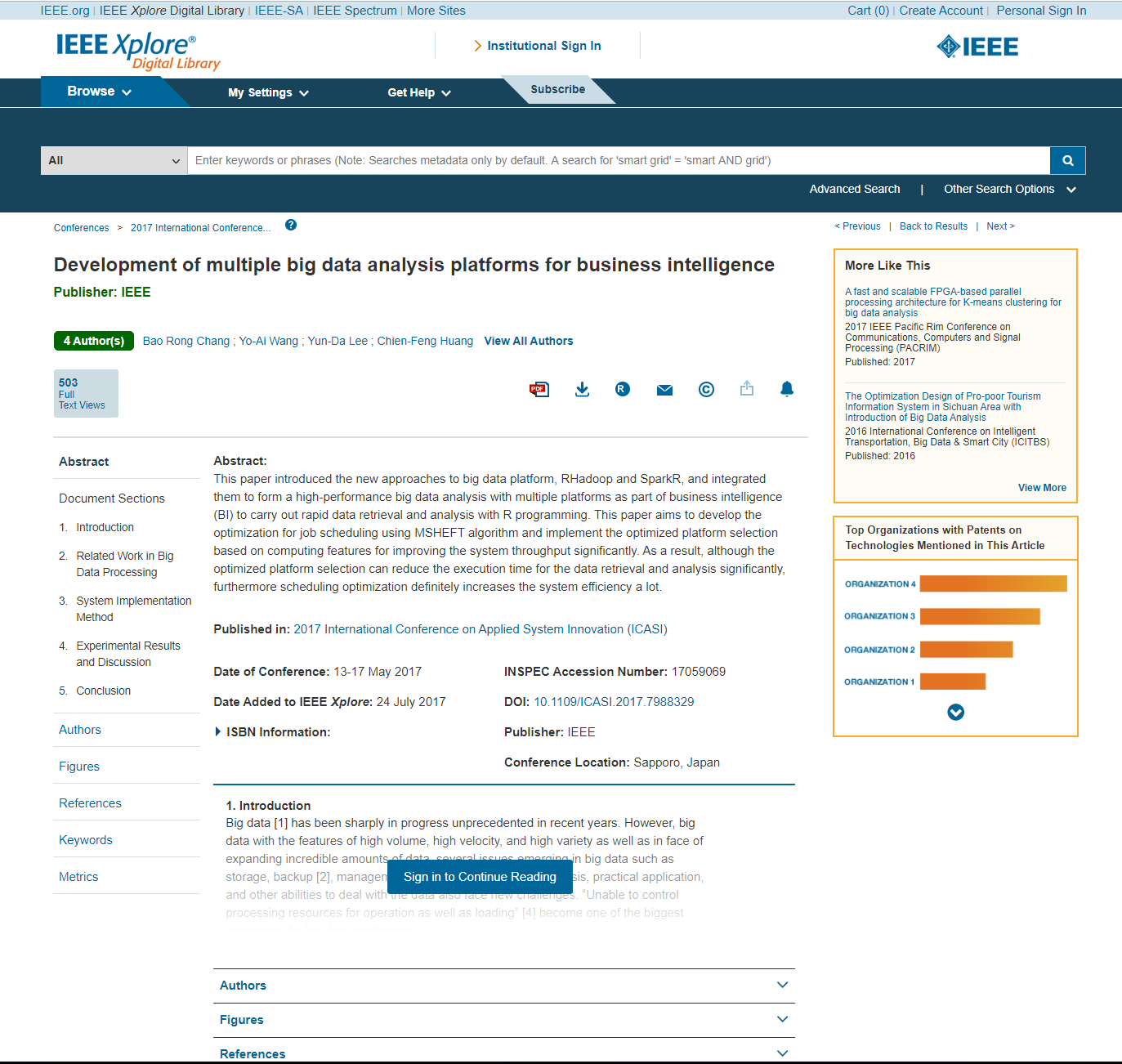
* scrapy runspider myspider.py -o blog.json



On peut enregistrer les données collectées dans un fichier JSON indiqué par l’option –o

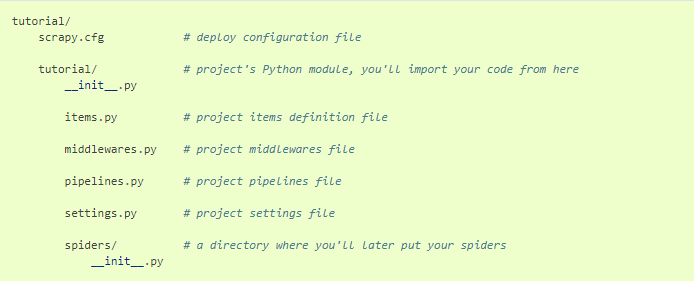
1. **Exemple Scraper IEEE :**

On collecte les articles scientifiques à partir de la bibliothèque IEEE Xplore en utilisant Scrapy et Splash, le script suivant montre un exemple d’un scraper qui a pour objectif l’extraction des champs suivants : Title, auhtors, date de publication…



Pour créer un projet Scrapy  visiter le lien suivant : <https://docs.scrapy.org/en/latest/intro/tutorial.html>

L’architecture d’un projet Scrapy contient les éléments suivants :



