Data Mining Homework 2

Fabio Fraschetti 1834942

November 2023

[Fabio Fraschetti]

1 Exercise

The code is a Python script for scraping product information from Amazon based on a given keyword. I stored the results in *amazon_products.tsv* file. The script uses libraries like BeautifulSoup to investigate the html taken from the request library. The code includes a sleep of 4 seconds between each page request to avoid overloading the server. Additionally, a user agent is randomly chosen for each request to mimic different devices accessing the website. The script outputs the cleaned dataset to a new TSV file ('*amazon_productsClean.tsv*'). After I stored the file I start to make an Exploratory Data Analysis (EDA):
The EDA1 returns me:

The ranges are:
Range Gpu: MIN 7.99, MAX 10905.0
Range Supports for Gpu: MIN 2.99, MAX 233.85
Range Cable: MIN 7.0, MAX 47.0
Range Thermal paste: MIN 3.99, MAX 39.99
Range Adapter for Gpu: MIN 8.0, MAX 25.95

I divided the categories into GPU, Supports for gpu, Cables, Thermal Paste and Adaptors.

The EDA2 returns me:

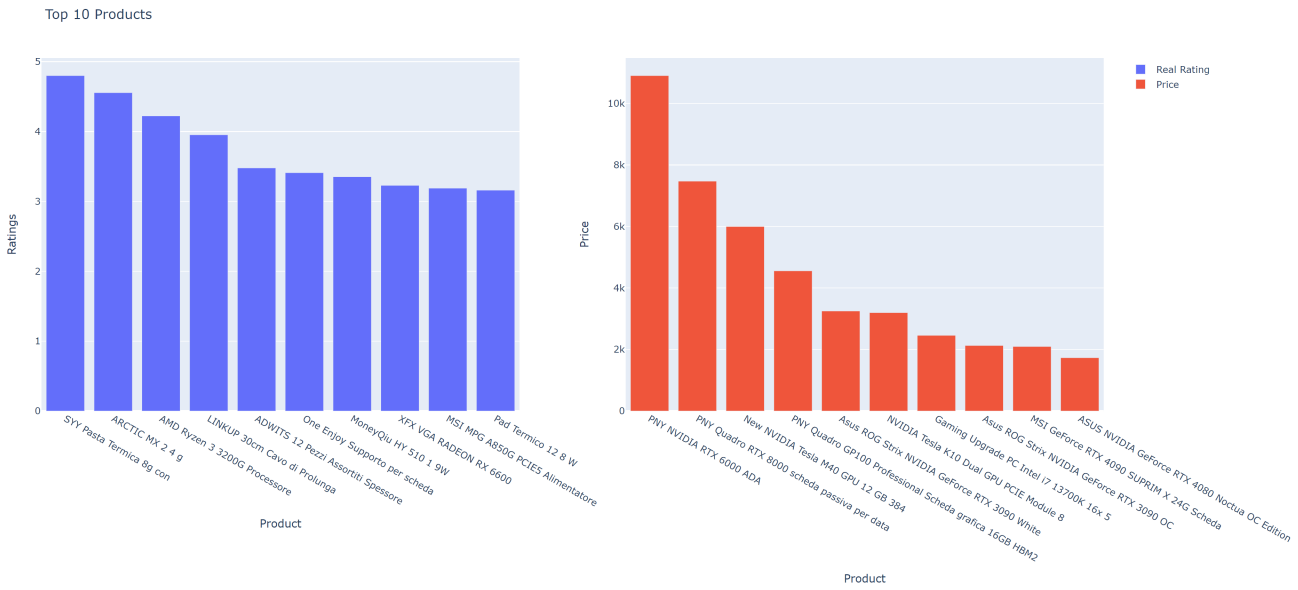
	Product Description	Price	Prime Status	Product URL	Stars	Number of Reviews	Real rating
303	SYY Pasta Termica 8g con Toolkit, Thermal Past...	6.00	Non-Prime	https://www.amazon.it/SYY-aggiornamento-Condut...	4.6	11741.0	4.800000
243	ARCTIC MX-2 (4 g) - Performance Pasta Termica ...	5.99	Non-Prime	https://www.amazon.it/ARCTIC-MX-2-Edizione-201...	4.6	10307.0	4.555728
371	AMD Ryzen 3 3200G, Processore PC, 3,6 GHz (fre...	92.49	Non-Prime	https://www.amazon.it/AMD-Ryzen-3200G-Processo...	4.7	8063.0	4.223478
59	LINKUP - 30cm Cavo di Prolunga PSU Super Morbi...	46.00	Prime	https://www.amazon.it/sspa/click?ie=UTF8&spc=M...	4.5	7062.0	3.952964
172	ADWITS 12 Pezzi Assortiti Spessore 0,5 1,0 1,5...	8.00	Non-Prime	https://www.amazon.it/ADWITS-Silicone-condutti...	4.6	3985.0	3.478818
35	One Enjoy Supporto per scheda grafica GPU, sup...	10.00	Non-Prime	https://www.amazon.it/sspa/click?ie=UTF8&spc=M...	4.5	3876.0	3.410250
305	MoneyQiu HY-510>1.9W/m·K 100g (4 * 25g) Pasta T...	17.00	Non-Prime	https://www.amazon.it/MoneyQiu-HY-510-Termica-...	4.5	3540.0	3.353015
29	XFX VGA RADEON RX 6600 SPEEDSTER SWFT 210 8GB ...	219.00	Non-Prime	https://www.amazon.it/XFX-SPEEDSTER-RADEON-gra...	4.6	2519.0	3.229095
136	MSI MPG A850G PCIE5, Alimentatore 850W, certif...	137.00	Prime	https://www.amazon.it/MSI-Alimentatore-certifi...	4.8	1699.0	3.189413
242	Pad Termico 12,8 W/MK, 85x45x1mm Thermal Pad, ...	12.29	Non-Prime	https://www.amazon.it/One-enjoy-Thermalright-8...	4.4	2703.0	3.160438

