

IIW – 5 Data Cleaning

Dataset from the BestBuy.com

1) Field : URL

Description : URL of the product.

Operation : Unfolding operation was performed on this field to unfold/extract the unique product id of each product. A new column called product_id was created which have product id from the URL. Initially the product id was part of the url. Add a column based on current column operation was performed. Below is the regex/operation performed:
value.match(/.*?(id=\w+).*/)[0].replace("id=", "")

Edit Column-> Add Column based on current column.

Example :

Before :

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/appliance-parts-
accessories/abcat0916000.c?id=abcat0916000",
  "product" : "Appliance Parts & Accessories",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "N/A"
}
```

After :

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/appliance-parts-
accessories/abcat0916000.c?id=abcat0916000",
  "product_id" : "abcat0916000",
  "product" : "Appliance Parts & Accessories",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "N/A"
}
```

2) Field : Price

Description : Price of the product.

Operation: The current records are in the form of 5, 60.5, 50.99 and empty record. Standardized the record so that it has format of two decimal values and N/A if the record is empty. The operation performed is Text Facet to convert into appropriate format.

Facet->Text Facet

Example:

Before:

```
{
  "URL" : "http://www.bestbuy.com/site/buying-
guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat3253000500
16",
  "product_id" : "pcmcat325300050016",
  "product" : "Vacuum Cleaner Buying Guide",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "60"
}
```

After:

```
{
  "URL" : "http://www.bestbuy.com/site/buying-
guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat3253000500
16",
  "product_id" : "pcmcat325300050016",
  "product" : "Vacuum Cleaner Buying Guide",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "60.00"
}
```

3) **Field** : Model_No

Description: The Model number of the product.

Operation: The model number had some incorrect data like "Some sentence" which was not a model number. Hence all the model number which had spaces require removed as incorrect Model no. Transform operation was performed to achieve this.

Text Filter(\.*?\s+.*)->Text Filter

Example:

Before:

```
{"URL": "http://www.bestbuy.com/site/home-appliances/ranges-ovens-
stoves/abcat0904000.c?id=abcat0904000", "product": "Ranges, Cooktops & Ovens",
"model": "Shop All \u203a"}
```

After:

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/ranges-ovens-
stoves/abcat0904000.c?id=abcat0904000",
  "product_id" : "abcat0904000",
  "product" : "Ranges, Cooktops & Ovens",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
}
```

```
    "price" : "N/A"
  }
```

4) **Field** : SKU

Description: SKU number of the product.

Operation: Whichever blank SKU number were there they were converted to N/A using facet operation.

Facet->Text Facet

Example:

Before:

The SKU field was missing.

After:

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/ranges-ovens-
stoves/abcat0904000.c?id=abcat0904000",
  "product_id" : "abcat0904000",
  "product" : "Ranges, Cooktops & Ovens",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "N/A"
}
```

5) **Field** : Rating

Description: Rating of the product.

Operation: Text Facet was performed for empty record as "0.0"

Facet->Text Facet.

Example:

Before:

There was no empty rating field for this record.

After:

```
{
  "URL" : "http://www.bestbuy.com/site/buying-
guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat3253000500
16",
  "product_id" : "pcmcat325300050016",
  "product" : "Vacuum Cleaner Buying Guide",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "60.00"
}
```

Before:

Facebook | Inbox (170) | neel_shah | Inbox (26) | WhatsApp | USC-CSC | hv6-clean | Original - C | 2016-10-15 | Nimesh

127.0.0.1:3333/project?project=1899054657683

Refine Original Permalink

Facet / Filter Undo / Redo 0

473 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: first previous 51 100 next last

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	product	model	sku	rating	price
ibuy.com/site/cyber-acoustics-2-1-speaker-system-3-piece-p7skuld=1430483	Cyber Acoustics - 2.1 Speaker System (3-Piece) - Black	Ca3001	1430483	3.2	24.99
ibuy.com/site/creative-gigaworks-t40-series-ii-2-0-speaker-system-2-piece-316421 p7skuld=9616421	Creative - GigaWorks T40 Series II 2.0 Speaker System (2-Piece) - Black/Yellow	T40	9616421	4.4	132.99
ibuy.com/site/key-digital-digital-iq-series-3-5mm-stereo-male-to-3-5mm-rca-arc-jumper-cables-4-pack-black-red/8413028 p7skuld=8413028	Key Digital - Digital IQ Series 3.5mm Stereo Male-to-3.5mm Stereo Female/RCA ARC Jumper Cables (4-Pack) - Black/Red	KD-35MMAIRH	8413028		
ibuy.com/site/mobile-phones-promotions/Cell-Phone-Buying-113200050000 c7id=pcmc:at313200050000&type=category	Cell Phone Buying Guide				50
ibuy.com/site/logitech-mk320-wireless-keyboard-and-mouse-p7skuld=1294039	Logitech - MK320 Wireless Keyboard and Mouse - Black	920-002836	1294039	4.5	24.99
ibuy.com/site/theaterseatstore-magnolia-4-seat-curved-leather-home-theater-2796242 p7skuld=2796242	TheaterSeatStore - Magnolia 4-Seat Curved Leather Home Theater Seating - Brown	Magnolia-R4CP-BR	2796242	5.0	
ibuy.com/site/theaterseatstore-magnolia-4-seat-curved-leather-home-theater-2796233 p7skuld=2796233	TheaterSeatStore - Magnolia 4-Seat Curved Leather Home Theater Seating - Black	Magnolia-R4CM-BL	2796233	5.0	
ibuy.com/site/computing-promotions/designer-942300050000 c7100050000&pageType=REDIRECT&isole=1&searchRedirect=designer+series					0
ibuy.com/site/logitech-m310-wireless-mouse-silver/1297054 p7skuld=1297054	Logitech - M310 Wireless Mouse - Silver	M310	1297054	4.7	14.99
ibuy.com/site/promo/designer-gift-sets-146163		JS-GFTSET-001	4563400	4.2	29.99
ibuy.com/site/jack-spade-gift-box-hard-shell-case-with-portable-charger-for-6-and-6s-black-gray/4563400 p7skuld=4563400	JACK SPADE - Gift Box - Hard Shell Case with Portable Charger for Apple® iPhone® 6 and 6s - Black/Gray	JS-GFTSET-001	4563400	4.2	
ibuy.com/site/logitech-k400-plus-wireless-keyboard-black/7575039 p7skuld=7575039	Logitech - K400 Plus Wireless Keyboard - Black	920-007119	7575039	4.5	29.99
ibuy.com/site/kate-spade-new-york-moto-mod-portable-charger-stripe-2-16-16-406130 p7skuld=406130	kate spade new york - Moto Mod Portable Charger - Stripe 2	KSMO-006-STTGF			
ibuy.com/site/logitech-m310-wireless-optical-mouse-peacock-p7skuld=9927955	Logitech - M310 Wireless Optical Mouse - Peacock Blue	910-001917	9927955	4.7	14.99
ibuy.com/site/mobile-phone-accessories/cell-phone-				3	edit

