Ingegneria Dei Dati - Homework 4

Davide Molitierno - 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 pagine 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 seguente:

- 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()")
- Nome: $x("//*[contains(@id,'productTitle')]/text()")
- Prezzo: $x("(//*[contains(@id,'corePriceDisplay_desktop_feature_div')]/div[1]/*/span[1]/tex

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

3 WDE per piloti di Formula 1