First 10 products by rating

I calculated the real ratings by normalizing the number of reviews in a range from 1 to 5 (as the stars) and the made a simple mean by this normalized number of ratings and stars of the product.

The EDA3 returns me:
The number of products with Prime are: 70 with an average Price of: 268.3634285714285 and an average of Stars of: 4.038571428571429
The number of products with Non-Prime are: 339 with an average Price of: 244.77533923303835 and an average of Stars of: 3.8117994100294985
Here we can say that the primes articles are less with an higher prices and higher ratings respect to the non-prime one.

The EDA4 returns me:



First 10 products by rating and price

Here I plot on the left the first 10 products in terms of best rating and on the right the higher prices in the dataset.

2 Exercise

In this exercise I implemented a search engine that first of all compute the inverted index and save it in a file '$inverted_index.pkl$' then I compute the tf-idf matrix and search the documents using the cosine similarity. The below pictures represents the results for the querys: gpu, ASUS, an entire product description and termico.

```
Enter your search query: gpu
One Enjoy Supporto per scheda grafica GPU, supporto per scheda video, supporto GPU (S)
One Enjoy Supporto per scheda grafica GPU, supporto per scheda video, supporto GPU (L)
GWAWG Supporto per scheda grafica GPU Supporto nero GPU Sag staffa supporto per scheda grafica
Uyubao Supporto Scheda Grafica GPU, Supporto Scheda Video, Supporto GPU, M(50mm-80mm)
Uyubao Supporto Scheda Grafica GPU, Supporto Scheda Video, Supporto GPU, M(50mm-80mm)
```

