

Information Integration on the Web Assignment 3 : Wrapper

1) Base URI used for extracting information: <http://www.bestbuy.com/>

Best Buy is the ecommerce website similar to Amazon.com which is used to buy and sell electronics and other goods item.

Field	Description
product	Name of the product to be sold. Eg. G50 Black Lenovo Laptop
price	Price of the product. Eg. \$1250
Rating	Rating of the product. Eg. 4.4
Model	Model of the product Eg. 65656GNO33
sku	SKU ID of the product Eg. 2121212122

The screenshot shows the Best Buy website interface. The URL in the browser is www.bestbuy.com/site/searchpage.jsp?st=shoes&_dyncharset=UTF-8&id=pcat17071&type=page&sc=Global&cp=1&nrp=&sp=&qp=&list=n&af=tn. The page displays a search result for "In Her Shoes [WS] (DVD) (Enhanced Widescreen for 16x9 TV) (English/French/Spanish) 2005". Hand-drawn arrows highlight specific fields: "Product" points to the product title, "Release Date" points to the release date, "Rating" points to the star rating, and "Price" points to the price tag. The price is \$7.99. The page also shows a "Price Match Guarantee" badge and an "Add to Cart" button. The left sidebar contains navigation links for "Shop Category" (Music (CDs & Vinyl), Movies & TV Shows, Laptop Accessories), "Current Deals", "Price", "Subgenre", and "Award Year".

2/3) Tool used for scrapping as well as wrapping was Portia.

Link : <https://portia-beta.scrapinghub.com/#/projects/105366>

Project Name : bestbuy

Spiders : www.bestbuy.com

Dataformat :

iphone-case---best-buy

food---best-buy

shoes---best-buy

ninja-master-prep-pro-food-and-drink-mixer-black

The wrapper used was the one provided by the scrappinghub. For using this wrapper initially, I provided some training examples like what is the field, type of field, how and where are they located in the HTML. The training data helps to identify the pattern formed within subsequent pages in terms like how the fields are extracted and what is the pattern before the extracted label and after the extracted label.

In this context the algorithm used is **Inductive Supervised Learning** since we are providing the training data in the beginning. Though I didn't have access to the code of Portia and have implementation details looking at the Inspector on the Portia looks like the Wrapper model they have used is **HLRT**.