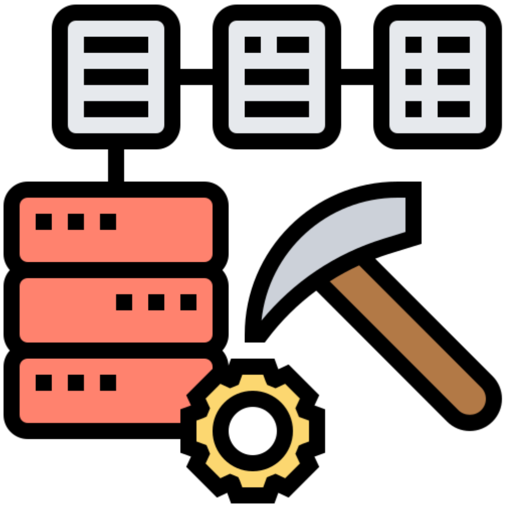

Data Mining

Homework 2

Ashkan Ansarifard

1970082

A.Y. 2023/24

Problem 1

In this report, I present the implementation and analysis of a Python program designed to scrape and analyze product data from Amazon using web scraping techniques. The objective of this problem is to collect information about products related to the keyword ***gpu*** from the Amazon.it website, parse the data, and perform an Exploratory Data Analysis. [1]

Web Scraping and Data Collection

The `AmazonScraper` class has been implemented to do the web scraping and data collection. The `scrape_amazon_products` method utilizes the `Requests` library to download web pages and `BeautifulSoup` for HTML parsing. It iterates through the specified number of pages, extracts relevant information for each product, and stores it in the `self.data` list.

Tab-Separated Value (TSV) File

The `save_to_tsv` method converts the collected data into a Pandas DataFrame and saves it to a TSV file using `pd.to_csv`. Each product's information is stored in a separate line, as written in the requirement of the problem statement.

Delay to Prevent Blocking

To prevent being blocked by Amazon due to excessive requests, a delay of 5 seconds (`time.sleep(5)`) has been introduced between different web page requests.

Exploratory Data Analysis (EDA)

The `analyze_data` method has been implemented to perform an EDA on the collected dataset. The analysis includes[2]:

- **Price Ranges**: Utilizing the Pandas `cut` function to categorize products into price ranges and visualizing the distribution using a box plot. The dataset has

[1] example usage class is `problem1_main.py`
[2] The EDA is from analyzing the first 10 pages

been categorized into the following price ranges:

Price Range	Count
500+	72
≤100	33
100-200	11
200-300	4
300-400	0
400-500	0

Table 1: Distribution of Products Across Price Ranges (Console Output)

