Data Mining Homework 2

Alessandro Monteleone 1883922

Academic Year: 2023/2024

Contents

1 Introduction

I have decided to realize this homework as a flask application in order to have a better view on the results.
Now I will explain one by one all the solutions to the problems, at the end of the document there will be images taken from the application showing the results and there will also be a little guide to use the application.
All the python files are placed in the folder named "server"

2 Problem 1

The functions involved in this peroblem are in amazon_product_retrieval.py, amazon_product_analysis.py and in utils_and_classes.py.
I have decided to use the library lxml to do scraping because it works well also with bigger data, this is not the case but it might be useful in future.
My code does the following for the part of retrieving data:

- it iterates on the number of pages to retrieve (I have chosen 10) and each time sends a request for a page
  - if the request is refused it tries again to retrieve the page
- if the response is ok then it finds all the products in the page looking for a certain class in the elements of the html
- for each item it builds an object Product which upon creation search inside the html of the element looking for the classes of the interesting informations (description,url,number of reviews, number of stars,primeness) and writes the informations in a tsv file (amazon_products_gpu.tsv)

All the classes used for doing scraping are saved as constants inside utils_and_classes.py since amazon could decide to change them and this would make this code not working anymore, instead in this way it is sufficient to change the constants to make it working.
For the analysis part let's say something about every requested feature:

- for the price ranges i have plotted the maximum and the minimum price for all the products which include in the description the word describing the category. For the categories I have chosen the series of the gpus like 3080,3090 etc.
- To evaluate the products throught their ratings I have chosen to compute a score based on the ratings times the logarithm of the number of reviews, in this way products with lots of reviews are not advantaged too much
- I have analyzed the primeness looking at the mean price and the mean ratings of the product with it and the ones without it and I have noticed that there is not a big difference between them
- for the top 10 products in term of price or ratings I have simply plotted the table ordered by the required term

All the result of the analysis are visible in the menu "DEFAULT QUERIES"

3 Problem 2

For this problem the code is inside index_building.py, amazon_product_analysis.py and in utils_and_classes.py.
For building the index in my code I iterate on the descriptions, for each description I use a class for preprocess the description, this class removes the punctuation and the stopwords and tokenize the description. After preprocessing the list of token is iterated to count the number of times that each word appear in the list, when the iteration finishes the dictionary word:num of appearences is stored in a list since it is useful to compute the TFIDF later. Then for each word in the dictionary built in the iteration it appends the tuple (product id, num of appearences of the word) to the list of the same word in the inverted index. It also computes the normas of each description, to do so it iterates on the list of word:num of appearences computed in previous step and for each description compute the TFIDF of the words (using also tthe inverted index), sums them squared and computes the squared root of the result and put the final result in a list. In this way I can get the norma of a description simply indicating the id of the product.
Now it stores the inverted index and the normas in a json in order to not to have to recompute it every time.
When it is required to compute cosine similarity the index and the normas are retrieved and the similarity is computed as follows:

- the query is preprocessed as the descriptions were
- for each word in the query using the inverted index it is retrieved each description that contains it and it is placed inside a dictionary as id description: cos similarity, each time that the same description appears in the list of any of the words of the query its cosine similarity is updated
- when the iteration on the query has finished all the descriptions with their cosine similarities are put inside an heap
- the number of required descriptions is then retrieved from the heap

The cosine similarity function is available in the menu "QUERY MENU"

4 Problem 3

The code for this problem is inside hashing_techniques.py and utils_for_hashing.py.
I have realized this problem using 5 classes, I will explain each one.

4.1 Shingling

this class is used to compute the shingles of a string, it takes in input an hash function and the length of the shingles (as k) to be initializad, then it exposes a function to compute the shingles. In this function it iterates on the string that takes in input and for each slice of length k it applies the hash function to it and put the result in a list. At the end of the iterations it returns the list.

4.2 MinwiseHashing

This class is built giving in input to it an integer t and a list of hash functions. The class exposes a function that given a list of shingles returns a list of hashed values of length t, in fact for t times it apply one of the hash functions (possibly always different) to every shingle of the list taking the minimum at each iteration and putting it in another list. After t iterations it returns the list of hashed values.

4.3 LSH

This class is built giving in input to it 2 integers r and b and a list of hash functions. This class exposes a function that given a list of elements represented as list of objects of length b*r (if the length is different then it raises an error) returns a set of pairs which are the nearest according to LSH. The function iterates on fragments of the lists, these fragments has length r and therefore there are b fragments for each element, when it is considering the i-th fragment it iterates on the elements, it takes the i-th fragment, hashes it and puts it in a dictionary of kind hash:list of elements with same hash, if there are elements inside the list for the hash of the actual element for each pair actual element, element in list it builds a tuple in which the lower element is always the first in the pair and then add the tuple to the resulting set. In this way there are no duplicates in the result and it is returned when the iterations finishes.

4.4 CompareWithLSH

This class takes in input an instance of each of the previous classes to be initialized. This class simply put together all the 3 above and keep trace of the execution time.

4.5 CompareWithJaccard

This class takes in input an instance of the class Shingles and a value for the threshold to be initialized. The class exposes a method for computing the nearest pairs of elements according to the Jaccard similarity and the given threshold. this method is preatty easy, as first it computes the shingles then it start iterating on all the possible pairs of elements, for each pair it compute the jaccard similarity and if it is higher than the threshold then the pair is put in the result set. At the end the function returns the result set.
The function also keep trace of the execution time.

