**Dokumentáció a Web Scraper Programhoz**

**Program Neve: CPU Extractor**

**Célja:**

Ez a Python program az **arukereso.hu** oldalról gyűjt adatokat AMD processzorokról, beleértve a termék nevét, műszaki adatait és árát. Az adatokat egy Cpu\_extract.csv nevű fájlba exportálja táblázatos formában.

**Működési Leírás**

1. **Adatok Gyűjtése:**
	* A program egy előre megadott URL-t használ (jelen esetben az AMD processzorok kategóriaoldalát az arukereso.hu-n).
	* A program a requests könyvtár segítségével HTTP GET kérést küld, hogy letöltse az oldal HTML-tartalmát.
2. **HTML Elemzés:**
	* A letöltött HTML-t a BeautifulSoup könyvtár dolgozza fel.
	* A program kinyeri az oldalról a releváns adatokat, például a processzorok nevét, magok számát, szálak számát, foglalat típusát, órajelet, Turbo órajelet és az árat.
3. **Adatok Mentése CSV-be:**
	* Az adatokat egy Cpu\_extract.csv nevű fájlba menti.
	* Minden termék adatai külön sorban jelennek meg, az oszlopok pedig a következőket tartalmazzák:
		+ Termék neve
		+ Magok száma
		+ Szálak száma
		+ Processzor foglalat
		+ Órajel
		+ Turbo órajel
		+ Ár

**Fájlok és Mappa Struktúra**

* Cpu\_extract.csv: Az összegyűjtött adatokat tartalmazó CSV fájl.
* A program futtatásához csak a web\_scraper.py fájl szükséges, amely tartalmazza a kódot.

**Követelmények**

**Könyvtárak Telepítése**

A program működéséhez az alábbi Python könyvtárak szükségesek:

* requests: HTTP kérések küldéséhez.
* BeautifulSoup a bs4 modulból: HTML feldolgozáshoz.
* csv: Az adatok CSV formátumba írásához (ez a Python beépített könyvtára).

Telepítésük:



**Forráskód Magyarázata**

**1. Alapvető Beállítások**

* A program az url változóban megadott kategóriaoldalról gyűjt adatokat.
* A headers változó biztosítja, hogy a kérés egy böngészőt utánozzon, ezzel elkerülve az oldalak blokkolását.



**2. HTTP Kérés és HTML Elemzés**

* A requests.get() segítségével a program letölti az oldalt.
* Ha a kérés sikeres, a BeautifulSoup elemzi a letöltött HTML-t.



**3. Adatok Kinyerése**

A program a következő adatokat gyűjti:

* **Termék neve**: A product-box osztályú <div> elem belsejében található <img> alt attribútumából.
* **Ár**: A price osztályú <div> elemből.
* **Tulajdonságok**:
	+ A description osztályú <div> elemek <li> tagekből.
	+ A <b> tagek tartalmazzák az értékeket, például a magok számát vagy az órajelet.



**4. CSV Fájlba Írás**

A program az összegyűjtött adatokat egy CSV fájlba menti. Minden adat egy sorban jelenik meg, oszlopokba rendezve.



**Kimenet Példa**

A generált Cpu\_extract.csv fájl tartalma:

Termék neve,Magok száma,Szálak száma,Foglalat,Órajel,Turbo órajel,Ár

AMD Ryzen 5 7600X,6 magos,12,AMD Socket AM5,3900 MHz,5400 MHz,104 890 Ft-tól

AMD Ryzen 5 5600,6 magos,12,AMD Socket AM4,3500 MHz,4400 MHz,95 000 Ft-tól

**Lehetséges Hibák és Megoldások**

1. **404 vagy 403 HTTP Kód**:
	* Ellenőrizd az URL-t, vagy frissítsd a headers tartalmát.
2. **Hiányzó Adatok**:
	* Ha a product-box, price vagy description elemek nem találhatók, az oldal HTML-struktúrája változhatott. Frissítsd a kódot az aktuális struktúrához.

**Továbbfejlesztési Lehetőségek**

* Több termékkategória egyidejű feldolgozása.
* A dinamikusan betöltődő tartalom (JavaScript) kezelésére Selenium integrálása.