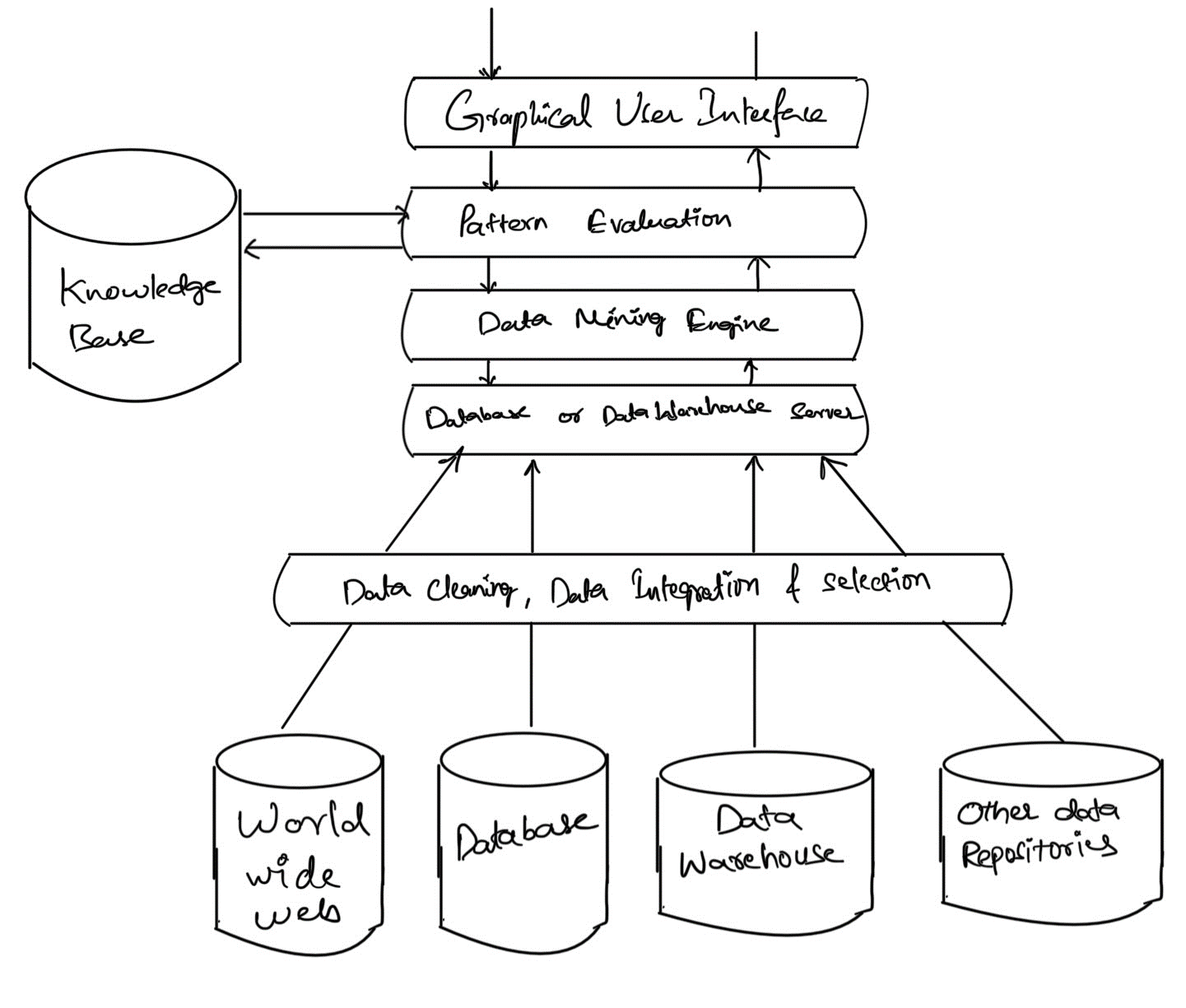
**CMPE 256, Web and Big Data Mining (Final exam)**