4.6 Report

This is another class but it is made only to execute the comparing classes and to build a report with the required informations. The fundamental things to say for this class are the coiches for the parameters and the results of the experiment.

- for the hash functions it was used the family showed in the questions file, each function was generated with a different integer
- the values for b and r that I have chosen are respectively 7 and 10, that's because I have tried lots of values in Geogebra and it seems to me that these are the best for the given threshold, the resulting approximation function is showed in figure [9]

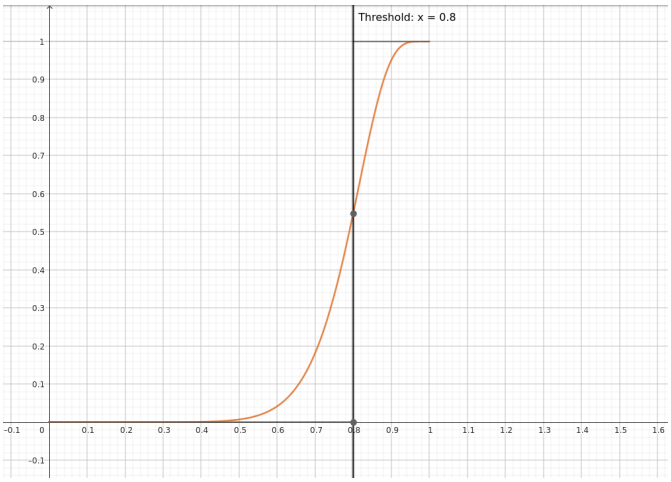


Figure 1: LSH approximation

For which concerns the results I have found that LSH finds 2 false positives and 0 false negative which is pretty good but it it surely dued by the kind of data, in first instance the descriptions of the products are short, they can't be compared with real documents, then a lot of products are duplicates, this is dued to the fact that often amazon puts some products at the beginning of the page since they are racomanded but they are not removed from their original place. Also the execution time is wasted by the bad data, in fact the LSH should be faster since it avoid to compare a lot of pairs of documents but in this case it is the opposite, this is probably caused also by the usage of cryptographic hash functions that are slower than normal hash functions. By the way the bare sesults are the following:

- Number of duplicates for LSH: 442
- Number of duplicates for Jaccard similarity: 440
- Size of intersection between Jaccard and LSH results: 440
- Time required for the execution of LSH algorithm: 5574
- Time required for the execution of Jaccard similarity algorithm: 1047

All the results of the report are also available in the menu "NEAREST PRODUCTS REPORTS" and the results of the nearest descriptions both for LSH and Jaccard similarity are available in the same menu

5 Problem 4

All the code for this problem is inside hashing_techniques_with_spark.py, hashing_techniques.py and utils_for_hashing.py. i have realized 2 classes to solve the problem, I will explain them separately

5.1 CompareWithLSHSpark

This class is built for comparing the descriptions using LSH and it is implemented with Spark. To be instantiated this class needs 3 integers: the length of the shingles (k), the number of bands (b) and the number of rows (r), then it needs also the hash functions that have to be applied in the steps of the algorithm. The function that this class exposes takes in input a RDD object containing all the descriptions and apply the algorithm(As in previous excercize the function also keep trace of the execution time). Let's explain that through its steps.

5.1.1 Shingling

To produce the shingles as first it applies a flat map that returns set of tuples (id,hash(shingle)), each tuple is produced dividing each element of the RDD into its shingles of length k

5.1.2 Minwise-hashing

To produce the signature of the descriptions in the RDD minwise hashing is executed in 2 steps:

- in step 1 it is applied again a flat map to the shingles and it is returned set of tuples ((id, num hash), hashed element), each tuple is produced appling to each element the t hashes (if you apply i-th hash then num hash = i)
- in step 2 it is applied a reduce by key function , since the key is (id, num hash) then all the elements of the same document to which it was applied the same hash are reduced and the reduce function is minimum. In this way the result is elements of kind ((id, num hash), hashed element) but now there is only 1 element or each pair (id, num hash) and the "hashed element" is the minimum

5.1.3 LSH

To apply LSH more steps are needed:

- in step 1 each element is mapped to its band, the resulting is a tuple of kind (id, num band), array with the element placed in its position inside the signature of the document (id). For each element it is computed the band and the resulting position inside an array of length r in which all the elements are null except for the hashed element that is placed in the computed position
- In step 2 a reduce by key function is called and the function applied is a merging of arrays so that every element of the same document that belongs to the same band is merged in one single array
- In step 3 every element is mapped to a tuple ((num band, hash(element)), id) where hash(element) is the hash function applied to the array
- In step 4 it is used a reduce by key function which has to reduce the elements with same (num band, hash(element)) to a single set of ordered pairs , in this way now each element of our RDD is like a bucket containing all the pairs of elements which had same (num band, hash(element))
- In step 5 we simply filter all the elements which haven't a set as value, since there could be pairs (num band, hash(element)) that are not repeated in the RDD
- In step 6 we compute the result applying a reduce function which merge all the sets together