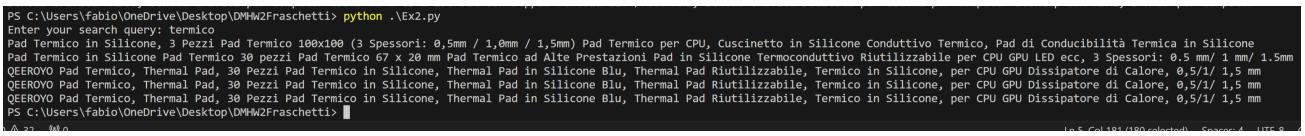
First 5 products on gpu

```
Enter your search query: ASUS
Asus Scheda Video, Nero, One Size
ASUS ROG Strix NVIDIA GeForce RTX 4080 Scheda Grafica Gaming, OpenGL 4.6, 16 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak III, Nero
ASUS DUAL NVIDIA GeForce RTX 3060 Ti OC Edition Scheda Grafica Gaming, OpenGL 4.6, 8 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak II, Nero
ASUS ROG Strix NVIDIA GeForce RTX 4090 Scheda Grafica Gaming, OpenGL 4.6, 24 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak III, Nero
ASUS DUAL NVIDIA GeForce RTX 3060 Ti OC Edition Scheda Grafica Gaming, OpenGL 4.6, 8 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak II, Bianco
PS C:\Users\fabio\OneDrive\Desktop\DMHW2Fraschetti>
```

First 5 products on ASUS

```
PS C:\Users\fabio\OneDrive\Desktop\DMHW2Fraschetti> python .\Ex2.py
Enter your search query: VBESTLIFE Scheda Grafica RX580 8G GDDR5, Doppia Ventola 1244 MHz 14000 MHz 256 Bit GPU Schede Grafiche da Gioco, Supporto 3 DP HD Interfaccia Multimediale Scheda Video Scheda Video
VBESTLIFE Scheda Grafica RX580 8G GDDR5, Doppia Ventola 1244 MHz 14000 MHz 256 Bit GPU Schede Grafiche da Gioco, Supporto 3 DP HD Interfaccia Multimediale Scheda Video Scheda Video
Scheda Grafica GDDR5 da 8 GB, Doppia Ventola 256 Bit 1284/7000 MHz Scheda Grafica da Gioco 8K 16 PCI Express 3.0, Scheda Video 3 X DP HDMI DVI HD
Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB 256 Bit con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Interfaccia DP HDMI DVI D
Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB 256 Bit con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Interfaccia DP HDMI DVI D
Scheda Grafica RX580, Scheda Grafica per Computer GDDR5 da 8 GB a 256 Bit con Doppia Ventola 1284/7000 MHz, Scheda Grafica Dedicata per Giochi per Computer Desktop con HDMI, 3 X DP per Computer
PS C:\Users\fabio\OneDrive\Desktop\DMHW2Fraschetti>
⚠ 32  ⚠ 0                                                    Ln 5, Col 181 (180 selected)  Spaces: 4  UTF-8  CRLF  P
```

First 5 products on VBESTLIFE Scheda Grafica RX580 8G GDDR5, Doppia Ventola 1244 MHz 14000 MHz 256 Bit GPU Schede Grafiche da Gioco, Supporto 3 DP HD Interfaccia Multimediale Scheda Video Scheda Video



```
PS C:\Users\fabio\OneDrive\Desktop\DMHW2Fraschetti> python .\Ex2.py
Enter your search query: termico
Pad Termico in Silicone, 3 Pezzi Pad Termico 100x100 (3 Spessori: 0,5mm / 1,0mm / 1,5mm) Pad Termico per CPU, Cuscinetto in Silicone Conduttivo Termico, Pad di Conducibilità Termica in Silicone
Pad Termico in Silicone Pad Termico 30 pezzi Pad Termico 67 x 20 mm Pad Termico ad Alte Prestazioni Pad in Silicone Termoconduttivo Riutilizzabile per CPU GPU LED ecc, 3 Spessori: 0.5 mm/ 1 mm/ 1.5mm
QEEROYO Pad Termico, Thermal Pad, 30 Pezzi Pad Termico in Silicone, Thermal Pad in Silicone Blu, Thermal Pad Riutilizzabile, Termico in Silicone, per CPU GPU Dissipatore di Calore, 0,5/1/ 1,5 mm
QEEROYO Pad Termico, Thermal Pad, 30 Pezzi Pad Termico in Silicone, Thermal Pad in Silicone Blu, Thermal Pad Riutilizzabile, Termico in Silicone, per CPU GPU Dissipatore di Calore, 0,5/1/ 1,5 mm
QEEROYO Pad Termico, Thermal Pad, 30 Pezzi Pad Termico in Silicone, Thermal Pad in Silicone Blu, Thermal Pad Riutilizzabile, Termico in Silicone, per CPU GPU Dissipatore di Calore, 0,5/1/ 1,5 mm
PS C:\Users\fabio\OneDrive\Desktop\DMHW2Fraschetti>
```