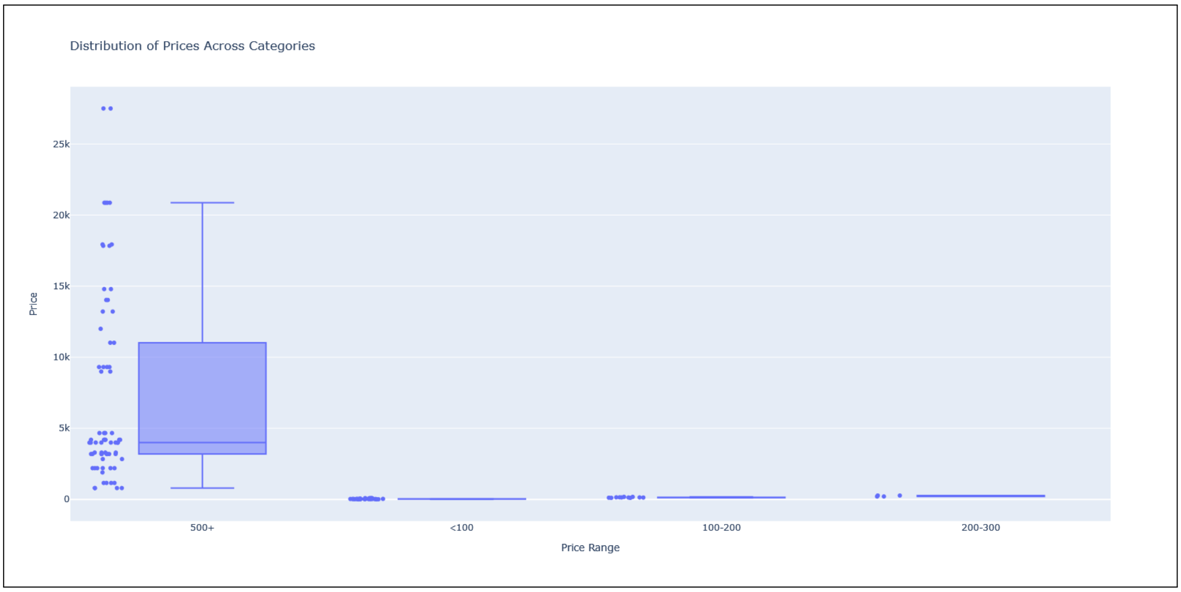

Figure 1: Distribution of prices across different categories

- **Customer Reviews**: Identifying and printing the top-rated products based on star ratings.

Figure 2: Top Rated Products with Customer Reviews

- **Primeness**: Separating the dataset into Prime and Non-Prime products and providing summary statistics for each category.

	Price	Stars	Reviews
count	31.000000	31.000000	31.000000
mean	3569.806452	1.458065	1409.645161
std	5977.225764	2.148142	2630.628120
min	39.000000	0.000000	0.000000
25%	41.000000	0.000000	1.500000
50%	147.000000	0.000000	5.000000
75%	4195.000000	4.500000	2010.000000
max	20870.000000	4.700000	7114.000000

Table 2: Summary statistics for Prime Products. (Console Output)

	Price	Stars	Reviews
count	89.000000	89.000000	89.000000
mean	3292.707865	1.767416	526.483146
std	11187.100932	2.209112	1671.567141
min	7.000000	0.000000	0.000000
25%	31.000000	0.000000	1.000000
50%	132.000000	0.000000	7.000000
75%	3195.000000	4.400000	62.000000
max	99990.000000	4.700000	7114.000000

Table 3: Summary statistics for Non-Prime Products. (Console Output)

- **Top 10 Products by Rating and Price**: Plotting the top 10 products based on both star rating and price using Plotly Express bar charts.

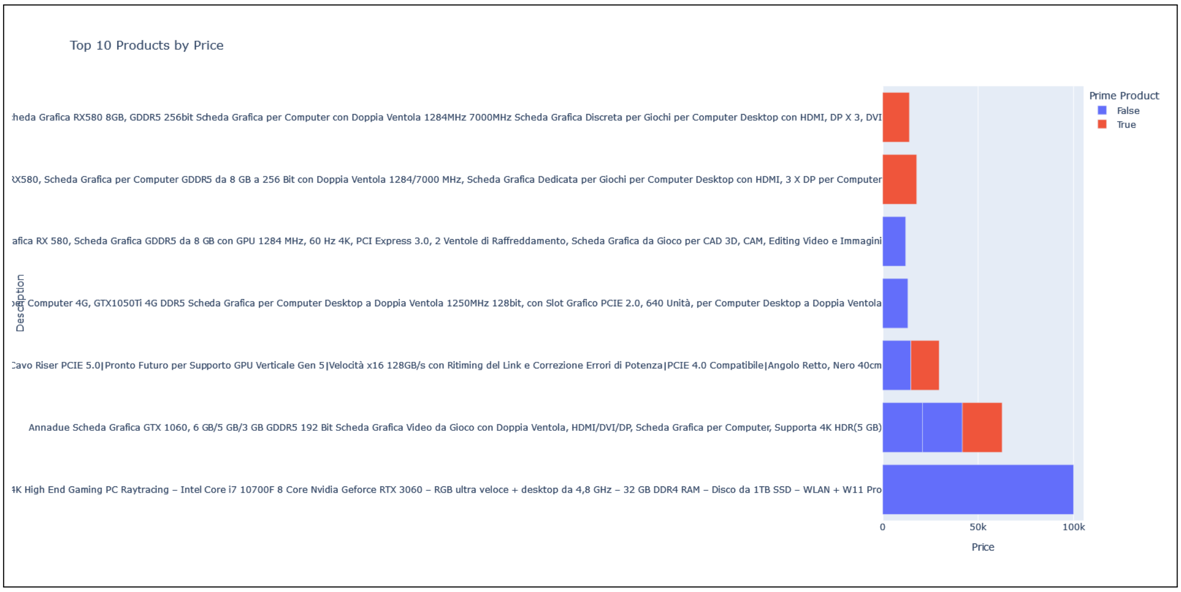

Figure 3: Top 10 Products by Price

Figure 4: Top 10 Products by Star Rating

- **Scatter Plot of Price vs. Star Rating**: Creating a scatter plot to visualize the relationship between product price and star rating, with marker size indicating the number of reviews.