1. Draw the reference architecture of the Data Mining System? (**5**)



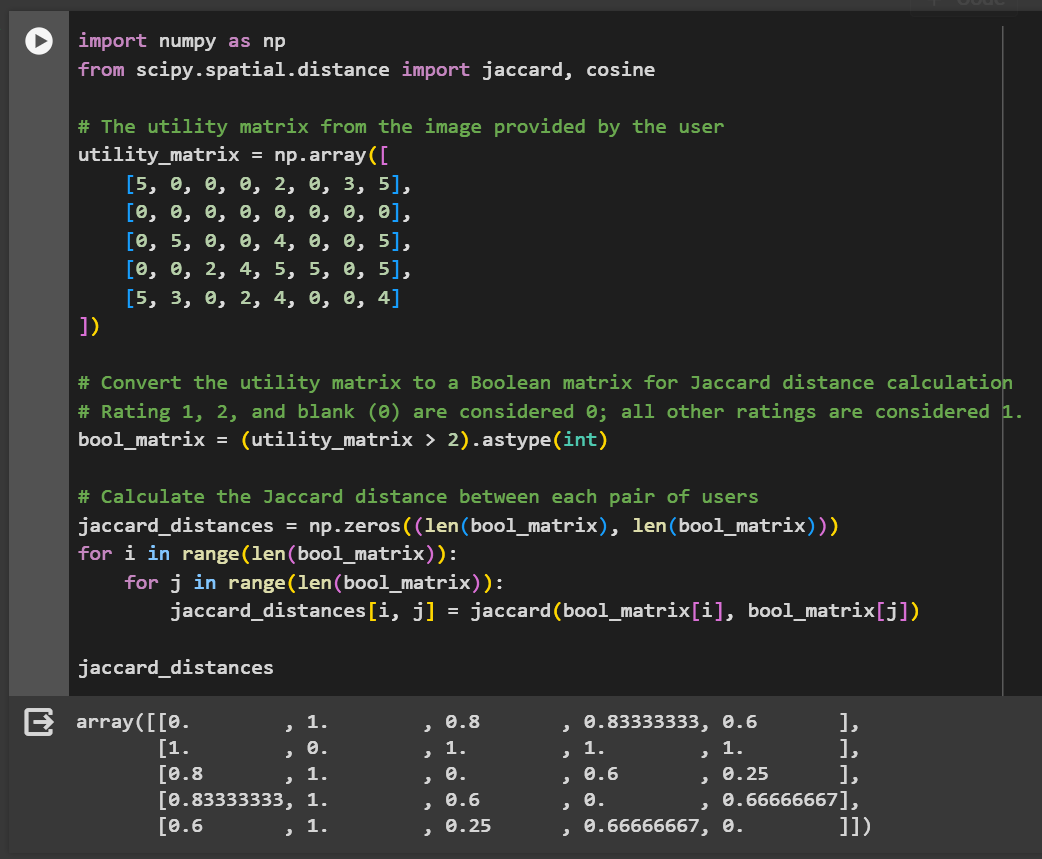
1. After graduating from San Jose State University, you have started Big Data Analytics Consulting Company - DataCo, inc. DataCo is specialized in delivering recommendation systems for online and Brick-Mortar companies. During the first month of your business, publisher from Harper Collins Publishers visited your company to develop a special recommendation system for their existing clients. As part of meeting, Harper Collins noted that they publish books in various categories, and it is immensely difficult to develop an effective marketing campaign based on User past rating behavior. As part of the business agreement, Harper wanted develop Proof Of Concept (POC) recommendation system for the following table (Note: 1 to 5 is Star scale – 5 the best) .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **User vs.**  **Book**  **Reading**  **Table** | **Built to Last** | **The**  **HP**  **Way** | **Physics**  **Made**  **Easy** | **The**  **Wisdom**  **of**  **Teams** | **Baby Bear** | **Experimental Chemistry** | **Charlie the Ranch**  **Dog** | **The Good**  **to**  **Great** | **Electronic devices and**  **circuits –**  **Millman**  **& Halkias** |
| **User 1** | 5 |  |  |  | 2 |  | 3 | 5 |  |
| **User 2**  **User 3**  **User 4**  **User 5** |  |  | 5 |  |  | 4 |  |  | 5 |
|  |  | 2 |  | 4 | 5 | 5 |  | 5 |
|  | 5 |  | 4 |  |  |  | 5 | 1 |
| 5 | 3 |  | 2 | 4 |  |  | 4 |  |

Develop recommendations system? (Hint: an effective recommendation system should consider following: (**15**)

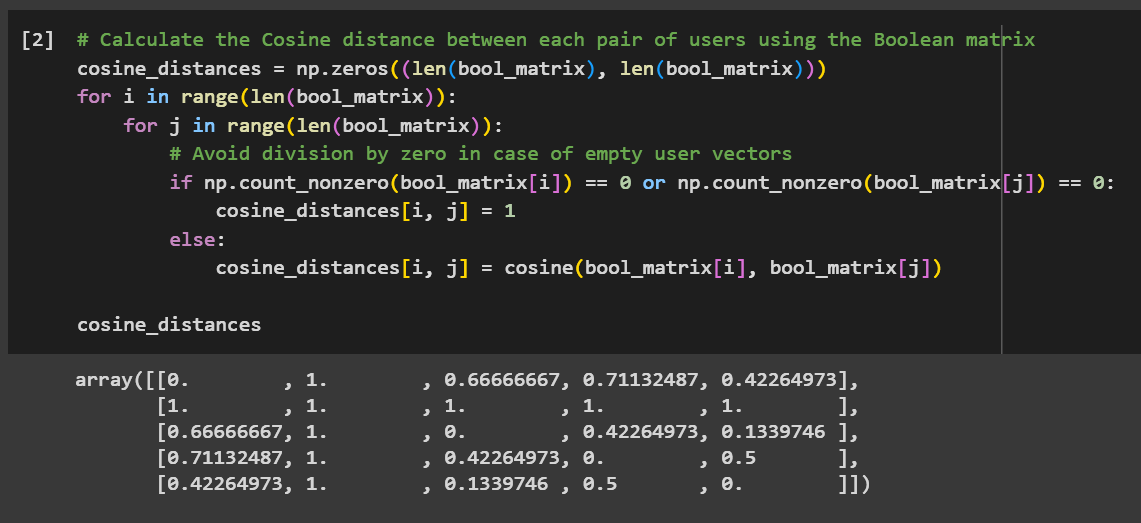
* + 1. Utility matrix as a Boolean, compute Jaccard distance between each pair
    2. Repeat a with Cosine distance
    3. Treat 3, 4, 5 as 1 and 1,2, and blank as 0. Compute Jaccard distance between each pair
    4. Repeat C with Cosine distance
    5. Normalize the matrix by subtracting from each nonblank entry the average value for its user.
    6. Using the normalized matrix from e, compute the cosine distance between each pair?

