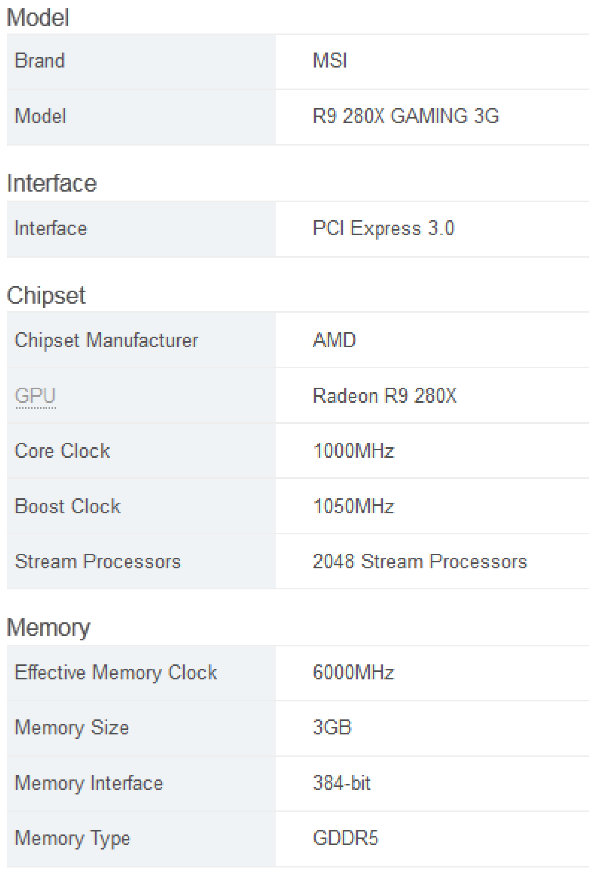
Newegg CA4002 Data mining - Slava Feoktistov - 10336661

The website I used to scrape and import into a MySQL database was newegg.com Newegg is a site that sells a wide variety of items which include PC parts. I chose to retrieve Graphic cards and scrape the specs, user rating, number of reviews and price.

The goal of this is to get the probability of which spec category (High-end, medium-end and low-end GPU's) does best in each GPU type (AMD, ATI or NVIDIA). This can be used to predict which GPU type to buy from according to the GPU specs and Ratings.

Web Scraping was done using Scrapy and Python. Each GPU has a separate page which contains specs, price, rating and number of people that rated it. The GPU's are contained within a main page that displays 20 GPU's per page and currently contains around 32 pages of GPU's so there is around 600 GPU's mined.
Each of the pages are accessed using the Scrapy Rule var that is managed by the parameters and Regex.

The spec section within the GPU page is generally structured the same, which looks like:



Model	
Brand	MSI
Model	R9 280X GAMING 3G
Interface	
Interface	PCI Express 3.0
Chipset	
Chipset Manufacturer	AMD
GPU	Radeon R9 280X
Core Clock	1000MHz
Boost Clock	1050MHz
Stream Processors	2048 Stream Processors
Memory	
Effective Memory Clock	6000MHz
Memory Size	3GB
Memory Interface	384-bit
Memory Type	GDDR5

Some GPU pages are missing vital information like prices and specs are either missing or sorted differently. The unvital specs like Core Clock or Cuda Cores that are missing are just set to default blank or zero.

Vital specs that are missing from an item are dropped within the Pipeline of the Scrapy project using the DropItem exception that is raised.

Pipeline

The Pipeline is used to create the table within MySQL DB, process the items and insert the items into the DB.

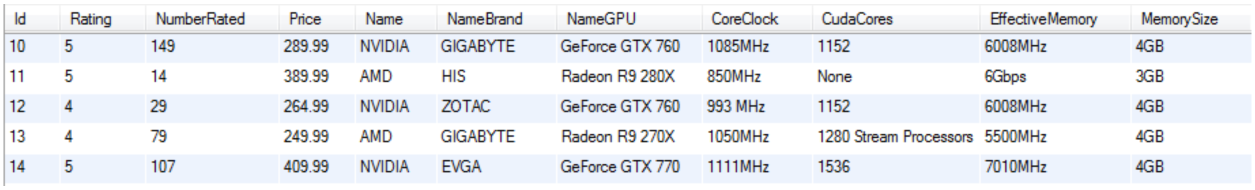
The table “items” is created within __init__() that is run once at the initialization of the spider.

Table Columns:

- Id - INT, AUTO_INCREMENT
- Rating - INT
- NumberRated - INT
- Price - FLOAT
- Name - VARCHAR(45), NOT NULL
- NameBrand - VARCHAR(45), NOT NULL
- NameGPU - VARCHAR(45), NOT NULL
- EffectiveMemory - VARCHAR(45)
- MemorySize - VARCHAR(25)
- CoreClock - VARCHAR(25)
- CudaCores - VARCHAR(25)

Each item is processed within process _item() which refines each item and sets default values if eg. a GPU spec is missing, then calls process_sql() which inserts the processed items into the table.

Table example:



Id	Rating	NumberRated	Price	Name	NameBrand	NameGPU	CoreClock	CudaCores	EffectiveMemory	MemorySize
10	5	149	289.99	NVIDIA	GIGABYTE	GeForce GTX 760	1085MHz	1152	6008MHz	4GB
11	5	14	389.99	AMD	HIS	Radeon R9 280X	850MHz	None	6Gbps	3GB
12	4	29	264.99	NVIDIA	ZOTAC	GeForce GTX 760	993 MHz	1152	6008MHz	4GB
13	4	79	249.99	AMD	GIGABYTE	Radeon R9 270X	1050MHz	1280 Stream Processors	5500MHz	4GB
14	5	107	409.99	NVIDIA	EVGA	GeForce GTX 770	1111MHz	1536	7010MHz	4GB