Pratique de la data science

April 4, 2025

1 TP 6 - Scrapping de News financières

Librairies à installer :

```
import requests
import json
from datetime import datetime, timedelta
import os
import pandas as pd
```

Objectif : Pour chaque entreprise, récupérer les dernières actualités à partir de NewsAPI, les filtrer par nom de l'entreprise, et stocker les articles pertinents (titre, description, source, date de publication) dans un fichier JSON.

Tout doit être codé sous forme de fonctions.

1.1 Obtention d'une clé NewsAPI

NewAPI est une API de scrapping d'articles de presse permettant d'accéder à des centaines de sources internationales fiables.

Elle permet notamment de :

- Rechercher des articles en fonction de mots-clés (ex: nom d'une entreprise),
- Filtrer par date de publication, langue, ou source d'information,
- Obtenir pour chaque article : le titre, la description, la date, la source et l'URL.

Créer un compte News API et récupérer sa clé personnelle (dans le tableau de bord).

Remarque : La version gratuite permet un usage limité (100 requêtes/jour et 100 articles/requête). Pour entrainement et mise en production d'un modèle, il faut un scrapping quotidien.

1.2 Scrapping de News

Création d'une fonction `get_news_by_date(company_name)` pour scrapping et sauvegarde de News.

1.2.1 Initialisation des paramètres

Exemple d'initialisation de paramètres pour la requête (Possibilité de choisir d'autres sources, plages de temps....)

```
url = 'https://newsapi.org/v2/everything'
last_day = datetime.today().strftime('%Y-%m-%d')
first_day = datetime.today() - timedelta(days=10)
news_dict = {}
```

```
api_key = ""

params = {
"sources": 'financial-post, the-wall-street-journal, bloomberg, the-washington-post, australian-finan
"q": company_name,
"apiKey": api_key,
"language": "en",
"pageSize": 100,
"from": first_day,
"to": last_day,
}
```

1.3 Extraction de news

1. Lancer une requête `GET` avec `requests.get(url, params=params)`.
2. Si la requête a réussi (`status_code == 200`), parcourir la liste d'articles reçus :
   - Récupérer : `title`, `description`, `publishedAt`, `source.name`.
   - Vérifier que l'article mentionne bien l'entreprise dans le titre ou la description.
   - Extraire la date (`publishedAt.split("T")[0]`).
   - Ajouter l'article au dictionnaire `news_dict`
3. Retourner le dictionnaire `news_dict` organisé par date et le sauvegarder dans un fichier JSON.

Remarque : Mise à jour quotidienne des actualités
Possibilité d'automatiser ce scrapping tous les jours (par exemple via une tâche planifiée) en modifiant la fonction `get_news_by_date(company_name)` :

- Avant chaque requête, charger le fichier JSON existant contenant les actualités précédentes via une fonction comme `load_existing_news(company)`.
- Comparer les titres des articles reçus avec ceux déjà présents pour une même date dans le fichier.
- Ajouter uniquement les nouveaux articles (par exemple si le titre n'existe pas déjà dans la liste du jour).
- Réécrire le fichier JSON mis à jour.