Answer:



The Jaccard distances between each pair of users have been computed using a Boolean utility matrix, where ratings of 3, 4, and 5 are considered a '1' (positive) and ratings of 1, 2, and blanks are considered a '0' (negative). Here are the distances:

* User 1 has Jaccard distances of 0.8, 0.833, and 0.6 with Users 3, 4, and 5, respectively.
* User 2 has not rated any books and hence has a distance of 1 with all other users.
* User 3 has distances of 0.6 and 0.25 with Users 4 and 5.
* User 4 has a distance of 0.666 with User 5.



The Cosine distances between each pair of users have been computed using the same Boolean matrix. The results are as follows:

* User 1 has Cosine distances of 0.667, 0.711, and 0.423 with Users 3, 4, and 5, respectively.
* User 2 has not rated any books, resulting in a Cosine distance of 1 with all other users.
* User 3 has Cosine distances of 0.423 and 0.134 with Users 4 and 5.
* User 4 has a Cosine distance of 0.5 with User 5.

1. Explain clearly all the steps involved in Knowledge Discovery Process (KDD) (**5**)?

**Answer:**

The Knowledge Discovery Process (KDD): Unlocking Insights from Data

The Knowledge Discovery Process (KDD) is an iterative, multi-step journey that transforms raw data into valuable knowledge. It's not just about crunching numbers, but about extracting meaningful patterns, trends, and insights that can inform better decision-making. Here's a breakdown of the key steps involved:

1. Goal Definition and Application Understanding:

* Identify the business problem or opportunity: What are you trying to achieve with this data analysis? Are you aiming to optimize costs, predict customer behavior, or detect fraud?
* Define the target audience: Who will be using the insights generated? Understanding their needs and expectations helps tailor the analysis approach.

2. Data Selection and Integration:

* Identify relevant data sources: This could include internal databases, external datasets, sensor readings, or even text documents.
* Data selection and pre-processing: Choose the data that best aligns with your goals and clean it by removing inconsistencies, missing values, and irrelevant information.
* Data integration: Combine data from various sources into a single, consistent format for analysis.

3. Data Transformation and Feature Engineering:

* Data transformation: Apply techniques like normalization, scaling, and discretization to prepare the data for specific algorithms.
* Feature engineering: Extract relevant features (characteristics) from the data that contribute to the desired outcome. This might involve creating combinations of existing features or deriving new ones.

4. Data Mining and Pattern Recognition:

* Choose appropriate data mining techniques: Based on your goals and data characteristics, select algorithms like clustering, classification, regression, or association rule learning.
* Apply data mining techniques: Train and run the chosen algorithms on the prepared data to identify patterns, relationships, and trends.
* Interpret and evaluate patterns: Analyze the results, assess their validity and significance, and identify actionable insights.

5. Knowledge Consolidation and Deployment:

* Present and communicate insights: Translate the technical findings into a clear and concise format understandable to the target audience.
* Validate and refine knowledge: Share the insights with stakeholders, gather feedback, and refine the knowledge based on real-world application.
* Deploy and integrate knowledge: Implement the findings into decision-making processes, business operations, or predictive models for real-world impact.

Iteration and Continuous Improvement:

The KDD process is not linear; it's an iterative journey. Throughout the process, you may need to revisit previous steps, refine your approach, or incorporate new data and insights. This continuous loop of learning and improvement ensures you're extracting the most valuable knowledge from your data.

1. After graduating from San Jose State University, you have joined data science company that provides insights to governments and public agencies. The company has a flagship Data science product for analyzing day to day trends in the city. Company clients includes City of San Jose, Mountain View, Santa Clara, Dubai, Delhi, Hyderabad, London, and many.



Given your Data Science background, they would like to provide a Data Mining solution architecture to their new Data and Analytics Platform (DnA) that gets the following data into their data warehouse and data lake estate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Frequency** | **Number of**  **Sources** | **Size** | **Intended Usage** |
| **Macroeconomic data (For example, GDP, Inflation, and other details)** | Monthly | 30 Key  Indicators | Parameters time 1 record per month | Pricing Models & Economic Models |
| **Sensor Hourly Data**  **- Audi Data** | Hourly | 20000 Sensors | 1 Mb File per Hour  (360 x # of Hours( | Audio Analysis to help emerging needs (customer sentiments & other choices) |
| **Video Data** | Hourly | Street Traffic  Lights –  30,000 | 5 MB Video data | Violations &  Public Safety  Analytics |
| **Emergency Call**  **Data (Example, 911**  **System)** | Per minute | 100000 | Audio and Call Data – length 1 minute interval | Crime and Signal Analysis |
| **Marketing & Public Events data. For example, new attractions in city, entertainment event data, and Sport & Others.** | Per Weekly | 10 | 10 MB Time Series  Event Data | Send event coupons and  recommendations |

* + What would be your design principles for developing advanced data mining architecture? (**6**)
  + Please provide what consideration do you take for handling data frequencies? (**5**) • Please provide overall Data Mining Architecture? (**10**)

1. Explain Information Retrieval? Describe detailed IR System architecture? Explain how Google IR System works? (Hint: Architecture diagram) (**10**)

**Answer:**

Information Retrieval (IR): Finding What You Need

Information retrieval (IR) deals with the search and retrieval of relevant information from large collections of data. It's the technology behind search engines like Google that help you find what you're looking for in the vast ocean of online information.