First 5 products on termico

3 Exercise

3.1

In the first part I implement character shingles. First I divided the descriptions in shingles of parameter k that you can choose. Then I hash this shingles with sha1 hash and save it into a file called *DataFrame.tsv*

3.2

In the second part I implemented a minwise hashing, taking 2 sets I compute the minhash matrix by taking the lower hash compared on all the elements. Once I have this two sets I compare the 2 sets with Jaccard similarity.

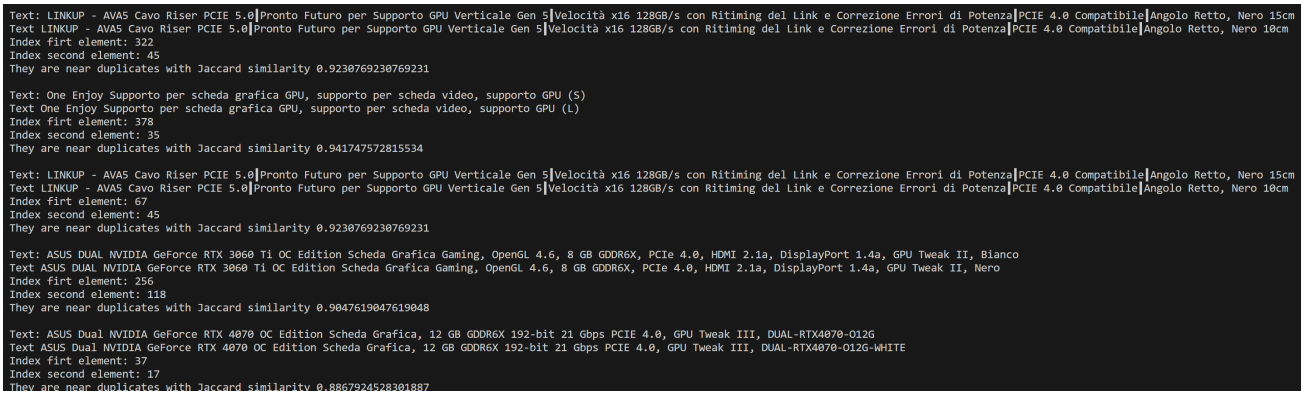
3.3

In the third part I implemtented an LSH with bands 20 and rows are 5. First of all i computed the shinglings and minwise hashing and then I divide it into bands and raws. Then I also compute the Brute force way and it's much slower.
This are the times and number of duplicates for LSH and brute force:

Results with no LSH:
Number of near duplicates: 286
Time taken with no LSH: 6.3647 seconds