5.2 CompareWithJaccardSpark

This class is built for comparing descriptions using Jaccard similarity and it is implemented with Spark.
To be instantiated this class needs an integer k which represents the length of the shingles, a float which is the threshold that tells us which descriptions are considered near duplicates and which not and a hash function to be applied during shingling.
As the other class also this exposes a function to compute the nearest descriptions using Jaccard similarity and it also keep trace of the execution time. The first part of the algorithm is shared with the previous class (Shingling part 5.1.1) the real algorithm is executed after and here it is explained in few steps:

- In step 1 the shingles of a same document are put together using a reduce by key function, the result is for each document a set of shingles

- In step 2 it is applied a flat map and for each element it is created a set of elements with the same value (the set) but different id, the id is a ordered pair that represent the comparing to be executed in the next step
- In step 3 it is applied a reduce by key function that for each pair of elements with the same key compute the Jaccard similarity and return True if it is higher than the threshold, false otherwise. So now the elements are ((id1,id2) are near?)
- In step 4 it is simply applied a filter on the tuples that are near duplicates (are near? = true)
- In step 5 the result is computed through a reduce function that puts all the pairs in a set

5.3 Report

The class report is also used in this case to execute the 2 classes and build the report. From the result we can notice that the resulting sets are the same as the ones computed by the code which doesn't use Spark so same number of elements in each result ans same intersections between them. What change is the execution time which is anyways higher than the one of the algorithm which doesn't use Spark. It is probably dued to the fact that the data are not so suitable to be treated in a distributed way or to the fact that my machine has only 4 cores so the parallelization is not so good. Anyway we can notice that beside the Jaccard similarity algorithm execution time has been more than duplicated the execution time of LSH algorithm is increased only of nearly 50%, then we could say that the LSH algoritm has shown to be more suitable for parallel execution. These are the results of the execution times:

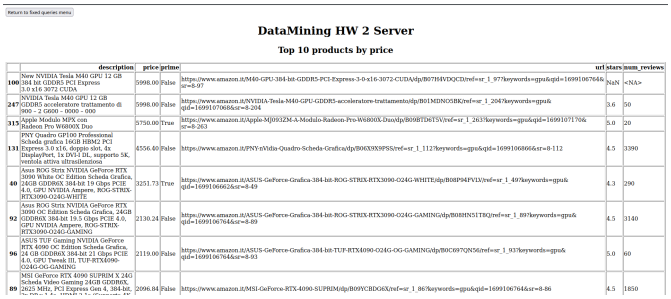
- Time required for the execution of LSH algorithm: 7873
- Time required for the execution of Jaccard similarity algorithm: 2888

All the results of the report are also available in the menu "NEAREST PRODUCTS REPORTS" and the results of the nearest descriptions both for LSH and Jaccard similarity are available in the same menu

6 Guide to use the application

The application is made by 2 fundamental parts the script for installing th erequired python libreries install-requirements.sh and the application: app.py . This one is inside the folder "server" and has to be executed using the command python3 app.py (easy) now it should launch a server available on local host at url http://127.0.0.1:5000. If you click on the link (after launching the app) you should find all the result and the functions visualized in a better way (nothing spectacular but way better than a txt file), you can navigate through menus and try all the functions.

7 Images



Return to fixed queries menu

DataMining HW 2 Server

Top 10 products by price

	description	price	prime	url	stars	num_reviews
100	Nvw NVIDIA Tesla M40 GPU 12 GB 384 bit GDDR5 PCI Express 3.0 x16 3072 CUDA	5998.00	False	https://www.amazon.it/M40-GPU-384-bit-GDDR5-PCI-Express-3-0-x16-3072-CUDA/dp/B07H4VDQCD/ref=sr_1_97?keywords=gpu&qid=1699106764&sr=8-97	NaN	<NA>
247	NVIDIA Tesla M40 GPU 12 GB GDDR5 acceleratore trattamento di 900 - 2 G600 - 0000 - 000	5998.00	False	https://www.amazon.it/NVIDIA-Tesla-M40-GPU-GDDR5-acceleratore-trattamento/dp/B01MDN05BK/ref=sr_1_204?keywords=gpu&qid=1699107668&sr=8-204	3.6	50
317	Apple Modulo MPX con Radeon Pro W6800X Duo	5750.00	True	https://www.amazon.it/Apple-MJ093ZM-A-Modulo-Radeon-Pro-W6800X-Duo/dp/B098BTD6TS/ref=sr_1_263?keywords=gpu&qid=1699107170&sr=8-263	5.0	20
131	PNY Quadro GP100 Professional Scheda grafica 16GB HBM2 PCI Express 3.0 x16, doppio slot, 4x DisplayPort, 1x DVI-I DL, supporto 5K, ventola attiva ultrasilenziosa	4556.40	False	https://www.amazon.it/PNY-nVidia-Quadro-Scheda-Grafica/dp/B06X9X9PSS/ref=sr_1_112?keywords=gpu&qid=1699106866&sr=8-112	4.5	3390
40	Asus ROG Strix NVIDIA GeForce RTX 3090 White OC Edition Scheda Grafica, 24GB GDDR6X 384-bit 19 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3090-O24G-WHITE	3251.73	True	https://www.amazon.it/ASUS-GeForce-Grafica-384-bit-ROG-STRIX-RTX3090-O24G-WHITE/dp/B08P94FV1J/ref=sr_1_40?keywords=gpu&qid=1699106662&sr=8-40	4.3	290
92	Asus ROG Strix NVIDIA GeForce RTX 3090 OC Edition Scheda Grafica, 24GB GDDR6X 384-bit 19.5 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3090-O24G-GAMING	2130.24	False	https://www.amazon.it/ASUS-GeForce-Grafica-384-bit-ROG-STRIX-RTX3090-O24G-GAMING/dp/B08HNS1TBQ/ref=sr_1_89?keywords=gpu&qid=1699106764&sr=8-89	4.5	3140
96	ASUS TUF Gaming NVIDIA GeForce RTX 4090 OC Edition 24GB Grafica 24 GB GDDR6X 384-bit 21 Gbps PCIE 4.0, GPU Tweak III, TUF-RTX4090-O24G-OG-GAMING	2119.00	False	https://www.amazon.it/ASUS-GeForce-Grafica-384-bit-TUF-RTX4090-O24G-OG-GAMING/dp/B0C697QN56/ref=sr_1_93?keywords=gpu&qid=1699106764&sr=8-93	5.0	60
89	MSI GeForce RTX 4090 SUPRIM X 24G Scheda Video Gaming 24GB GDDR6X, 2625 MHz, PCI Express Gen 4, 384 bit,	2096.84	False	https://www.amazon.it/MSI-GeForce-RTX-4090-SUPRIM/dp/B09YCBDG6X/ref=sr_1_86?keywords=gpu&qid=1699106764&sr=8-86	4.5	1850