Detailed IR System Architecture:

Imagine an IR system as a pipeline with four main stages:

1. Ingestion and Preprocessing:

* Data Collection: Web crawlers gather documents from the web, databases, or other sources.
* Parsing and Cleaning: Documents are cleaned by removing noise, extracting text, and applying basic language processing techniques like tokenization and stemming.
* Indexing: Documents are analyzed and key terms are extracted to create an index, acting as a map for efficient searching.

2. Query Understanding:

* Query Parsing: User queries are analyzed to understand their intent, identify keywords, and remove ambiguity.
* Query Expansion: Synonyms and related terms are added to expand the search scope and improve recall (finding relevant documents).
* Query Weighting: Terms are assigned weights based on their importance in the query, influencing relevance ranking.

3. Matching and Ranking:

* Document Retrieval: Indexed documents are retrieved based on their keyword match with the expanded query.
* Relevance Scoring: Documents are scored based on various factors like keyword frequency, term proximity, document popularity, and other ranking algorithms.
* Ranking and Filtering: Documents are ranked by their relevance score and filtered to return the most relevant ones to the user.

4. Presentation and Evaluation:

* Result Presentation: Top-ranked documents are presented to the user in a clear and informative way, often with snippets and additional information.
* User Feedback and System Learning: User feedback (clicks, dwell time, etc.) is used to improve the relevance ranking and personalize future searches.
* Evaluation Metrics: IR systems are evaluated using metrics like precision (relevant documents retrieved), recall (relevant documents not missed), and F1 score (balance of both).

Google's IR System:

Google's IR system is incredibly complex and constantly evolving. However, it follows a similar architecture to the one described above. Here are some key highlights:

* Massive Data Ingestion: Google crawls and indexes billions of web pages constantly, creating a vast and dynamic index.
* Sophisticated Ranking Algorithms: Google uses a complex algorithm called PageRank, along with other factors like user behavior and content quality, to rank search results.
* Personalization: Google personalizes search results based on your past search history, location, and other factors.
* Continuous Learning and Improvement: Google's IR system is constantly learning and adapting based on user feedback and new data, making it one of the most powerful search engines in the world.

Architecture Diagram:

Here's a simplified diagram of the IR system architecture:

+--------------------+

| Data Collection |

+--------------------+

|

| Preprocessing &

| Indexing

|

+--------------------+

| Query Understanding |

+--------------------+

|

| Matching &

| Ranking

|

+--------------------+

| Presentation & |

| Evaluation |

+--------------------+

1. Answer following Objective Type questions (**5**)

|  |
| --- |
| Using collaborative filtering to generate recommendations is computationally expensive.    a) True b) False |
| Association rule mining considers the sequence in which the items are purchased. a) True b) False |
| According to the Apriori Principle, If an itemset is frequent, then all its \_\_\_\_\_\_\_\_\_\_\_\_\_ must also be frequent. |
| In Social Network Analysis , centrality focuses on in-links while prestige focuses on out-links a) True b) False |
| \_\_\_\_\_\_\_\_\_\_\_\_\_ Algorithm works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. |

Using collaborative filtering to generate recommendations is computationally expensive.

**a) True**

Association rule mining considers the sequence in which the items are purchased.

**b) False**

According to the Apriori Principle, If an itemset is frequent, then all its \_\_\_\_\_\_\_\_\_\_\_\_\_ must also be frequent.

