

Ingegneria Dei Dati - Homework 4

Davide Moliterno - 537969

16 Novembre 2022

Link github del progetto: <https://github.com/chila99/IDDHomeworks>

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 prodotti del medesimo campo. Nello specifico nella sezione [2](#) viene analizzato come è stato svolto il processo di Web Data Extraction(WDE) per la categoria videogiochi dal sito web di Amazon mentre nella sezione [3](#) la ricerca di informazioni relativa a i piloti di Formula 1.

2 WDE per videogiochi da Amazon

Tale sezione ha lo scopo di analizzare il lavoro svolto per la realizzazione del primo obiettivo postosi 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. Nello specifico la categoria sulla quale è stato svolto il lavoro è quella relativa ai [videogiochi](#).

2.1 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:

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

Nonché, come richiesto:

- Nome;
- Prezzo.

Occorre far notare come alcune di queste caratteristiche non è detto siano presenti per ciascun prodotto a causa della mancanza di scelta ad esempio per quella che è l'edizione o per la mancata informazione da parte del sito web di quella informazione ad esempio relativamente al paese di origini.

2.2 Formato delle istruzioni

Ciascuna espressione ideata si basa sullo stesso 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 espressione XPath comincia per un `//` per andare a ricerca all'interno dell'intera pagina e successivamente per `contains` o `text()`. La parte successiva invece è data dal percorso necessario per arrivare all'informazione di interesse. Le 5 istruzioni XPath individuate sono quindi le seguenti:

Tali espressioni sono state verificate su ulteriori prodotti quali: **"God of War"**, **"Cities Skylines"**, **"Uncharted 4"**, **"The Last of Us 2"**, **"**

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>

3 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.

3.1 Formula 1 Official

Il primo sito scelto è quello ufficiale di [Formula 1](#) e ciascuna informazione è stata ideata a partire dalla pagina web di [Lewis Hamilton](#). Di seguito sono mostrate le espressioni XPath utilizzate per estrarre le informazioni:

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](#)

3.2 Al Volante

Il secondo sito scelto è quello di [Al Volante](#) e ciascuna informazione è stata ideata a partire dalla pagina web di [Lewis Hamilton](#). Di seguito sono mostrate le espressioni XPath utilizzate per estrarre le informazioni:

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 champions	<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](#)

3.3 Formula1.lne

Il terzo sito scelto è quello di [Formula1.lne](#) e ciascuna informazione è stata ideata a partire dalla pagina web di [Lewis Hamilton](#). Di seguito sono mostrate le espressioni XPath utilizzate per estrarre le informazioni:

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](#)

3.4 Motorsport Total

Il quarto sito scelto è quello di [Motorsport Total](#) e ciascuna informazione è stata ideata a partire dalla pagina web di [Lewis Hamilton](#). Di seguito sono mostrate le espressioni XPath utilizzate per estrarre le informazioni:

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

Espressioni XPath per i piloti di Formula 1 da Motorsport Total	
Caratteristica	Espressione XPath
Team	<code>\$x("//*[contains(text(), 'Aktuelles Team')]/following-sibling::*/*/text())"</code>
Numero	<code>\$x("//*[text()='Startnummer']/following-sibling::*/*/text())"</code>
Data di nascita	<code>\$x("//*[text()='Geburtsdatum']/following-sibling::*/*/text())"</code>
Nazione	<code>\$x("//*[text()='Nationalität']/following-sibling::*/*/text())"</code>
World championships	N.D.

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

3.5 Motor.es

Il quinto sito scelto è quello di [Motor.es](#) e ciascuna informazione è stata ideata a partire dalla pagina web di [Lewis Hamilton](#). Di seguito sono mostrate le espressioni XPath utilizzate per estrarre le informazioni:

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

- [Max Verstappen](#)
- [Sergio Perez](#)
- [George Russell](#)

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>

- [Fernando Alonso](#)