Figure 2: Top 10 products by price

Return to fixed queries menu

DataMining HW 2 Server

Top 10 products by ratings

	description	price	prime	url	stars	num_reviews
118	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	124.29	False	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo4NDE5ODU5MTY1MDc2NzI6MTY5OTEwNjc2NDpzcF9idGY6MzAwMDA1Mzc1NTAxMjMyOjowOjo&url=%2FGrafica-Risoluzione-4096x2160-Computer-RaffreddamentoPCI%2Fdp%2FB0BV6KYKKN%2Fref%3Dsr_1_107_sspa%3Fkeywords%3Dgpu%26qid%3D1699106764%26sr%3D8-107-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9idGY%26psc%3D1	5.0	10
406	EKWB Kit di disconnessione rapida GPU EK-Pro	135.30	False	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo1MjI3NjI2MzIyODExNjcyOjE2OTkxMDc0NzU6c3BfbXRmOjIwMTI0MzQ5MzIyMDk4OjowOjo&url=%2FEKWB-EK-Pro-Kit-disconnessione-rapida%2Fdp%2FB0855BBLPM%2Fref%3Dsr_1_389_sspa%3Fkeywords%3Dgpu%26qid%3D1699107475%26sr%3D8-389-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9tdGY%26psc%3D1	5.0	30
54	EKWB Kit di disconnessione rapida GPU EK-Pro	135.30	False	https://www.amazon.it/sspa/click?ie=UTF8&spc=MToyMDQ0NDk1Mzk5MzgxNjcyOjE2OTkxMDY2NjI6c3BfbXRmOjIwMTI0MzQ5MzIyMDk4OjowOjo&url=%2FEKWB-EK-Pro-Kit-disconnessione-rapida%2Fdp%2FB0855BBLPM%2Fref%3Dsr_1_35_sspa%3Fkeywords%3Dgpu%26qid%3D1699106662%26sr%3D8-35-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9tdGY%26psc%3D1	5.0	30
382	EKWB Kit di disconnessione rapida GPU EK-Pro	135.30	False	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo2NTg2MzI3NzEzODg5NjcyOjE2OTkxMDcyNzI6c3BfbXRmOjIwMTI0MzQ5MzIyMDk4OjowOjo&url=%2FEKWB-EK-Pro-Kit-disconnessione-rapida%2Fdp%2FB0855BBLPM%2Fref%3Dsr_1_305_sspa%3Fkeywords%3Dgpu%26qid%3D1699107272%26sr%3D8-305-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9tdGY%26psc%3D1	5.0	30
376	Gintai - Cavo di alimentazione da 8 pin a 8 pin per Dell R730 e Nvidia K80 / M40 / M60 / P40 / P100 PCIE GPU	20.99	False	https://www.amazon.it/Gintai-Cavo-alimentazione-Nvidia-PCIE-GPU/dp/B08749VZTL/ref=sr_1_312?keywords=gpu&qid=1699107272&sr=8-312	5.0	10
375	YOYIAG Supporto Scheda Grafica GPU Supporto: Supporto Scheda Grafica in Lega Di Alluminio Regolabile in Altezza con Magnete Supporto per Scheda Grafica Supporto Scheda Video Durevole Estendi Il Nero	7.99	False	https://www.amazon.it/YOYIAG-Supporto-Scheda-Grafica-GPU/dp/B0BTLXK7Q6/ref=sr_1_311?keywords=gpu&qid=1699107272&sr=8-311	5.0	50
70	ASUS TUF Gaming NVIDIA GeForce RTX 4060 Ti OC Edition Scheda Grafica, 8 GB GDDR6 [illegible]	512.15	False	https://www.amazon.it/VGA-Asus-RTX4060TI-8GB-TUF/dp/B0BRYY1JX8/ref=sr_1_63?keywords=gpu&qid=1699106764&sr=8-63	5.0	120