**Subsets**

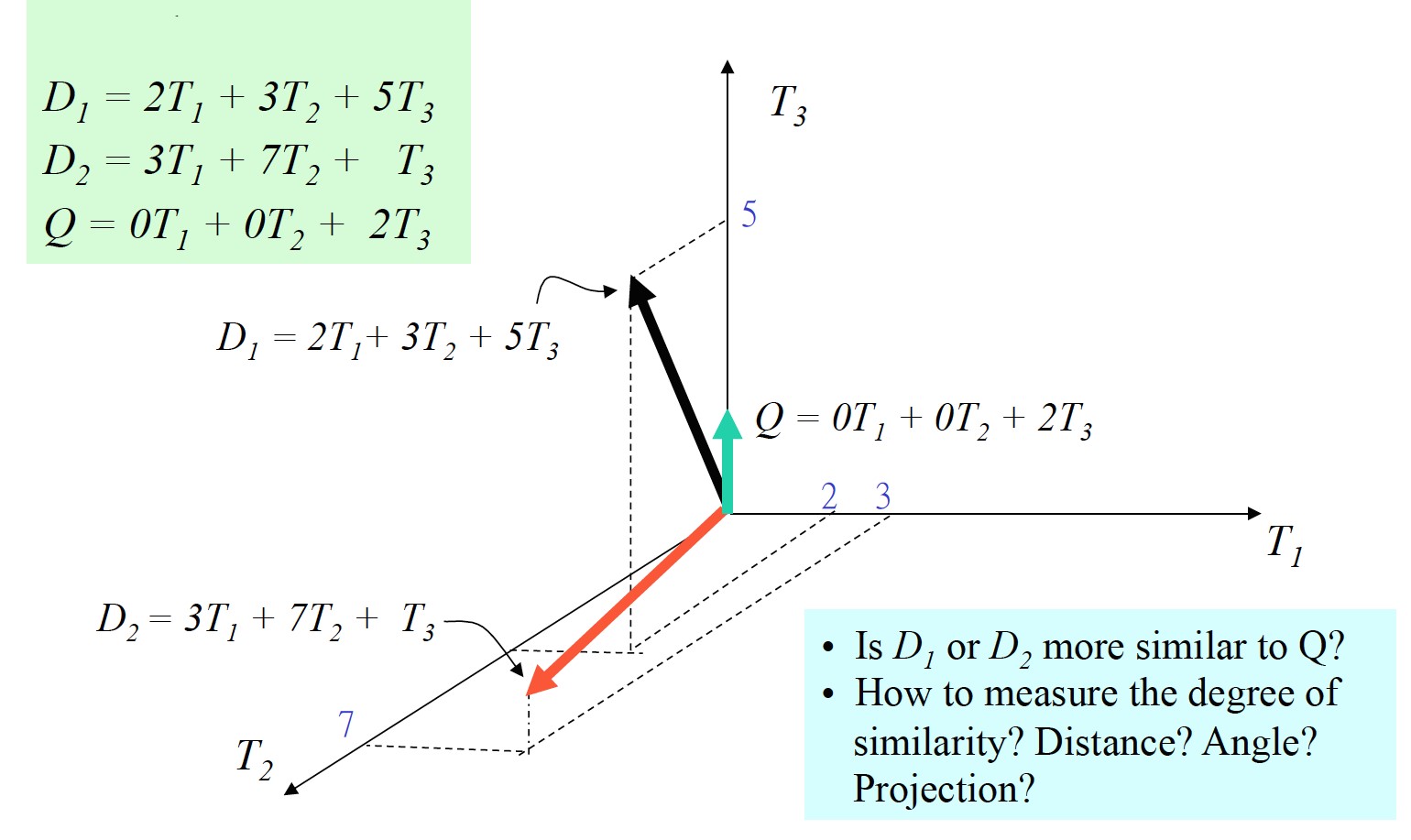
In Social Network Analysis , centrality focuses on in-links while prestige focuses on out-links

**b) False**

\_\_\_\_\_\_\_\_\_\_\_\_\_ Algorithm works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.

**PageRank**

1. Calculate (**6**)



**Answer:**

The image appears to be a diagram representing vectors in a three-dimensional space, defined by the basis vectors  *T*1​,*T*2​, and *T*3​. There are three vectors given:

* *D*1​=2*T*1​+3*T*2​+5*T*3​
* *D*2​=3*T*1​+7*T*2​+*T*3​
* *Q*=0*T*1​+0*T*2​+2*T*3​

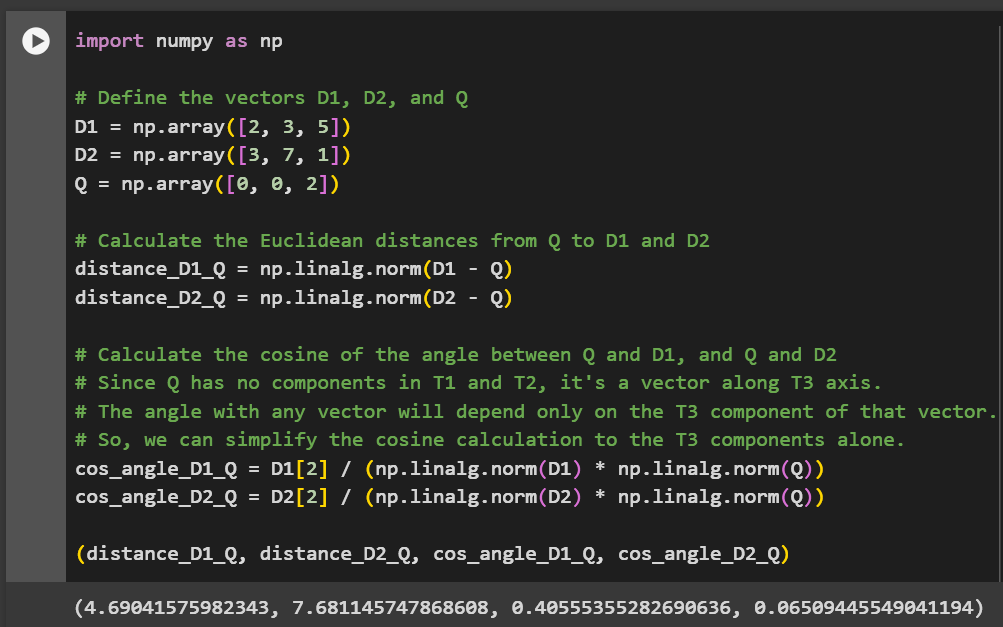
The questions posed are:

* Is *D*1​ or *D*2​ more similar to *Q*?
* How to measure the degree of similarity? Distance? Angle? Projection?

