PROJET PYTHON

SCRAPPER IMDB

https://github.com/Naimau-del/scrapper

NEMO TORRES

Dans le contexte de cet exercice nous devions mettre en place un logiciel de scrapping pour récupérer les données d'un site, nous étions libre du choix du site à condition que ce site soit fait pour se faire scrapper et donc ne pas être dans l'illégalité.
Notre choix s'est porté sur IMDB et nous avons décidé de récupérer les 100 meilleurs films de chaque genre pour ensuite les analyser.

Nous avons décidé d'utiliser la librairie Selenium qui possède les outils nécessaire au scrapping. En inspectant des pages et sélectionnant les bons éléments on peut faire une boucle qui prend toutes les infos des différents films.

```
for genre in genres:
    data = {}
    log(f"Getting data for {genre}")
    driver.get(f"https://www.imdb.com/search/title/?title_type=feature&genres={genre}%2C%21documentary%2C%21short")
    for i in range(1, 51):
        name = driver.find_element(By.CSS_SELECTOR, f"li:nth-child({i}) > div > div > div > div > div.dli-title > a > h3").text
        name = re.sub(r".*.\..", "", name)
        try:
            data[name] = {"year": driver.find_element(By.CSS_SELECTOR, f"li:nth-child({i}) > div > div > div > div > div.dli-title-metadata > span:nth-child(1)"
        except:
            data[name] = {"year": "0"}

        try:
            r = str(driver.find_element(By.CSS_SELECTOR, f"li:nth-child({i}) > div > div > div > div > div > span > div > span").text)
        except:
            r = "0"
        m = re.search(r"(\d+\,\d+)", r)
        if m:
            data[name]["rating"] = m.group(1).replace(",", ".")
        else:
            data[name]["rating"] = 0.0
        if "genres" in data[name]:
            data[name]["genres"].append(genre)
        else:
            data[name]["genres"] = [genre]
        log(f"{name} : {data[name]}")
    datas.append(data)
```

Nous avons ensuite utilisé Matplotlib pour réaliser des graphes.

```
# Charger les données du fichier JSON
with open("datas.json", "r", encoding="utf-8") as file:
    datas = json.load(file)

# Créer une liste d'objets Film à partir des données chargées
films = []
for film_data in datas:
    for film_name, data in film_data.items():
        film = Film(film_name, int(data["year"]), float(data["rating"]), data["genres"])
        films.append(film)

# Créer un dictionnaire pour stocker les notes moyennes par année pour chaque genre
notes_moyennes_par_genre_et_annee = {}
for film in films:
    for genre in film.genres:
        if genre not in notes_moyennes_par_genre_et_annee:
            notes_moyennes_par_genre_et_annee[genre] = {}
        if film.year not in notes_moyennes_par_genre_et_annee[genre]:
            notes_moyennes_par_genre_et_annee[genre][film.year] = [film.rating]
        else:
            notes_moyennes_par_genre_et_annee[genre][film.year].append(film.rating)

# Trier les années par ordre croissant
sorted_years = sorted(set(year for genre in notes_moyennes_par_genre_et_annee for year in notes_moyennes_par_genre_et_annee[genre]))

# Créer un graphique pour chaque genre
for genre, notes_moyennes_par_annee in notes_moyennes_par_genre_et_annee.items():
    plt.figure(figsize=(10, 5))
    moyennes_par_annee = [np.mean(notes_moyennes_par_annee[year]) if year in notes_moyennes_par_annee else np.nan for year in sorted_years]
    plt.plot(sorted_years, moyennes_par_annee, marker='o', linestyle='-')
    plt.title(f'Note moyenne par année pour le genre {genre}')
    plt.xlabel('Année')
    plt.ylabel('Note moyenne')
    plt.grid(True)
    plt.show()
```

Gérer le patrimoine informatique	Travailler en mode projet	Organiser son développement professionnel
Nous avons cherché quel scrapper utilisé entre Selenium ou d’autres tels que BeautifulSoup nous avons fait le pour et le contre de chaque librairie	Nous avons mis en place un repository Github qui nous a permis de travailler ensemble et se répartir le travail. Cela a été utile quand j’ai voulu commencer à faire un graphe avec un échantillon des résultats du scrapper	Ce projet était très intéressant car il traitait de l’éthique du scrapping et nous a demandé des recherches pour trouver qu’es ce qui est scrappable ou pas. Je pense que cette notion est importante pour tout informaticien