**Items.py :** présentation Objet de l’élément à scraper



**middlewares.py :** définie les modifications à effectuer sur l’objet avant ou après les transactions Scrapy

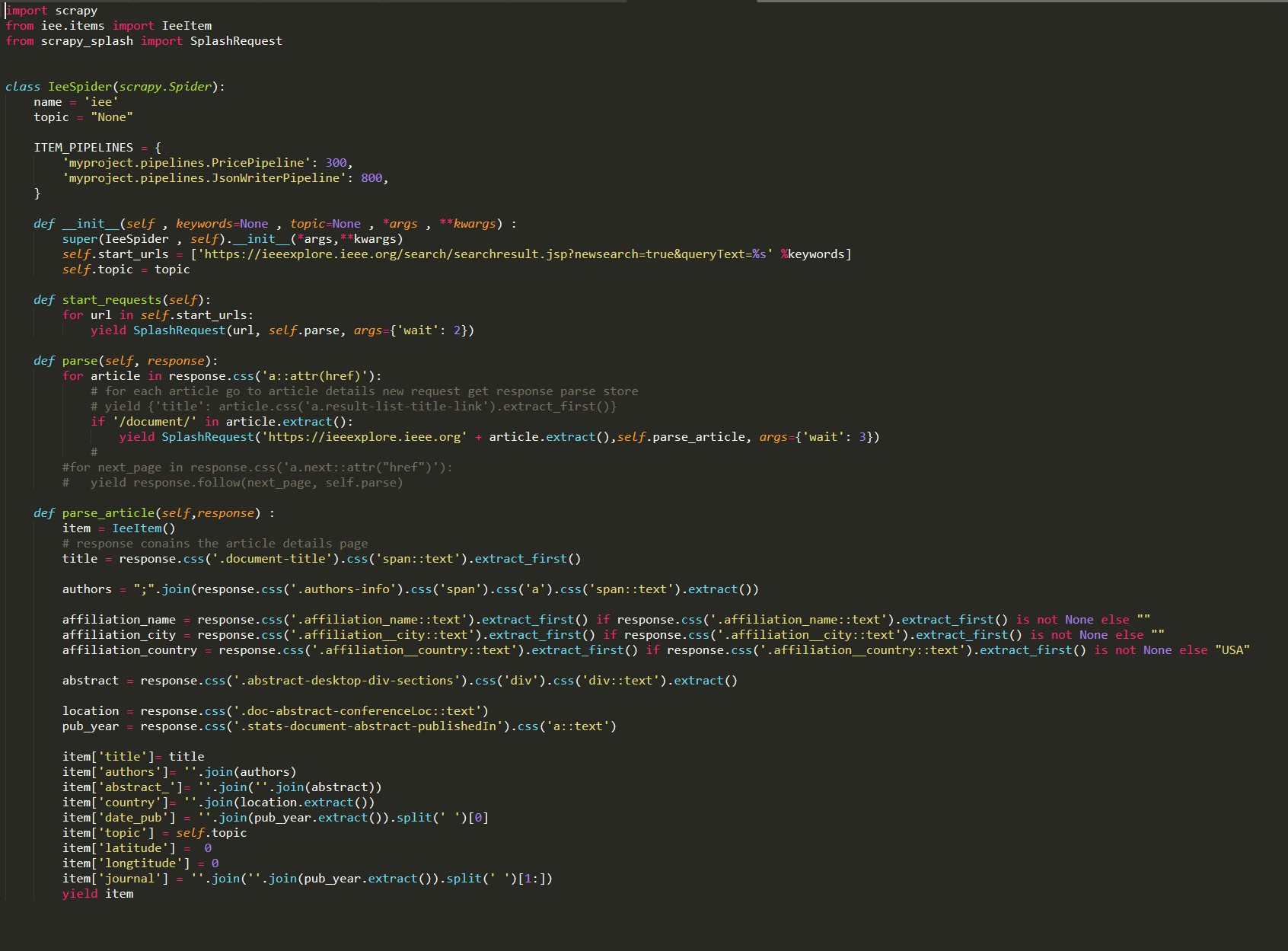
**piplines.py :** permet d’enchainer des étapes de traitement sur un item collecté, exemple MongoDB pipeline pour stocker l’item dans la collections « items »



**settings.py :** fichier de configuration du projet Scrapy.



**spiders :** dossier contenant les spiders à lancer pour scraper un site.



**Pour lancer le scraper :**

* cd emplacementProjet/iee
* scrapy crawl iee –a keywords= “big data”

1. **Travail à faire :** 
   1. **Collecter les données à partir de IEEE Xplor en utilisant le script fourni**
   2. **Créer un nouveau projet scrapy pour collecter les données de Science Direct**
   3. **Créer un nouveau projet scrapy pour collecter les données de ACM Digital Library**
   4. **Pensez à faire du nettoyage avant le stockage des données.**

**Webographie :** pour démarrer

<https://www.consultingit.fr/es/tutorial-scrapping-web-en-python-avec-scrapy>

<https://www.youtube.com/watch?v=TtCVYZRBv-M>

<https://www.edureka.co/blog/scrapy-tutorial/>

<https://steemit.com/howto/@ertinfagor/scrapy-in-docker-install>

<https://realpython.com/web-scraping-with-scrapy-and-mongodb/>

<https://medium.com/@loo.zingzai/write-your-web-crawler-using-scrapy-37c51aafa30b>