Figure 3: Top 10 products by rating

Return to fixed queries menu

DataMining HW 2 Server

Top 10 products by weighted ratings

	description	price	prime	url	stars	num_reviews	weighted_ratings
308	Corsair TM30 Pasta Termica ad Alte Prestazioni, Collante Dissipatore di Calore, Pasta Termoconduttiva High per CPU e GPU, Materiale Interfaccia Termica, 3 grammi	14.40	False	https://www.amazon.it/Corsair-Prestazioni-Dissipatore-Termoconduttiva-Interfaccia/dp/B07KQ1T158/ref=sr_1_253?keywords=gpu&qid=1699107170&sr=8-253	4.6	20183	65.783924
261	Corsair TM30 Pasta Termica ad Alte Prestazioni, Collante Dissipatore di Calore, Pasta Termoconduttiva High per CPU e GPU, Materiale Interfaccia Termica, 3 grammi	14.40	False	https://www.amazon.it/Corsair-Prestazioni-Dissipatore-Termoconduttiva-Interfaccia/dp/B07KQ1T158/ref=sr_1_218?keywords=gpu&qid=1699107068&sr=8-218	4.6	20183	65.783924
340	AMD Ryzen 3 3200G, Processore PC, 3,6 GHz (frequenza massima: 4,0 GHz), 4 MB L3 Cache, Socket AM4, GPU integrato Radeon Vega 8	91.99	False	https://www.amazon.it/AMD-Ryzen-3200G-Processore-frequenza/dp/B07STGHZK8/ref=sr_1_292?keywords=gpu&qid=1699107170&sr=8-292	4.7	9211	61.894968
366	Corsair Cavi Protetti Singolarmente con Guaina PCIe (Connettore Singolo) Type 4 Gen 4 Premium - Nero	22.90	False	https://www.amazon.it/Corsair-Protetti-Singolarmente-Connettore-Singolo/dp/B07L68FVP8/ref=sr_1_299?keywords=gpu&qid=1699107272&sr=8-299	4.7	5740	58.688125
99	Evga Geforce Rtx 2060 Sc Gaming, 6Gb Gddr6, Ventilatore Hdb , 06G-P4-2062-Kr, Nero, 18.98 x 4.45 x 11.12 cm, 880 grammi	NaN	False	https://www.amazon.it/EVGA-GeForce-GAMING-ventilatore-06G-P4-2062-KR/dp/B07MVNPKWN/ref=sr_1_96?keywords=gpu&qid=1699106764&sr=8-96	4.6	6720	58.485529
21	EZDIY-FAB GPU VGA PCIe 8 pin U Turn 180 Gradi Angolo Connettore Scheda Adattatore di Alimentazione per Scheda Grafica Desktop Tipo Inverso 3Pack	11.99	False	https://www.amazon.it/EZDIY-FAB-Connettore-Adattatore-Alimentazione-Grafica/dp/B096VB8MB5/ref=sr_1_26?keywords=gpu&qid=1699106662&sr=8-26	4.6	6600	58.365951
24	XFX SPEEDSTER QICK319 RADEON RX 6750XT CORE Gaming Scheda Grafica con 12GB GDDR6 HDMI 3xDP, AMD RDNA™ 2 (RX-675XYJFDP)	379.00	False	https://www.amazon.it/XFX-SPEEDSTER-QICK319-Grafica-RX-675XYJFDP/dp/B0B34M1YLW/ref=sr_1_29?keywords=gpu&qid=1699106662&sr=8-29	4.6	6540	58.305345
253	EZDIY-FAB Personalizzato PSU Estensione Cavo di Prolunga Manicotto MOD GPU PC Alimentatore per PC in Nylon Morbido Intrecciato con Kit Pettine 24PIN/8PIN a 6+2Pin/8PIN a 4+4PIN-300MM Bianco	25.99	False	https://www.amazon.it/EZDIY-FAB-IT2-WS-18AWG-300/dp/B082XNT8N1/ref=sr_1_210?keywords=gpu&qid=1699107068&sr=8-210	4.7	5300	58.147350
103	EZDIY-FAB Personalizzato PSU Estensione Cavo di Prolunga Manicotto MOD GPU PC Alimentatore per PC in Nylon Morbido Intrecciato con Kit Pettine 24PIN/8PIN a 6+2Pin/8PIN a 4+4PIN-300MM Bianco	25.99	False	https://www.amazon.it/EZDIY-FAB-IT2-WS-18AWG-300/dp/B082XNT8N1/ref=sr_1_100?keywords=gpu&qid=1699106764&sr=8-100	4.7	5300	58.147350
162	EZDIY-FAB Personalizzato PSU Estensione Cavo di Prolunga Manicotto MOD GPU PC Alimentatore per PC in Nylon Morbido Intrecciato con Kit Pettine 24PIN/8PIN a 6+2Pin/8PIN a 4+4PIN-300MM Bianco	25.99	False	https://www.amazon.it/EZDIY-FAB-IT2-WS-18AWG-300/dp/B082XNT8N1/ref=sr_1_150?keywords=gpu&qid=1699106866&sr=8-150	4.7	5300	58.147350