To determine which vector is more similar to *Q*, we can use various methods:

1. **Distance**: Calculate the Euclidean distance between each *D* vector and *Q*. The vector with the smallest distance to *Q* is the most similar.
2. **Angle**: Calculate the angle between each *D* vector and *Q*. The vector with the smallest angle to *Q* is the most similar.
3. **Projection**: Calculate the projection of each *D* vector onto *Q*. The vector with the largest projection onto *Q* is the most similar.

Let's calculate both the Euclidean distance and the cosine of the angle between each *D* vector and *Q* to answer the question.



Based on the calculations:

* The Euclidean distance between *D*1​ and *Q* is approximately 4.69.
* The Euclidean distance between *D*2​ and *Q* is approximately 7.68.

Since the distance from *D*1​ to *Q* is shorter than from *D*2​ to *Q*, *D*1​ is more similar to *Q* in terms of distance.

In terms of the angle:

* The cosine of the angle between *D*1​ and *Q* is approximately 0.406.
* The cosine of the angle between *D*2​ and *Q* is approximately 0.065.

The cosine of the angle is higher for *D*1​ and *Q*, which means that the angle between *D*1​ and *Q* is smaller than the angle between *D*2​ and *Q*. Therefore, *D*1​ is also more similar to *Q* in terms of angle.

Overall, *D*1​ is more similar to *Q* than *D*2​ is, based on both distance and angle measures.

1. Please list fundamental Stream Processing Rules. (5)

**Answer:**

Stream processing is a crucial component in modern data-driven enterprises. It enables real-time analysis and action on data streams as they are generated, contrasting with traditional batch processing, which works on static datasets. The power of stream processing lies in its ability to handle and process data in motion, providing businesses with timely insights and responses.

Here are some fundamental rules and concepts of stream processing:

1. Event Time vs. Processing Time: Event time signifies the moment an event actually occurs in the real world. This can be crucial in scenarios where events arrive out of order, ensuring that data is processed accurately, reflecting the actual sequence of occurrences. Processing time, on the other hand, is when the system processes the event, which may not always align with its actual occurrence.
2. Windowing and Time Windows: One of the fundamental concepts in stream processing is the notion of time windows. This allows us to segment the continuous flow of data into discrete, manageable chunks for analysis. There are various types of time windows, each suited for different analytical purposes:

* Tumbling Windows: These windows do not overlap, and each event belongs to a single window. Tumbling windows are well-suited for scenarios where discrete, non-overlapping analysis is required.
* Hopping Windows: Unlike tumbling windows, hopping windows can overlap. This means that an event can belong to multiple windows. In a hopping window, a fixed-size time interval is defined, and the window “hops” forward by a specified increment. This allows for continuous and dynamic analysis of data.

1. Temporal and Event-Time Processing: The ability to handle event time processing is becoming crucial for many use cases. This involves accurately processing events based on the time they occurred, rather than when they were ingested into the system. Temporal processing ensures accurate analysis and decision-making, especially in scenarios where event ordering is critical.
2. Unified Data Platforms: The convergence of stream processing with other data processing paradigms (such as batch processing and interactive querying) within unified data platforms is gaining prominence. This provides a holistic view of data processing across various time scales and use cases, enabling organizations to leverage a single platform for diverse analytical needs.
3. AI and Machine Learning Integration: The integration of artificial intelligence (AI) and machine learning (ML) capabilities into stream processing workflows is a significant trend. This enables real-time data-driven decision-making, anomaly detection, and predictive analytics. Stream processing platforms are increasingly offering native support for deploying and executing ML models within the processing pipeline.

These rules and concepts collectively indicate an exciting future for stream processing, with advancements in technology and methodologies enabling organizations to derive even greater value from real-time data streams.

Rewrite

Top of Form

1. As a Data Analytics Consultant, your expertise is to customize Search Engine IR systems to retrieve documents. Your consulting company accepted a new project from leading networking company, Cisco Systems. Cisco Systems defined internal document space by following key terms: Software, IoE, Edge,

Network, and Hardware. **V = { Software, IoE, Edge, Network, Hardware}** And Set of documents defined as: (**5**)

|  |
| --- |
| A1 = (1,0,0,1,1) |
| A2 = (0,1,1,1,1) |
| A3 = (1,1,1,0,0) |
| A4 = (0,0,1,1,1) |
| A5 = (1,0,0,1,0) |
| A6 = (0,0,0,1,1) |
| A7 = (0,1,0,1,1) |
| A8 = (0,1,1,0,0) |
| A9 = (1,1,0,0,1) |

* 1. If the query is “IOE and Edge” what documents should be retrieved? (Hint: Use Boolean Query

Matching)

* 1. If the Query is “Software, IoE, Edge) what documents should be retrieved? (Hint: Use Vector

Space Matching)

* 1. If the query is “Edge and Network”, what documents should be retrieved? (Either Boolean or Vector Space Matching)

**Answer:**

Retrieving Documents for Cisco Systems:

a) Query: "IOE and Edge" (Boolean Query Matching):

1. Identify relevant terms: "IOE" and "Edge" are both present in the vocabulary V.
2. Construct the Boolean query: IOE AND Edge
3. Match against documents: Apply the query to each document vector. A document is retrieved if it satisfies the query (both terms present with value 1).

