HW03 - Web Scrapping

```
# install gazpacho library
!pip install gazpacho

## import function
from gazpacho import Soup
import requests

# IMDB website
url = "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
html = requests.get(url)

imdb = Soup(html.text)
```

```
Requirement already satisfied: gazpacho in /usr/local/lib/python3.10/dist-packages (1.1)
```

```
# Movie Title
titles = imdb.find("h3", {"class": "lister-item-header"})
clean_titles = [title.strip() for title in titles]
```

```
# Movie Rate Type
rate_types = imdb.find("span", {"class": "certificate"})
clean_rate_types = [rate_type.strip() for rate_type in rate_types]
```

```
# Movie Length
lengths = imdb.find("span", {"class": "runtime"})
clean_lengths = [length.strip() for length in lengths]
```

```
# Movie Genre
genres = imdb.find("span", {"class": "genre"})
clean_genres = [genre.strip() for genre in genres]
```

```
# Movie Rating
ratings = imdb.find("div", {"class": "ratings-imdb-rating"})
clean_ratings = [float(rating.strip()) for rating in ratings]
```

```
# # Movie Release Year
years = imdb.find("span", {"class": "lister-item-year"})
clean_years = [int(year.strip().replace('(','').replace(')','')) for year in years]
```

```
# create DataFrame
import pandas as pd

# movie_database
movie_db = pd.DataFrame(data ={
    "Title": clean_titles,
    "Rate_type": clean_rate_types,
    "Length": clean_lengths,
    "Genre": clean_genres,
    "Rating": clean_ratings,
    "Released Year": clean_years
    })

movie_db.head()
```

	Title	Rate_type	Length	Genre	Rating	Released Year
0	1. The Shawshank Redemption (1994)	R	142 min	Drama	9.3	1994
1	2. The Godfather (1972)	R	175 min	Crime, Drama	9.2	1972
2	3. The Dark Knight (2008)	PG-13	152 min	Action, Crime, Drama	9.0	2008
3	4. Schindler's List (1993)	R	195 min	Biography, Drama, History	9.0	1993
4	5. The Lord of the Rings: The Return of the Ki...	PG-13	201 min	Action, Adventure, Drama	9.0	2003