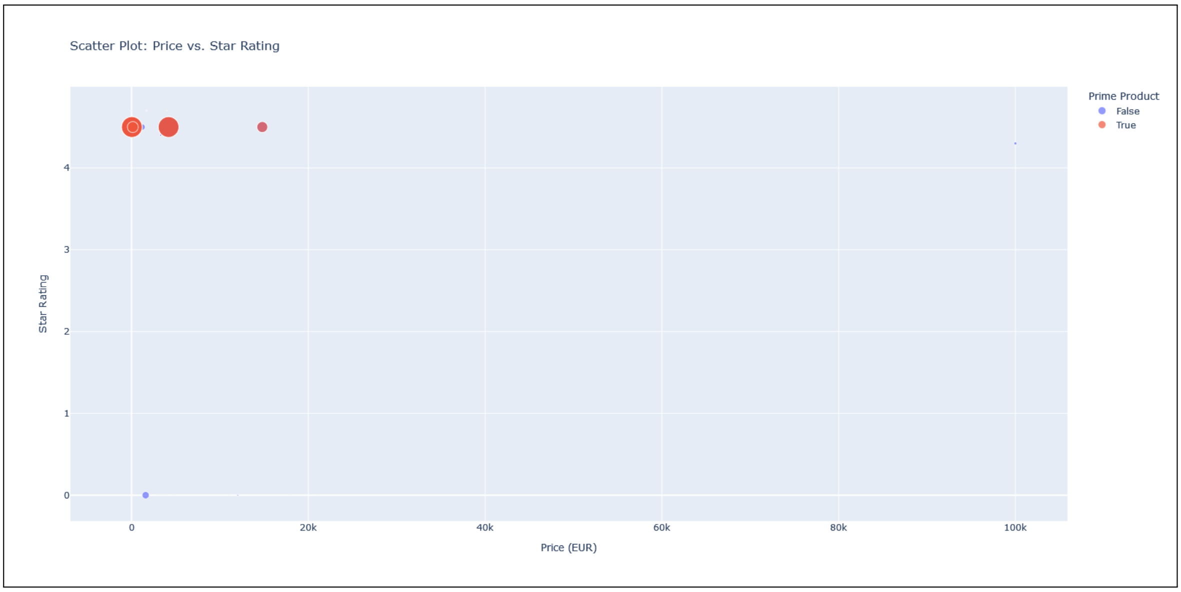

Figure 5: Scatter Plot Price vs. Star Rating

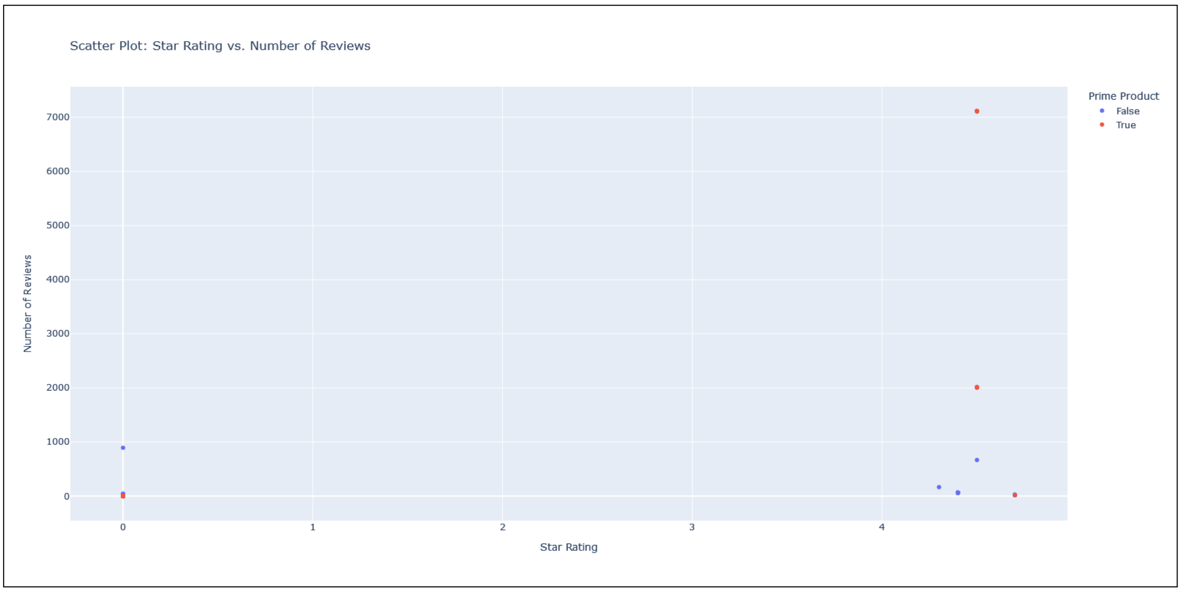

Figure 6: Scatter Plot Star Rating vs. Number of Reviews

Problem 2

In this section, I present the implementation of a search engine extension for the `AmazonScraper` class. The search engine focuses on evaluating queries based on the product's description, returning the top 10 most related products along with their information.[3]

Class Implementation: AmazonScraperWithSearchEngine

The `AmazonScraperWithSearchEngine` class extends the `AmazonScraper` class to incorporate search engine functionalities. Below, I detail how each requirement in the problem statement has been fulfilled.

Inverted Index and Cosine Similarity

The `evaluate_query` method has been overridden to include vectorization and cosine similarity calculation. This method takes a user query, transforms it into a TF-IDF vector, and computes cosine similarities with the TF-IDF matrix of product descriptions. The related product indices are then sorted based on cosine similarities, and the top products are returned.

Build Inverted Index

The `build_inverted_index` method initializes a TF-IDF vectorizer and computes the TF-IDF matrix based on the product descriptions. This serves as the inverted index for the search engine. It checks for the availability of data and prompts the user to run `scrape_amazon_products` if no data is present.

Usage and Example

To use the search engine, first, instantiate the `AmazonScraperWithSearchEngine` class with a keyword and the number of pages to scrape. Run `scrape_amazon_products`

[3] example usage class is `problem2_main.py`

to collect data, and then execute `build_inverted_index` to build the inverted index. Once the inverted index is available, queries can be evaluated using `evaluate_query`.

Example Usage

To demonstrate the usage of the `AmazonScraperWithSearchEngine` class, consider the following example which can be found also in `problem2_main.py`:

1. **Instantiate the Search Engine:**

   ```
   search_engine = AmazonScraperWithSearchEngine(keyword='gpu
   num_pages=5)
   ```

2. **Scrape Amazon Products:**

   ```
   search_engine.scrape_amazon_products()
   ```

3. **Build the Inverted Index:**

   ```
   search_engine.build_inverted_index(search_engine.df['Descri
   ```

4. **Evaluate a Query:**

   ```
   user_query = "high-performance graphics card"
   query_result = search_engine.evaluate_query(user_query,
   top_N=5)
   print("Query Result:")
