Fabacus
12 Soho Square, London, W1D 3QF

Data Engineer/Scientist technical task

Objective

The main goal is to create a script(s) that collect data from different sources and produce a consistent output from where further data analysis can take place.

Description

You have been asked to collect/scrape all identifiable product listings related to the Jurassic Park franchise (e.g. all associated IP, including the more recent "Jurassic World" reboot) from 3 different retailer sites. You must then organise the data set by "unique" product, using and referencing an appropriate product ID (where possible, your dataset must include the product barcode number (i.e. GTIN/UPC/EAN) regardless).

Ouce you have the data in a single repository, we would like you to analyse the data set to be able to answer the following questions:

- Are there any individual products that are listed across the 3 sites? If so, please provide a list with links to each site.
- Which brands and/or manufacturers are associated to the products?
- Bonus question: Can you identify any specific IP or movie characters are part of the products (i.e. "Blue" Raptor from Jurassic World) from each of the products?
- What other additional interesting insights you think may be relevant to the Jurassic Park Franchise owners (NBC Universal)?

The list of retailer sites we would like you to use are:

`www.amazon.co.uk`

`www.argos.co.uk`

`www.smythstoys.com`

Considerations

To keep the set of records to a manageable size, please search on the retailer sites for:

`jurassic world toys`

Deliverables

- Please provide the list of matching products from each retailer site.
- Please provide the answers in a document, including any assumptions and decisions you have made.
- Please provide the source code for any scripts you needed to complete your task.