* Retrieved documents: A2, A3, A7

b) Query: "Software, IoE, Edge" (Vector Space Matching):

1. Represent the query as a vector: Q = (1, 1, 1, 0, 0).
2. Calculate cosine similarity between the query and each document: Cosine similarity measures the angle between vectors, with higher values indicating closer resemblance.
3. Set a threshold: Choose a threshold for relevance (e.g., 0.5). Documents with cosine similarity above the threshold are retrieved.

* Retrieved documents (depending on the chosen threshold): A2, A3, A9 (possible, depending on threshold).

c) Query: "Edge and Network" (Either Boolean or Vector Space Matching):

Boolean Matching:

1. Construct the Boolean query: Edge AND Network
2. Apply the query as in part a).

* Retrieved documents: A2, A4, A7

Vector Space Matching:

1. Represent the query as a vector: Q = (0, 0, 1, 1, 1).
2. Calculate cosine similarity as in part b).
3. Set a threshold and retrieve relevant documents.

* Retrieved documents (depending on the chosen threshold): A2, A4, A7, A9 (possible, depending on threshold).

Comparison:

* Boolean matching is simpler to understand but can be too restrictive (missing documents with one term but high similarity).
* Vector space matching is more flexible and captures document relevance based on overall similarity.

The choice between Boolean and Vector Space Matching depends on the specific needs and desired level of precision vs. recall.

1. Explain Page Rank Algorithm in detail? What are the assumptions? (**5)**

Calculate Page Rank in following cases (assume damping factor 1)?

|  |  |
| --- | --- |
| Page Rank at A: - assume small universe of five pages – A, B, C, D, and E.  (**3**) |  |
| Page Rank at F: - Assume small universe of five pages – A, B, C, D, E and F.  (**3**) |  |

|  |  |
| --- | --- |
| Page Rank at F: - Assume small universe of five pages – A, B, C, D, E, F, G, and H. (**3**) |  |

1. Explain Mathematical Optimization & Solver Technologies? What are the three major steps? (5)

**Answer:**

Mathematical Optimization & Solver Technologies: Optimizing Decisions with Math

Mathematical optimization is a powerful technique for finding the best solution to a problem, given a set of constraints and objectives. It uses mathematical models and algorithms to systematically search for the option that maximizes benefits or minimizes costs. Solver technologies are software tools that implement these algorithms to find optimal solutions.

Here are the three major steps involved in using mathematical optimization and solver technologies:

1. Model Construction:

* Define the problem: This involves clearly identifying the decision variables (e.g., production quantities, delivery routes), constraints (e.g., budget limitations, resource availability), and objective function (e.g., maximize profit, minimize travel time).
* Formulate the model: Translate the problem into a mathematical format using equations and inequalities. This often involves linear, nonlinear, or mixed-integer programming depending on the problem complexity.
* Validate the model: Ensure the model accurately represents the real-world problem and produces meaningful solutions.

2. Solver Selection & Configuration:

* Choose the appropriate solver: Different solvers excel at specific types of problems (e.g., linear vs. nonlinear, small vs. large scale). Popular options include CPLEX, Gurobi, and SCIP.
* Configure the solver: Set parameters like accuracy tolerances, memory limitations, and solution quality requirements.
* Integrate the solver: Connect the solver with the model and any additional data sources used for optimization.

3. Solution Analysis & Implementation:

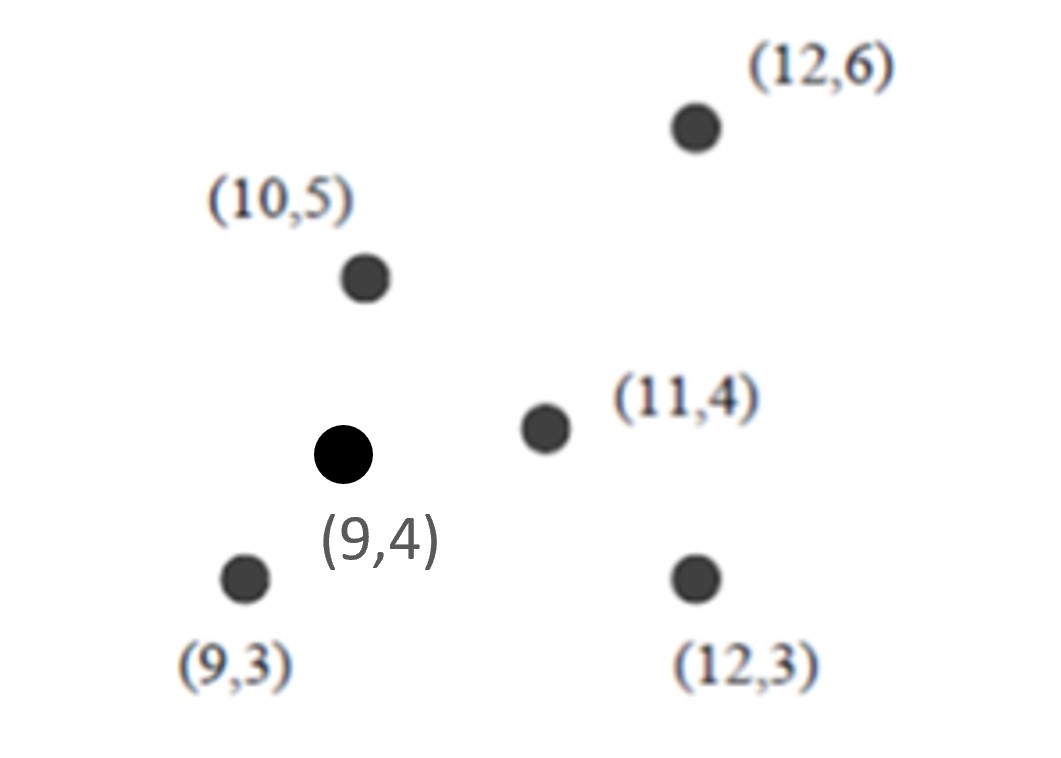
* Run the solver: Execute the optimization process to find the optimal solution.
* Analyse the results: Interpret the solution in terms of the original problem, evaluate its feasibility and optimality, and identify potential trade-offs.
* Implement the solution: Put the optimal decision variables into action, monitor the outcomes, and adjust the model as needed.