   print(query_result)
   ```

This example demonstrates the workflow of using the *AmazonScraperWithSearchEngine* class. It begins by instantiating the class with a specified keyword and the number of pages to scrape. The class is then used to scrape Amazon products, build the inverted index based on product descriptions, and finally, evaluate a sample query using the cosine similarity measure. The resulting query result includes the top-related products along with their information.

Example Usage Results

In the following, result of the search engine for different query lengths are shown. The list of queries are:

1. Gaming GPU

1	Description	Price	Prime Product	Stars	Reviews	URL
2	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www
3	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www
4	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www
5	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www
6	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www
7	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www
8	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www
9	PUSOKEI Scheda Grafica RX 580, 8 GB GDDR5 256 Bit 1284/7000 MHz PCI Express 3.0 Scheda Grafica per Giochi Minerari, Doppia Ventola Scheda Video con 3 DP, HDMI, DVI, per PC Gaming Mining	129	FALSE	0	0	https://www
10	GreedÂ® Mk2 4K High End Gaming PC Raytracing â€“ Intel Core i7 10700F 8 Core Nvidia Geforce RTX 3060 â€“ RGB ultra veloce + desktop da 4,8 GHz â€“ 32 GB DDR4 RAM â€“ Disco da 1TB SSD â€“ WLAN + V	999	FALSE	4.3	167	https://www
11	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13.99	FALSE	4.5	706	https://www
12						

Figure 7: Query result - Gaming GPU

2. GPU for video editing

1	Description	Price	Prime Product	Stars	Reviews	URL
2	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www.amazon.it/ss
3	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www.amazon.it/ss
4	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www.amazon.it/ss
5	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	0	7	https://www.amazon.it/ss
6	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.amazon.it/ss
7	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.amazon.it/ss
8	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.amazon.it/ss
9	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39	TRUE	0	0	https://www.amazon.it/ss
10	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39	TRUE	0	0	https://www.amazon.it/ss
11	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39.94	FALSE	0	0	https://www.amazon.it/ss
12						

