

# Ingegneria Dei Dati - Homework 4

Davide Moliterno - 537969

16 Novembre 2022

## 1 Introduzione

Il seguente progetto ha lo scopo di effettuare l'operazione di "**Web Data Extraction**", ossia di estrarre informazioni d'interesse su pagine relative allo stesso campo da differenti sorgenti web. Ciascuna caratteristica saliente è stata ottenuta per mezzo di specifiche espressioni **XPath** e testata successivamente su differenti entità del medesimo tipo. Nello specifico nella sezione [2](#) viene analizzato come sono state individuate le migliori espressioni XPath mentre nelle sezioni [3](#) e [4](#) vengono mostrate le informazioni di interesse per i due differenti esercizi e le relative istruzioni utilizzate.

## 2 Formato delle istruzioni

Ciascuna espressione ideata si basa sul medesimo principio per cui, ci si posiziona prima su un elemento invariante il più possibile vicino alla caratteristica di interesse e successivamente ci si muove verso essa. Per tale motivo ciascuna istruzione XPath comincia per un `"/"` per effettuare la ricerca all'interno dell'intera pagina. Successivamente una parte contenente `"contains"` o `"text()="` per posizionarsi in un punto vicino alla caratteristica di interesse che sia presente in ogni pagina relativa a quella tipologia di prodotto. Infine la parte terminale è costituita dal percorso complementare necessario per arrivare all'informazione di interesse. Di seguito viene mostrato un esempio di espressione che utilizza un formato della tipologia sopracitata (l'espressione ricava il titolo dei prodotti della tipologia videogiochi all'interno del sito Amazon.it).

```
$x("//*[contains(@id,'productTitle')]/text())"
```

### 3 WDE per videogiochi da Amazon

Tale sezione ha lo scopo di analizzare il lavoro svolto per la realizzazione del primo esercizio richiesto all'interno del progetto, quello di individuare una categoria di prodotti di interesse all'interno del sito web di [Amazon](#) ed estrapolare almeno cinque caratteristiche per ciascun prodotto. Per tale esercizio la categoria sulla quale è stato svolto il lavoro è quella relativa ai [videogiochi](#). Nello specifico di seguito vengono analizzate le caratteristiche di interesse individuate ed il prodotto su cui sono state ideate e testate le istruzioni XPath che seguono lo schema introdotto nella sezione [2](#).

#### 3.1 Prodotto tipo e caratteristiche scelte

Le istruzioni utilizzate sono state ideate sulla base della pagina web di [Fifa 23](#) e successivamente sono state testate su diversi altri prodotti simili. Le caratteristiche fondamentali di interesse individuate sono state le seguenti:

- Nome;
- Prezzo;
- Età consigliata;
- Lingua;
- Data di uscita;
- Piattaforma;
- Edizione;
- Paese di origine.

Occorre far notare come alcune di queste caratteristiche non sempre sono presenti per ciascun prodotto. Ad esempio alcune entità non presentano edizioni differenti e per tale motivo la proprietà è assente dal sito mentre per altre Amazon non fornisce alcuna informazione come ad esempio per il paese di origine. Le istruzioni XPath individuate sono quindi le seguenti:

Espressioni XPath per videogiochi su Amazon	
Caratteristica	Espressione XPath
Nome	<code>\$x("//*[contains(@id,'productTitle')]/text())"</code>
Prezzo	<code>\$x("//*[contains(@id,'corePriceDisplay_desktop_feature_div')]/div[1]/*span[1]/text())[1]"</code>
Età consigliata	<code>\$x("//span[contains(text(),'Età consigliata')]/following-sibling::*text())"</code>
Lingua	<code>\$x("//span[contains(text(),'Lingua')]/following-sibling::*text())"</code>
Data d'uscita	<code>\$x("//span[contains(text(),'Data')]/following-sibling::*text())"</code>
Piattaforma	<code>\$x("//*[contains(text(),'Piattaforma:')]../text()[2]"</code>
Edizione	<code>\$x("//*[text()=' Edizione: ']/*text())"</code>
Paese di origine	<code>\$x("//*[contains(text(),'Paese di origine')]/following-sibling::*text())"</code>

Infine tali espressioni sono state verificate sui seguenti prodotti:

- "God of War";
- "Cities Skylines";
- "Uncharted 4";
- "The Last of Us 2";
- "Horizon: Forbidden West";
- "Death Stranding";
- "Spider-Man Miles Morales";
- "Ratchet & Clank: Rift Apart";

- ”Gotham Knights”.