Figure 4: Top 10 products by rating weighted by the number of reviews

Return to fixed queries menu

DataMining HW 2 Server

Analyze primeness

	average price	average price prime products	average stars	average stars prime products
0	247.9	232.8	4.38	4.2

Figure 5: Analysis of primeness with respect to price and ratings

Return to fixed queries menu

DataMining HW 2 Server

Price range for different series of gpu

	categories	min price	max price
0	4090	17.99	2119.00
1	4080	19.99	1739.99
2	7900	1204.38	1204.38
3	4070	22.86	1079.23
4	3090	8.90	3251.73
5	3080	167.63	1015.68
6	6800	5750.00	5750.00
7	3070	502.00	931.90
8	2080	31.98	31.98
9	6700	377.90	407.89
10	4060	346.00	563.00
11	3060	318.00	377.89
12	6650	285.54	394.57
13	7600	316.58	361.47
14	6600	217.00	253.00
15	1080	15.95	147.30
16	2060	307.00	369.88
17	3050	258.62	317.20
18	1660	233.85	233.85
19	590	19.99	205.99
20	6500	179.65	179.65
21	2060	307.00	369.88
22	580	109.90	148.75
23	1650	179.90	189.96
24	480	109.76	109.76
25	780	800.42	800.42

Figure 6: Price range for different series of gpus

Return to main menu

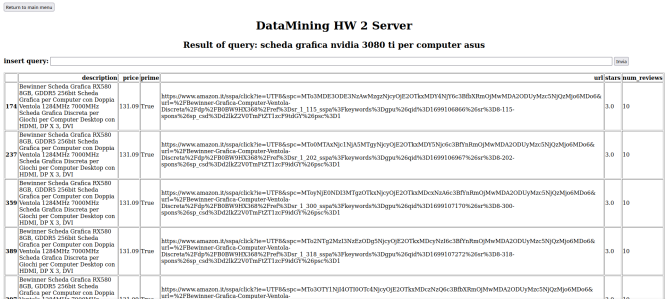
DataMining HW 2 Server

Result of query: 3080 nvidia

insert query: Invia

	description	price	prime	url	stars	num_reviews
227	Asus ROG Strix NVIDIA GeForce RTX 3080 OC Edition Scheda Grafica, 12GB GDDR6X 384-bit 19 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3080-O12G-GAMING	NaN	False	https://www.amazon.it/ASUS-GeForce-Grafica-DisplayPort-Consigliata/dp/B09Q952FNZ/ref=sr_1_200?keywords=gpu&qid=1699106967&sr=8-200	4.5	1200
241	Asus ROG Strix NVIDIA GeForce RTX 3080 OC Edition Scheda Grafica, 12GB GDDR6X 384-bit 19 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3080-O12G-GAMING	NaN	False	https://www.amazon.it/ASUS-GeForce-Grafica-DisplayPort-Consigliata/dp/B09Q952FNZ/ref=sr_1_198?keywords=gpu&qid=1699107068&sr=8-198	4.5	1200
75	ASUS Dual Nvidia GeForce RTX 2060 6GB EVO Gaming Scheda Grafica, 6GB GDDR6 192-bit PCIE 4.0, GPU NVIDIA Ampere	307.00	True	https://www.amazon.it/ASUS-Nvidia-GeForce-Grafica-192-bit/dp/B07V32M9SQ/ref=sr_1_68?keywords=gpu&qid=1699106764&sr=8-68	4.6	3060
8	Asus Dual NVIDIA GeForce RTX 3060 V2 OC Edition Scheda Grafica, 12GB GDDR6 192-bit 15 Gbps PCIE 4.0, GPU NVIDIA Ampere, DUAL-RTX3060-O12G-V2	334.94	False	https://www.amazon.it/ASUS-GeForce-Grafica-192-bit-DUAL-RTX3060-O12G-V2/dp/B096658ZWP/ref=sr_1_13?keywords=gpu&qid=1699106662&sr=8-13	4.7	1145
215	Asus Dual NVIDIA GeForce RTX 3070 V2 OC Edition Scheda Grafica, 8GB GDDR6 256-bit 14 Gbps PCIE 4.0, GPU NVIDIA Ampere, DUAL-RTX3070-O8G-V2	509.90	False	https://www.amazon.it/Asus-GeForce-Grafica-256-bit-DUAL-RTX3070-O8G-V2/dp/B098JXFQD6/ref=sr_1_188?keywords=gpu&qid=1699106967&sr=8-188	4.5	1210
65	ASUS TUF Gaming NVIDIA GeForce RTX 4070 Ti OC Edition Scheda Grafica, 12GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU NVIDIA Ampere, TUF-RTX4070TI-O12G-GAMING	983.50	False	https://www.amazon.it/ASUS-GeForce-Grafica-192-bit-TUF-RTX4070TI-O12G-GAMING/dp/B0BNW6598Y/ref=sr_1_58?keywords=gpu&qid=1699106764&sr=8-58	4.5	1130
92	Asus ROG Strix NVIDIA GeForce RTX 3090 OC Edition Scheda Grafica, 24GB GDDR6X 384-bit 19.5 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3090-O24G-GAMING	2130.24	False	https://www.amazon.it/ASUS-GeForce-Grafica-384-bit-ROG-STRIX-RTX3090-O24G-GAMING/dp/B08HN51T8Q/ref=sr_1_89?keywords=gpu&qid=1699106764&sr=8-89	4.5	3140
40	Asus ROG Strix NVIDIA GeForce RTX 3090 White OC Edition Scheda Grafica, 24GB GDDR6X 384-bit 19 Gbps PCIE 4.0, GPU NVIDIA Ampere, ROG-STRIX-RTX3090-O24G-WHITE	3251.73	True	https://www.amazon.it/ASUS-GeForce-Grafica-384-bit-ROG-STRIX-RTX3090-O24G-WHITE/dp/B08P94FVLV/ref=sr_1_49?keywords=gpu&qid=1699106662&sr=8-49	4.3	290
26	ASUS TUF Gaming NVIDIA GeForce RTX 4070 OC Edition Scheda Grafica, 12GB GDDR6X 192-bit 21 Gbps PCIE 4.0, GPU NVIDIA Ada Lovelace, TUF-RTX4070-O12G-GAMING	752.55	False	https://www.amazon.it/ASUS-GeForce-Grafica-Lovelace-TUF-RTX4070-O12G-GAMING/dp/B0BVMM9VB1/ref=sr_1_31?keywords=gpu&qid=1699106662&sr=8-31	4.5	630
41	Asus ROG Strix LC NVIDIA GeForce RTX 3080 Ti OC Edition Scheda Grafica, 12 GB GDDR6X, PCIe 4.0, HDMI 2.1, DisplayPort 1.4a, Raffreddamento a Liquido, PSU Consigliata 850W, RGB, GPU Tweak II, Nero	NaN	False	https://www.amazon.it/ASUS-GeForce-DisplayPort-Raffreddamento-Consigliata/dp/B096KY13Q7/ref=sr_1_50?keywords=gpu&qid=1699106662&sr=8-50	4.3	40