Figure 8: Query result - GPU for video editingd

3. Best budget GPU

1	Description	Price	Prime Product	Stars	Reviews	URL
2	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	TRUE	4.5	706	https://www.amazon.it/sspa/click?
3	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	FALSE	4.5	706	https://www.amazon.it/sspa/click?
4	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13.99	FALSE	4.5	706	https://www.amazon.it/sspa/click?
5	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	FALSE	4.5	706	https://www.amazon.it/sspa/click?
6	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13.99	FALSE	4.5	706	https://www.amazon.it/sspa/click?
7	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13.99	FALSE	4.5	706	https://www.amazon.it/sspa/click?
8	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	FALSE	4.5	706	https://www.amazon.it/sspa/click?
9	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	FALSE	4.5	706	https://www.amazon.it/sspa/click?
10	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13	FALSE	4.5	706	https://www.amazon.it/sspa/click?
11	upHere Nero Supporto GPU Scheda Grafica,360 gradi di regolazione Supporto GPU in Alluminio,impedisce la Scheda Grafica Sag Holder, Supporta tutte le Schede Grafiche,GH05K	13.99	FALSE	4.5	706	https://www.amazon.it/sspa/click?
12						

Figure 9: Query result - Best budget GPU

4. NVIDIA GPU

1	Description	Price	Prime Product	Stars	Reviews	URL
2	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
3	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
4	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
5	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
6	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
7	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
8	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
9	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
10	HP AMD FirePro 2270 NVIDIA QUADRO RTX5000 16GB CTLR	1719	FALSE	4.7	2	https://www.am
11	GreedÂ® Mk2 4K High End Gaming PC Raytracing â€“ Intel Core i7 10700F 8 Core Nvidia Geforce RTX 3060 â€“ RGB ultra veloce + desktop da 4,8 GHz â€“ 32 GB DDR4 RAM â€“ Disco da 1TB SSD â€“ WLAN	999	FALSE	4.3	167	https://www.am
12						

Figure 10: Query result - NVIDIA GPU

5. AMD gaming graphics card

1	Description	Price	Prime Product	Reviews	Stars	URL
2	PCIE Riser Card 8 condensatori, GPU Extender Riser Card per Bitcoin Litecoin ETH Ethereum Mining, con cavo di prolunga USB 3.0 da 60 cm e cavo di alimentazione SATA 6PIN (V009s-PLUS, 2 pezzi)	18	FALSE	119	4.3	https://www.amaz
3	PCIE Riser Card 8 condensatori, GPU Extender Riser Card per Bitcoin Litecoin ETH Ethereum Mining, con cavo di prolunga USB 3.0 da 60 cm e cavo di alimentazione SATA 6PIN (V009s-PLUS, 2 pezzi)	18	FALSE	119	4.3	https://www.amaz
4	PCIE Riser Card 8 condensatori, GPU Extender Riser Card per Bitcoin Litecoin ETH Ethereum Mining, con cavo di prolunga USB 3.0 da 60 cm e cavo di alimentazione SATA 6PIN (V009s-PLUS, 2 pezzi)	18.99	FALSE	119	4.3	https://www.amaz
5	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	22	0	https://www.amaz
6	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	22	0	https://www.amaz
7	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	22	0	https://www.amaz
8	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	7	0	https://www.amaz
9	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	7	0	https://www.amaz
10	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	7	0	https://www.amaz
11	PELADN AMD Radeon RX580 8GB Scheda grafica (Gaming). per Giochi, Streaming, Media, Editing Video e Graphic Design. Basso consumo (130 W)	129	TRUE	7	0	https://www.amaz
12						