## 4 WDE per piloti di Formula 1

Tale sezione ha lo scopo di analizzare il lavoro svolto per la realizzazione del secondo esercizio del progetto, quello di individuare una entità di interesse e una serie di sorgenti web da cui poter estrarre un insieme di caratteristiche utili per ciascuna entità. Nello specifico l’entità sulla quale è stato svolto il lavoro è quella relativa ai **piloti di Formula 1**. Le informazioni principali individuate sono state le seguenti:

- Team;
- Numero;
- Data di nascita;
- Nazione;
- World championships.

e in ciascuna sottosezione verrà analizzato come tali informazioni sono state estrapolate per ciascun sito web indicandone il link e l’entità su cui sono state ideate le differenti espressioni. Anche in questo caso, così come fatto per l’esercizio precedente, le espressioni XPath seguono la struttura definita nella sezione [2](#).

### 4.1 Formula 1 Official

- Link per il sito web: [Formula 1](#)
- Pagina web di riferimento: [Lewis Hamilton](#).

Espressioni XPath per i piloti di Formula 1 da Formula 1 Official	
Caratteristica	Espressione XPath
Team	<code>\$x("//span[text()='Team']/../following-sibling::*[1]/text() ")</code>
Numero	<code>\$x("//*[contains(@class,'driver-number')]/*[1]/text() ")</code>
Data di nascita	<code>\$x("//span[text()='Date of birth']/../following-sibling::*[1]/text() ")</code>
Nazione	<code>\$x("//span[text()='Country']/../following-sibling::*[1]/text() ")</code>
World championships	<code>\$x("//span[text()='World Championships']/../following-sibling::*[1]/text() ")</code>

Ciascuna delle seguenti espressioni XPath è stata verificata sulle seguenti entità:

- [Max Verstappen](#);
- [Sergio Perez](#);
- [George Russell](#);
- [Fernando Alonso](#).

## 4.2 Al Volante

- Link per il sito web: [Al Volante](#)
- Pagina web di riferimento: [Lewis Hamilton](#).

Espressioni XPath per i piloti di Formula 1 da Al Volante	
Caratteristica	Espressione XPath
Team	<code>\$x("//*[contains(@class,'f1-node-car')]/preceding-sibling::*[2]/*/text())</code>
Numero	<code>\$x("//*[text()='Numero di gara']/following-sibling::*[2]/*/text() ")</code>
Data di nascita	<code>\$x("//*[text()='Data di nascita']/following-sibling::*[2]/*/text() ")</code>
Nazione	<code>\$x("//*[text()='Nazione']/following-sibling::*[2]/*/text() ")</code>
World championships	<code>\$x("//*[text()='Titoli iridati']/following-sibling::*[2]/*/text() ")</code>

Ciascuna delle seguenti espressioni XPath è stata verificata sulle seguenti entità:

- [Max Verstappen](#);
- [Sergio Perez](#);
- [George Russell](#);
- [Fernando Alonso](#).

### 4.3 Formula1.lne

- Link per il sito web: [Formula1.lne](#)
- Pagina web di riferimento: [Lewis Hamilton](#).

Espressioni XPath per i piloti di Formula 1 da Formula1.lne	
Caratteristica	Espressione XPath
Team	<code>\$x("//*[contains(@id, 'pilotos_escuderia')]/h2/*/text())</code>
Numero	<code>\$x("//*[contains(text(), 'Número')]/following-sibling::*/*text())</code>
Data di nascita	<code>\$x("//*[contains(text(), 'Fecha de Nacimiento')]/following-sibling::*/*text())</code>
Nazione	<code>\$x("//*[contains(text(), 'Nacionalidad')]/following-sibling::*/*text())</code>
World championships	<code>\$x("//*[contains(text(), 'Campeonatos')]/following-sibling::*/*text())</code>

Ciascuna delle seguenti espressioni XPath è stata verificata sulle seguenti entità:

- [Max Verstappen](#);
- [Sergio Perez](#);
- [George Russell](#);
- [Fernando Alonso](#).

#### 4.4 Motorsport Total

- Link per il sito web: [Motorsport Total](#)
- Pagina web di riferimento: [Lewis Hamilton](#).





Espressioni XPath per i piloti di Formula 1 da Motor.es	
Caratteristica	Espressione XPath
Team	<code>\$x("//*[contains(@class, 'bandera loading')]/following-sibling::*[2]/text())</code>
Numero	<code>\$x("//*[contains(text(), 'Dorsal')]/following-sibling::*[1]/text())</code>
Data di nascita	<code>\$x("//*[contains(text(), 'Fecha de nacimiento')]/following-sibling::*[1]/text())</code>
Nazione	<code>\$x("//*[contains(text(), 'Nacionalidad')]/following-sibling::*[1]/text())</code>
World championships	<code>\$x("//*[contains(text(), 'Campeonatos mundiales')]/following-sibling::*[1]/text())</code>

Ciascuna delle seguenti espressioni XPath è stata verificata sulle seguenti entità:

- [Max Verstappen](#);
- [Sergio Perez](#);
- [George Russell](#);
- [Fernando Alonso](#).