Figure 7: Result of a short query using cosine similarity



Return to main menu

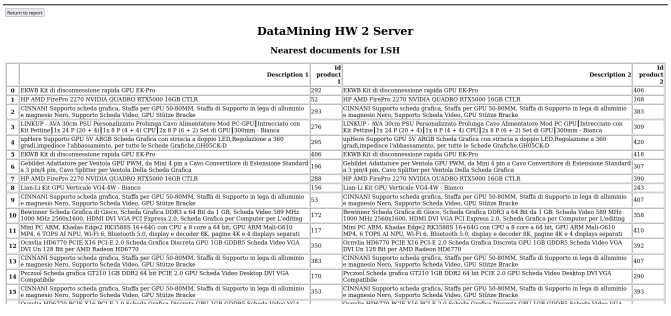
DataMining HW 2 Server

Result of query: scheda grafica nvidia 3080 ti per computer asus

insert query: Invia

	description	price	prime	url	stars	num_reviews
174	Bewinner Scheda Grafica RX580 8GB, GDDR5 256bit Scheda Grafica per Computer con Doppia Ventola 1284MHz 7000MHz Scheda Grafica Discreta per Giochi per Computer Desktop con HDMI, DP X 3, DVI	131.09	True	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo3MDE3ODE3NzAwMzgzNjcyOjE2OTkxMDY4NjY6c3BfbXRmOjMwMDA2ODUyMzc5NjQzMjo6MDo6&url=%2FBewinner-Grafica-Computer-Ventola-Discreta%2Fdp%2FB08W9HXL568%2Fref%3Dsr_1_115_sspa%3Fkeywords%3Dgpu%26qid%3D1699106866%26sr%3D8-115-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGY%26psc%3D1	3.0	10
237	Bewinner Scheda Grafica RX580 8GB, GDDR5 256bit Scheda Grafica per Computer con Doppia Ventola 1284MHz 7000MHz Scheda Grafica Discreta per Giochi per Computer Desktop con HDMI, DP X 3, DVI	131.09	True	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo0MTAxNjc1NjA5MTgyNjcyOjE2OTkxMDY5Njc6c3BfYnRmOjMwMDA2ODUyMzc5NjQzMjo6MDo6&url=%2FBewinner-Grafica-Computer-Ventola-Discreta%2Fdp%2FB08W9HXL568%2Fref%3Dsr_1_202_sspa%3Fkeywords%3Dgpu%26qid%3D1699106967%26sr%3D8-202-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGY%26psc%3D1	3.0	10
339	Bewinner Scheda Grafica RX580 8GB, GDDR5 256bit Scheda Grafica per Computer con Doppia Ventola 1284MHz 7000MHz Scheda Grafica Discreta per Giochi per Computer Desktop con HDMI, DP X 3, DVI	131.09	True	https://www.amazon.it/sspa/click?ie=UTF8&spc=MToyNjE0ND1MTgzOTkxNjcyOjE2OTkxMDcxNzA6c3BfYnRmOjMwMDA2ODUyMzc5NjQzMjo6MDo6&url=%2FBewinner-Grafica-Computer-Ventola-Discreta%2Fdp%2FB08W9HXL568%2Fref%3Dsr_1_300_sspa%3Fkeywords%3Dgpu%26qid%3D1699107170%26sr%3D8-300-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGY%26psc%3D1	3.0	10
389	Bewinner Scheda Grafica RX580 8GB, GDDR5 256bit Scheda Grafica per Computer con Doppia Ventola 1284MHz 7000MHz Scheda Grafica Discreta per Giochi per Computer Desktop con HDMI, DP X 3, DVI	131.09	True	https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo2NTg2Mzl3NzEzODg5NjcyOjE2OTkxMDcyNzI6c3BfYnRmOjMwMDA2ODUyMzc5NjQzMjo6MDo6&url=%2FBewinner-Grafica-Computer-Ventola-Discreta%2Fdp%2FB08W9HXL568%2Fref%3Dsr_1_318_sspa%3Fkeywords%3Dgpu%26qid%3D1699107272%26sr%3D8-318-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGY%26psc%3D1	3.0	10
	Bewinner Scheda Grafica RX580 8GB, GDDR5 256bit Scheda Grafica per Computer con Doppia			https://www.amazon.it/sspa/click?ie=UTF8&spc=MTo3OTY1NjI4OTI0OTc4NjcyOjE2OTkxMDcxNzQ6c3BfbXRmOjMwMDA2ODUyMzc5NjQzMjo6MDo6&url=%2FBewinner-Grafica-Computer-Ventola-		