Figure 11: Query result - AMD gaming graphics card

6. Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU

1	Description	Price	Prime Product	Stars	Reviews	URL
2	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.a
3	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.a
4	SAPLOS Radeon RX 550 Low Profile Scheda Video, 4GB, GDDR5, 128-bit, VGA DVI-D HDMI, Video Card PC Gaming, 4k Displays, Computer GPU, SFF Small Form Factor (includere mini staffa)	119	TRUE	0	22	https://www.a
5	Ocnvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA DVI Un 128 Bit per AMD Radeon HD6770	110	TRUE	0	3	https://www.a
6	Ocnvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA DVI Un 128 Bit per AMD Radeon HD6770	110.21	TRUE	0	3	https://www.a
7	Ocnvlia HD6770 PCIE X16 PCI-E 2.0 Scheda Grafica Discreta GPU 1GB GDDR5 Scheda Video VGA DVI Un 128 Bit per AMD Radeon HD6770	110	TRUE	0	3	https://www.a
8	PUSOKEI Scheda Grafica RX 580, 8 GB GDDR5 256 Bit 1284/7000 MHz PCI Express 3.0 Scheda Grafica per Giochi Minerari, Doppia Ventola Scheda Video con 3 DP, HDMI, DVI, per PC Gaming Mining	129	FALSE	0	0	https://www.a
9	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39.94	TRUE	0	0	https://www.a
10	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39	TRUE	0	0	https://www.a
11	Pvczool Scheda grafica GT210 1GB DDR2 64 bit PCIE 2.0 GPU Scheda Video Desktop DVI VGA Compatibile	39.94	FALSE	0	0	https://www.a
12						

Figure 12: Query result - Long Query GPU

Problem 3

Part 1: Shingling

The `Shingling` class generates shingles from a given document. Shingles are created by sliding a window of size k through the document. Each shingle is a substring of length k. The `generate_shingles` method creates a set of shingles from the input document.

Part 2: Minwise Hashing

The `MinwiseHashSignature` class is designed for minwise hashing. It generates multiple hash functions using the MD5 algorithm. For each set of elements, the class updates a signature matrix with hash values. The `generate_signatures` method takes a collection of sets and returns a signature matrix, where each column represents the minwise hash signature of a set.

Part 3: Locality-Sensitive Hashing (LSH)

The `LSH` class implements the Locality-Sensitive Hashing technique. It uses minwise hash signatures to build hash tables. The `index_signatures` method populates hash tables based on the signatures of the input sets. The `query_signatures` method retrieves candidate sets by querying hash tables with the signature of a query set. The `find_near_duplicates` method uses LSH to find near-duplicates within a collection of shingles.

Part 4: Threshold Intersection Analysis

The class contains a method named `s_curve_plot_and_analysis` that analyzes the threshold intersection for different values of r and b. It iterates through combinations of r and b, calculates the probability of becoming a candidate using the S-curve formula, and plots the results. The analysis involves checking for step-shaped curves that indicate optimal values for r and b.

Part 5: Choosing Optimal Parameters

The `choose_r_b_values` method in the `LSH` class helps in selecting suitable values of r and b based on a given threshold probability. It iterates over possible combinations of r and b, calculates the expected threshold probability, and selects values that closely match the given threshold.

Part 6: Jaccard Similarity Calculation

The `ShingleComparison` class calculates Jaccard similarity between sets of shingles. The `compare_shingle_sets` method takes a collection of shingles and descriptions, computes Jaccard similarities, and identifies near-duplicate pairs based on a specified threshold.

Performance Metrics

- The elapsed time for LSH execution is measured using the `elapsed_time_lsh` attribute.
- The Jaccard similarities between shingle sets are stored in the `jaccard_values` attribute.
- The near-duplicate pairs and their Jaccard similarities are stored in a DataFrame (`df`) in the `ShingleComparison` class.

Problem 4

The provided code implements the Locality-Sensitive Hashing technique using Apache Spark. The steps involve tokenizing product descriptions, creating Word2Vec representations, and using MinHashLSH for approximate similarity joins. Below is an overview of the key components of the Spark implementation:

1. **Initialization:** The Spark session is initialized with the name "AmazonScraperLSH".
2. **DataFrame Conversion:** The Pandas DataFrame obtained from web scraping is converted to a Spark DataFrame (spark_df).
3. **Tokenization:** Product descriptions are tokenized using the 'Tokenizer' class.
4. **Word2Vec Representation:** Word2Vec representations of product descriptions are created using the 'Word2Vec' model.
5. **MinHashLSH:** MinHashLSH is applied to generate hash values for the features, and the DataFrame is divided into two halves for efficient processing.
6. **Approximate Similarity Join:** The code performs an approximate similarity join using LSH with a specified Jaccard distance threshold.
7. **Result Presentation:** The results are presented in the form of a DataFrame showing near-duplicate pairs along with their Jaccard distances.
8. **File Export:** The final near-duplicate information is saved to a CSV file, and the file is made available for download.