Benefits of using Mathematical Optimization & Solver Technologies:

* Improved decision-making: Provides data-driven insights for more informed and efficient choices.
* Cost and resource optimization: Helps minimize costs, maximize profits, and optimize resource allocation.
* Enhanced planning and forecasting: Facilitates better planning for future scenarios and potential disruptions.
* Automation and scalability: Enables automation of routine optimization tasks and scales efficiently to handle large and complex problems.

Overall, mathematical optimization and solver technologies offer a powerful tool for tackling complex decision-making problems in various fields. By following the three major steps and leveraging the right tools, you can unlock significant benefits in terms of efficiency, cost savings, and improved outcomes.

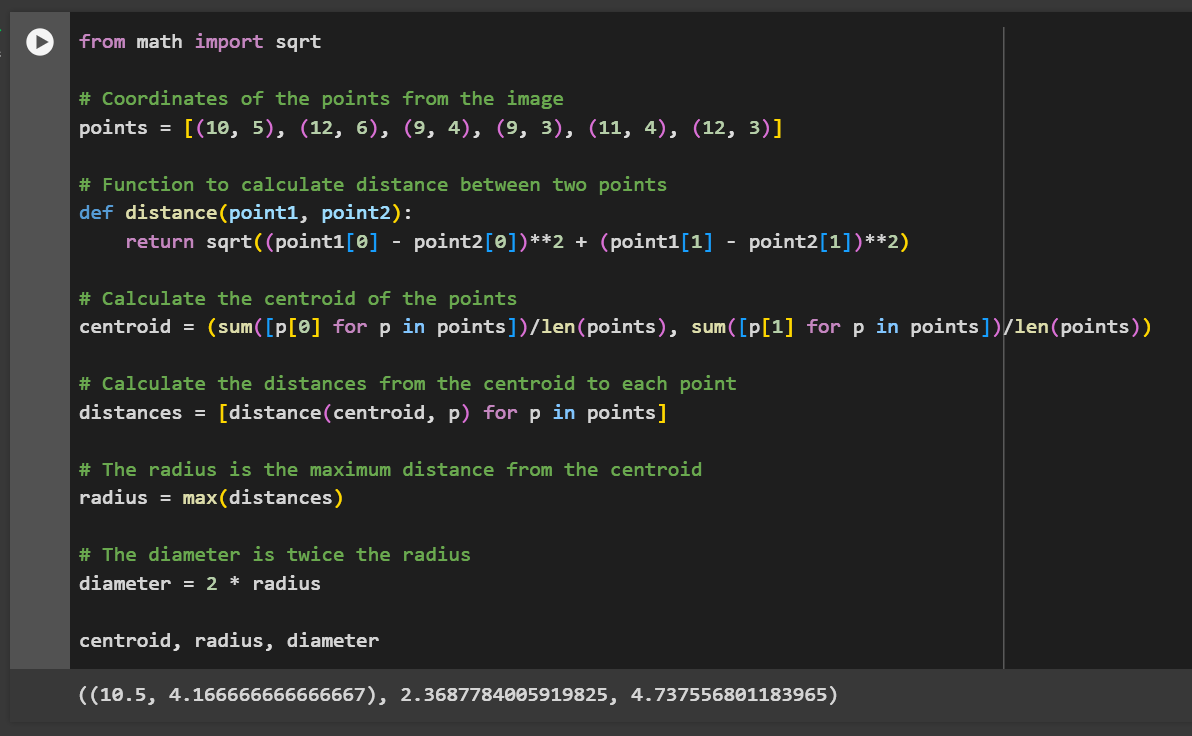
1. Compute Radius and Diameter? (4)



**Answer:** To do this, we would:

1. Compute the centroid or the mean point of all the given points, which serves as the center of the cluster.
2. Calculate the distance from the centroid to each point.
3. Find the maximum distance, which would be the radius.
4. Multiply the radius by 2 to get the diameter.

Let's compute the centroid, then the radius and diameter based on the points in the image.



The centroid (average point) of the points on the scatter plot is approximately (10.5, 4.17). The radius of the circle that would encompass all the points is approximately 2.37 units, and the diameter of that circle would be approximately 4.74 units.