Results with LSH :
Number of near duplicates: 107
Time taken LSH: 0.1906 seconds

Some examples of duplicates are:



```
Text: LINKUP - AVA5 Cavo Riser PCIE 5.0|Pronto Futuro per Supporto GPU Verticale Gen 5|Velocità x16 128GB/s con Ritiming del Link e Correzione Errori di Potenza|PCIE 4.0 Compatibile|Angolo Retto, Nero 15cm
Text LINKUP - AVA5 Cavo Riser PCIE 5.0|Pronto Futuro per Supporto GPU Verticale Gen 5|Velocità x16 128GB/s con Ritiming del Link e Correzione Errori di Potenza|PCIE 4.0 Compatibile|Angolo Retto, Nero 10cm
Index firt element: 322
Index second element: 45
They are near duplicates with Jaccard similarity 0.9230769230769231

Text: One Enjoy Supporto per scheda grafica GPU, supporto per scheda video, supporto GPU (S)
Text One Enjoy Supporto per scheda grafica GPU, supporto per scheda video, supporto GPU (L)
Index firt element: 378
Index second element: 35
They are near duplicates with Jaccard similarity 0.941747572815534

Text: LINKUP - AVA5 Cavo Riser PCIE 5.0|Pronto Futuro per Supporto GPU Verticale Gen 5|Velocità x16 128GB/s con Ritiming del Link e Correzione Errori di Potenza|PCIE 4.0 Compatibile|Angolo Retto, Nero 15cm
Text LINKUP - AVA5 Cavo Riser PCIE 5.0|Pronto Futuro per Supporto GPU Verticale Gen 5|Velocità x16 128GB/s con Ritiming del Link e Correzione Errori di Potenza|PCIE 4.0 Compatibile|Angolo Retto, Nero 10cm
Index firt element: 67
Index second element: 45
They are near duplicates with Jaccard similarity 0.9230769230769231

Text: ASUS DUAL NVIDIA GeForce RTX 3060 Ti OC Edition Scheda Grafica Gaming, OpenGL 4.6, 8 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak II, Bianco
Text ASUS DUAL NVIDIA GeForce RTX 3060 Ti OC Edition Scheda Grafica Gaming, OpenGL 4.6, 8 GB GDDR6X, PCIe 4.0, HDMI 2.1a, DisplayPort 1.4a, GPU Tweak II, Nero
Index firt element: 256
Index second element: 118
They are near duplicates with Jaccard similarity 0.9047619047619048

Text: ASUS Dual NVIDIA GeForce RTX 4070 OC Edition Scheda Grafica, 12 GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU Tweak III, DUAL-RTX4070-O12G
Text ASUS Dual NVIDIA GeForce RTX 4070 OC Edition Scheda Grafica, 12 GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU Tweak III, DUAL-RTX4070-O12G-WHITE
Index firt element: 37
Index second element: 17
They are near duplicates with Jaccard similarity 0.8867924528301887
```

4 Exercise

In this exercise I developed the same problem as before but in spark in the first part i cleaned a bit the documents, then I applied the shingling minwise hashing and LSH as before. The results are:
False positive rate: 0.07599064294321968
False negative rate: 0.0

```
(1, 276)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 276)]
(1, 74)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('EZDIY-FAB GPU Holder Brace Supporto Della Scheda Grafica GPU Supporto Della Scheda Video con 5V 3 pin ARGB LED,Video Card Sag Holder/Holster Bracket Support RX6700,RTX3090-309EZ-Nero', 74)]
(1, 126)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('EZDIY-FAB Scheda Grafica GPU Brace 5V 3Pin ARGB,Supporto Della Scheda Video Sag Holder Holster Bracket,Alluminio Anodizzato (Nero)', 126)]
(74, 126)
[('EZDIY-FAB GPU Holder Brace Supporto Della Scheda Grafica GPU Supporto Della Scheda Video con 5V 3 pin ARGB LED,Video Card Sag Holder/Holster Bracket Support RX6700,RTX3090-309EZ-Nero', 74)]
[('EZDIY-FAB Scheda Grafica GPU Brace 5V 3Pin ARGB,Supporto Della Scheda Video Sag Holder Holster Bracket,Alluminio Anodizzato (Nero)', 126)]
(74, 276)
[('EZDIY-FAB GPU Holder Brace Supporto Della Scheda Grafica GPU Supporto Della Scheda Video con 5V 3 pin ARGB LED,Video Card Sag Holder/Holster Bracket Support RX6700,RTX3090-309EZ-Nero', 74)]
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 276)]
(126, 276)
[('EZDIY-FAB Scheda Grafica GPU Brace 5V 3Pin ARGB,Supporto Della Scheda Video Sag Holder Holster Bracket,Alluminio Anodizzato (Nero)', 126)]
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 276)]
(1, 33)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('RX580 8GB GDDR5 Scheda Grafica da Gioco a 256 Bit, Risoluzione 4096x2160 8 Pin Computer PC Gaming Video Scheda Grafica GPU con Doppia Ventola di RaffreddamentoPCI 3.0', 33)]
(1, 60)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('Kit di Raffreddamento ad Acqua per Computer PC, Set di Raffreddamento a Liquido per CPU all-in-One Fai-da-Te: dissipatore di Calore da 240 mm CPU/GPU Blocco Serbatoio Pompa LED Ventola', 60)]
(1, 94)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('RX580 8GB GDDR5 Scheda Grafica da Gioco a 256 Bit, Risoluzione 4096x2160 8 Pin Computer PC Gaming Video Scheda Grafica GPU con Doppia Ventola di RaffreddamentoPCI 3.0', 94)]
(33, 60)
[('RX580 8GB GDDR5 Scheda Grafica da Gioco a 256 Bit, Risoluzione 4096x2160 8 Pin Computer PC Gaming Video Scheda Grafica GPU con Doppia Ventola di RaffreddamentoPCI 3.0', 33)]
[('Kit di Raffreddamento ad Acqua per Computer PC, Set di Raffreddamento a Liquido per CPU all-in-One Fai-da-Te: dissipatore di Calore da 240 mm CPU/GPU Blocco Serbatoio Pompa LED Ventola', 60)]
Time spent for LSH: 0.04412674903869629
```

