Ingegneria Dei Dati - Homework 4

Davide Molitierno - 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 prodotti del medesimo tipo. Nello specifico nella sezione 2 viene analizzato come è stato svolto il processo in relazione alla categoria videogiochi dal sito web di Amazon mentre nella sezione 3 per l'estrapolazione di informazioni relative ai 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 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. 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:

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

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 caratteristica come ad esempio il paese di origine.

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 effetuare la ricerca all'interno dell'intera pagina e successivamente **contains** o **text()=** per posizionarsi in una parte vicina alla caratteristica di interesse che fosse presente in ogni pagina relativa a quella tipologia di prodotto. La parte successiva invece è data dal percorso necessario per arrivare all'informazione di interesse. Le 5 istruzioni XPath individuate sono quindi le seguente:

Tali espressioni sono state verificate sui seguenti prodotti:

- **"God of War"**;
- **"Cities Skylines"**;
- **"Uncharted 4"**;

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

- **"The Last of Us 2"**;
- **"Horizon: Forbidden West"**;
- **"Death Stranding"**;
- **"Spider-Man Miles Morales"**;
- **"Ratchet & Clank: Rift Apart"**;
- **"Gotham Knights"**.

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 indicandone il link e l'entità su cui sono state ideate le differenti espressioni.

3.1 Formato delle istruzioni

Anche in questo esercizio le istruzioni XPath utilizzate seguono la stessa logica utilizzata per la realizzazione dell'esercizio 1 e per tale motivo si rimanda alla sezione 2.2 per una descrizione più dettagliata del processo.

3.2 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	$x("//span[text()='Team']/../following-sibling::*/text() ")
Numero	$x("//*[contains(@class,'driver-number')]/*[1]/text()")
Data di nascita	$x("//span[text()='Date of birth']/../following-sibling::*/text() ")
Nazione	$x("//span[text()='Country']/../following-sibling::*/text() ")
World championships	$x("//span[text()='World Championships']/../following-sibling::*/text() ")

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

- Max Verstappen;
- Sergio Perez;
- George Russell;
- Fernando Alonso.

3.3 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	$x("//*[contains(@class,'f1-node-car')]/preceding-sibling::*[2]/*/text()")
Numero	$x("//*[text()='Numero di gara']/following-sibling::*/text() ")
Data di nascita	$x("//*[text()='Data di nascita']/following-sibling::*/text() ")
Nazione	$x("//*[text()='Nazione']/following-sibling::*/text() ")
World championships	$x("//*[text()='Titoli iridati']/following-sibling::*/text() ")

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

- Max Verstappen;
- Sergio Perez;
- George Russell;
- Fernando Alonso.

3.4 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	$x("//*[contains(@id, 'pilotos_escuderia')]/h2/*/text()")
Numero	$x("//*[contains(text(), 'Número')]/following-sibling::*/text()")
Data di nascita	$x("//*[contains(text(), 'Fecha de Nacimiento')]/following-sibling::*/text()")
Nazione	$x("//*[contains(text(), 'Nacionalidad')]/following-sibling::*/text()")
World championships	$x("//*[contains(text(), 'Campeonatos')]/following-sibling::*/text()")

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

- Max Verstappen;
- Sergio Perez;
- George Russell;
- Fernando Alonso.

3.5 Motorsport Total

- Link per il sito web: Motorsport Total
- Pagina web di riferimento: Lewis Hamilton.

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

Da notare come in questa pagina web il dato relativo al numero di campionati mondiali vinti non fosse disponibile e per tale motivo la relativa espressione XPath non è stata riportata.

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

- Max Verstappen;
- Sergio Perez;
- George Russell;
- Fernando Alonso.

3.6 Motor.es

- Link per il sito web: Motor.es
- Pagina web di riferimento: Lewis Hamilton.

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

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

- Max Verstappen;
- Sergio Perez;
- George Russell;
- Fernando Alonso.