Scrapy TP

🕷️ Sujet : Extraction des Données KBO via Scrapy

🎯 Objectif :

Développez **trois spiders spécialisés**, chacun dédié à une source officielle différente, afin d'extraire les données publiques d'entreprises belges à partir d'une liste de numéros d'entreprises contenue dans un fichier CSV. Les données extraites seront structurées et sauvegardées dans une base de données **MongoDB**.

📄 Données d'entrée :

Un fichier CSV `entreprises.csv` contenant une colonne avec les numéros d'entreprises à traiter.

🕷️ Spider 1 : `kbo_spider` – Informations générales (KBO)

🔗 URL :

> https://kbopub.economie.fgov.be/kbopub/toonondernemingps.html?ondernemingsnummer=<NUMERO_ENTREPRISE>
>
> (autre example https://kbopub.economie.fgov.be/kbopub/toonondernemingps.html?ondernemingsnummer=203430576)

(Ajoutez un header de langue approprié pour obtenir la version francophone)

📦 Données à extraire (si elles existent) :

- **Généralités**
- **Fonctions**
- **Capacités entrepreneuriales**
- **Qualités**

- **Autorisations**
- **NACE code** 2025 / 2008 / 2003
- **Données financières**
- **Liens entre entités**
- **Liens externes**

Chaque entreprise sera sauvegardée dans MongoDB sous forme d'un **objet unique** avec tous ces champs, organisés en sections claires.

🕷️ Spider 2 : `ejustice` – Publications Moniteur Belge

🔗 URL :

```
https://www.ejustice.just.fgov.be/cgi_tsv/list.pl?btw=<NUMERO_ENTREPRISE>
```

📦 Données à extraire pour chaque publication :

- Numéro de publication
- Titre de la publication + Code
- Adresse de publication
- Type de publication
- Date de publication
- Référence de publication
- URL de l'image (s'il y a une image liée)

Les publications doivent être associées à l'entreprise correspondante dans MongoDB.

🕷️ Spider 3 : `consult` – Comptes annuels (NBB)

🔗 URL :

```
https://consult.cbso.nbb.be/consult-enterprise/<NUMERO_ENTREPRISE>
```

📦 Données à extraire pour chaque dépôt :

- Titre de la publication
- Référence de la publication
- **Date de dépôt**
- **Date de fin d’exercice**
- **Langue du document**