LSH

```
Time spent for brute force comparisons: 0.022305011749267578
(1, 276)
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 1)]
[('Schede Grafiche AMD per Radeon HD7670, 4GB GDDR5 Computer PC Gaming Video GPU Scheda Grafica, 128-Bit, Supporto DirectX 11 PCI Express X16 2.1 DVI, HDMI, VGA', 276)]
(3, 61)
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 3)]
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 61)]
(3, 310)
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 3)]
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 310)]
(3, 378)
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 3)]
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Scheda Grafica da Gioco per CAD 3D, CAM, Editing Video e Immagini', 378)]
(18, 37)
[('ASUS Dual NVIDIA GeForce RTX 4070 OC Edition Scheda Grafica, 12 GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU Tweak III, DUAL-RTX4070-O12G-WHITE', 18)]
[('ASUS Dual NVIDIA GeForce RTX 4070 OC Edition Scheda Grafica, 12 GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU Tweak III, DUAL-RTX4070-O12G', 37)]
(21, 75)
[('MUTOUHE Supporto per scheda grafica, supporto GPU in alluminio per schede grafiche regolabili per una vasta gamma di computer e schede grafiche (nero)', 21)]
[('MUTOUHE Supporto per scheda grafica, supporto GPU in alluminio per schede grafiche regolabili per una vasta gamma di computer e schede grafiche (nero)', 75)]
(24, 72)
[('adspow Supporto Scheda Grafica GPU, Supporto Scheda Video, Supporto in Alluminio per Scheda Video Nero Supporto GPU Altezza Regolabile, Mini (27-50 mm)', 24)]
[('adspow Supporto Scheda Grafica GPU, Supporto Scheda Video, Supporto in Alluminio per Scheda Video Nero Supporto GPU Altezza Regolabile, Mini (27-50 mm)', 72)]
(25, 247)
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB 256 Bit con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Interfaccia DP HDMI DVI D', 25)]
[('Scheda Grafica RX 580, Scheda Grafica GDDR5 da 8 GB 256 Bit con GPU 1284 MHz, 60 Hz 4K, PCI Express 3.0, 2 Ventole di Raffreddamento, Interfaccia DP HDMI DVI D', 247)]
(32, 93)
[('Gintai Cavo di alimentazione 10 Pin a 8 + 8 Pin per HP DL380 G9 e GPU 50 cm', 32)]
[('Gintai Cavo di alimentazione 10 Pin a 8 + 8 Pin per HP DL380 G9 e GPU 50 cm', 93)]
(33, 94)
[('RX580 8GB GDDR5 Scheda Grafica da Gioco a 256 Bit, Risoluzione 4096x2160 8 Pin Computer PC Gaming Video Scheda Grafica GPU con Doppia Ventola di RaffreddamentoPCI 3.0', 33)]
[('RX580 8GB GDDR5 Scheda Grafica da Gioco a 256 Bit, Risoluzione 4096x2160 8 Pin Computer PC Gaming Video Scheda Grafica GPU con Doppia Ventola di RaffreddamentoPCI 3.0', 94)]
```

Brute force