Figure 8: Result of a long query using cosine similarity



Return to report

DataMining HW 2 Server

Nearest documents for LSH

	Description 1	Id product 1	Description 2	Id product 2
0	EKWB Kit di disconnessione rapida GPU EK-Pro	292	EKWB Kit di disconnessione rapida GPU EK-Pro	406
1	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	52	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	168
2	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	293	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	383
3	LINKUP - AVA 30cm PSU Personalizzato Prolunga Cavo Alimentatore Mod PC GPU [Intrecciato con Kit Pettine] 1x 24 P (20 + 4) 1x 8 P (4 + 4) CPU 2x 8 P (6 + 2) Set di GPU 300mm - Bianco	276	LINKUP - AVA 30cm PSU Personalizzato Prolunga Cavo Alimentatore Mod PC GPU [Intrecciato con Kit Pettine] 1x 24 P (20 + 4) 1x 8 P (4 + 4) CPU 2x 8 P (6 + 2) Set di GPU 300mm - Bianco	309
4	upThere Supporto GPU 5V ARGB Scheda Grafica con striscia a doppio LED,Regolazione a 360 gradi,Impedisce l'abbassamento, per tutte le Schede Grafiche,GH05CK.D	295	upThere Supporto GPU 5V ARGB Scheda Grafica con striscia a doppio LED,Regolazione a 360 gradi,Impedisce l'abbassamento, per tutte le Schede Grafiche,GH05CK.D	420
5	EKWB Kit di disconnessione rapida GPU EK-Pro	406	EKWB Kit di disconnessione rapida GPU EK-Pro	418
6	Gebildet Adattatore per Ventola GPU PWM, da Mini 4 pin a Cavo Convertitore di Estensione Standard a 3 pin/4 pin, Cavo Splitter per Ventola Della Scheda Grafica	196	Gebildet Adattatore per Ventola GPU PWM, da Mini 4 pin a Cavo Convertitore di Estensione Standard a 3 pin/4 pin, Cavo Splitter per Ventola Della Scheda Grafica	307
7	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	288	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	390
8	Lian-Li Kit GPU Verticale VG4-4W - Bianco	156	Lian-Li Kit GPU Verticale VG4-4W - Bianco	243
9	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	53	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	407
10	Bewinner Scheda Grafica di Gioco, Scheda Grafica DDR3 a 64 Bit da 1 GB, Scheda Video 589 MHz 1000 MHz 2560x1600, HDMI DVI VGA PCI Express 2.0, Scheda Grafica per Computer per L'editing	172	Bewinner Scheda Grafica di Gioco, Scheda Grafica DDR3 a 64 Bit da 1 GB, Scheda Video 589 MHz 1000 MHz 2560x1600, HDMI DVI VGA PCI Express 2.0, Scheda Grafica per Computer per L'editing	358
11	Mini PC ARM, Khadas Edge2 RK3588S 16+64G con CPU a 8 core a 64 bit, GPU ARM Mali-G610 MP4, 6 TOPS AI NPU, Wi-Fi 6, Bluetooth 5.0, display e decoder 8K, paghe 4K a 4 displays separati	117	Mini PC ARM, Khadas Edge2 RK3588S 16+64G con CPU a 8 core a 64 bit, GPU ARM Mali-G610 MP4, 6 TOPS AI NPU, Wi-Fi 6, Bluetooth 5.0, display e decoder 8K, paghe 4K a 4 displays separati	410
12	Ocznvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA DVI Un 128 Bit per AMD Radeon HD6770	350	Ocznvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA DVI Un 128 Bit per AMD Radeon HD6770	392
13	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	383	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	407
14	Pvcznol Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	170	Pvcznol Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	290
15	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	353	CINNANI Supporto scheda grafica, Staffa per GPU 50-80MM, Staffa di Supporto in lega di alluminio e magnesio Nero, Supporto Scheda Video, GPU Stitze Bracke	393
	Ocznvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA		Ocznvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA	

Figure 9: Some resulting near duplicates computed using LSH