extraction.json.txt Show all

Ask me anything 7:09 PM 10/28/2016

After:

Facebook | Inbox (170) | neel_shah | Inbox (26) | WhatsApp | USC-CSC | hv6-clean | extraction | 2016-10-15 | Nimesh

127.0.0.1:3333/project?project=1875585962452

Refine extraction json Permalink

Facet / Filter Undo / Redo 18

473 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: first previous 51 100 next last

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

RL	product_id	product	model	sku	rating	price
esbuy.com/site/cyber-acoustics-2-1-speaker-system-3-piece-83 p7skuld=1430483	1430483	Cyber Acoustics - 2.1 Speaker System (3-Piece) - Black	Ca3001	1430483	3.2	24.99
esbuy.com/site/creative-gigaworks-t40-series-ii-2-0-speaker-system-2-piece-9616421 p7skuld=9616421	9616421	Creative - GigaWorks T40 Series II 2.0 Speaker System (2-Piece) - Black/Yellow	T40	9616421	4.4	132.99
esbuy.com/site/key-digital-digital-iq-series-3-5mm-stereo-male-to-3-5mm-rca-arc-jumper-cables-4-pack-black-red/8413028 p7skuld=8413028	8413028	Key Digital - Digital IQ Series 3.5mm Stereo Male-to-3.5mm Stereo Female/RCA ARC Jumper Cables (4-Pack) - Black/Red	KD-35MMAIRH	8413028	0.0	N/A
esbuy.com/site/mobile-phones-promotions/Cell-Phone-Buying-at313200050000 c7id=pcmc:at313200050000&type=category	pcmc:at313200050000	Cell Phone Buying Guide	N/A	N/A	0.0	50.00
esbuy.com/site/logitech-mk320-wireless-keyboard-and-mouse-39 p7skuld=1294039	1294039	Logitech - MK320 Wireless Keyboard and Mouse - Black	920-002836	1294039	4.5	24.99
esbuy.com/site/theaterseatstore-magnolia-4-seat-curved-leather-home-theater-m/2796242 p7skuld=2796242	2796242	TheaterSeatStore - Magnolia 4-Seat Curved Leather Home Theater Seating - Brown	Magnolia-R4CP-BR	2796242	5.0	N/A
esbuy.com/site/theaterseatstore-magnolia-4-seat-curved-leather-home-theater-k/2796233 p7skuld=2796233	2796233	TheaterSeatStore - Magnolia 4-Seat Curved Leather Home Theater Seating - Black	Magnolia-R4CM-BL	2796233	5.0	N/A
esbuy.com/site/computing-promotions/designer-at342300050000 c712300050000&pageType=REDIRECT&isole=1&searchRedirect=designer+series	pcmc:at342300050000		N/A	N/A	0.0	0.00
esbuy.com/site/logitech-m310-wireless-mouse-silver/1297054 p7skuld=1297054	1297054	Logitech - M310 Wireless Mouse - Silver	M310	1297054	4.7	14.99
esbuy.com/site/promo/designer-gift-sets-146163			JS-GFTSET-001	4563400	4.2	29.99
esbuy.com/site/jack-spade-gift-box-hard-shell-case-with-portable-charger-for-6-and-6s-black-gray/4563400 p7skuld=4563400	4563400	JACK SPADE - Gift Box - Hard Shell Case with Portable Charger for Apple® iPhone® 6 and 6s - Black/Gray	JS-GFTSET-001	4563400	4.2	N/A
esbuy.com/site/logitech-k400-plus-wireless-keyboard-black/7575039 p7skuld=7575039	7575039	Logitech - K400 Plus Wireless Keyboard - Black	920-007119	7575039	4.5	29.99

extraction.json.txt Show all

Ask me anything 7:08 PM 10/28/2016