	b)	Discuss various approaches to document sentiment classification. What are the limitations of document-level analysis?	L2	CO4	5 M
UNIT-V					
10	a)	Discuss the discovery and analysis of web usage patterns. How can these patterns improve website performance?	L2	CO3	5 M
	b)	Explain collaborative filtering using the K-Nearest Neighbor (KNN) method. How does it generate recommendations?	L2	CO4	5 M
OR					
11	a)	Explain the process of data modeling for web usage mining. What types of data are used in this process?	L2	CO3	5 M
	b)	Explain matrix factorization in collaborative filtering. Why is it considered effective for large-scale recommendation problems?	L2	CO4	5 M

Code: 23IT6501

III B.Tech - I Semester - Honors Examinations - NOVEMBER 2025

SOCIAL MEDIA ANALYTICS
(HONORS in INFORMATION TECHNOLOGY)

Duration: 3 hours Max. Marks: 70

Note: 1. This question paper contains two Parts A and B.
2. Part-A contains 10 short answer questions. Each Question carries 2 Marks.
3. Part-B contains 5 essay questions with an internal choice from each unit. Each Question carries 10 marks.
4. All parts of Question paper must be answered in one place.

BL – Blooms Level CO – Course Outcome

PART – A

		BL	CO
1.a)	Define the term "social network landscape." Why is it important in analytics?	L1	CO1
1.b)	State any two basic concepts of Information Retrieval (IR).	L1	CO1
1.c)	Outline stemming. How does it help in text pre-processing?	L2	CO2
1.d)	Define stop word removal and explain its role in text mining.	L1	CO2
1.e)	Differentiate between focused crawlers and universal crawlers.	L2	CO3
1.f)	Outline crawler ethics. Give one example of an ethical consideration.	L2	CO3
1.g)	Explain the main objective of opinion mining.	L2	CO3
1.h)	Outline any two techniques used to detect opinion spam.	L2	CO4
1.i)	Explain about content-based recommendation. Provide a simple example.	L2	CO4
1.j)	Outline web usage mining. Briefly explain its importance.	L2	CO4

PART – B

			BL	CO	Max. Marks
		UNIT-I			
2	a)	Discuss the importance of analytics in social media. How do businesses use it to drive decision-making and engagement?	L2	CO1	5 M
	b)	Explain the process of web mining. How does it differ from traditional data mining and what are its main components?	L2	CO1	5 M
		OR			
3	a)	Explain the future of social media analytics. Discuss upcoming trends and technologies that are shaping the field.	L2	CO1	5 M
	b)	Explain the analytics process used in social media analysis. Illustrate with an example.	L2	CO1	5 M
		UNIT-II			
4	a)	Explain the HITS algorithm. How does it differentiate between hubs and authorities in a network?	L2	CO2	5 M
	b)	Describe the process and significance of web page pre-processing. Include methods for handling noise and detecting duplicates.	L2	CO2	5 M
		OR			
5	a)	Summarize the relationship of the HITS algorithm with co-citation and bibliographic coupling. How are these concepts interconnected?	L2	CO2	5 M
	b)	Discuss the concept of duplicate detection in web mining. What are some common techniques used for identifying duplicate content?	L2	CO2	5 M
		UNIT-III			
6	a)	Outline the performance of a web crawler evaluated. Discuss the metrics used for evaluation.	L2	CO3	5 M
	b)	Summarize the ethical considerations in web crawling. What are the potential conflicts with website owners and users?	L2	CO3	5 M
		OR			
7	a)	Discuss the differences between universal, focused and topical crawlers. Provide examples and use cases for each.	L2	CO3	5 M
	b)	Discuss recent developments in web crawling technology. How are modern crawlers adapting to dynamic and large-scale web environments?	L2	CO3	5 M
		UNIT-IV			
8	a)	Differentiate between regular sentiment classification and comparative opinion mining. Provide examples.	L4	CO4	5 M
	b)	Explain the problem of opinion mining. Why is it challenging to extract sentiments from text?	L2	CO4	5 M
		OR			
9	a)	Explain how opinion search and retrieval systems work. What are the challenges in retrieving sentiment-rich content?	L2	CO4	5 M

Code: 23IT6501

PVP23

IIIB.Tech–I Semester–Honors Examinations-NOVEMBER2025

SOCIAL MEDIA ANALYTICS

(HONORS in INFORMATION TECHNOLOGY)

Scheme of Valuation

Duration: 3 Hours **Max. Marks: 70**

PART-A

1.a)Define the term "social network landscape." Why is it important in analytics? **2M**

Definition of Social Network Landscape – 1M

Importance in Analytics – 1M

1.b)State any two basic concepts of Information Retrieval (IR). **2M**

Any two Basic concepts of Information Retrieval (IR) – 1M + 1M = 2M

1.c)Outline stemming. How does it help in text pre-processing? **2M**

Stemming – 1M

Text pre-processing – 1M

1.d)Define stop word removal and explain its role in text mining. **2M**

Definition of Stop Word Removal – 1M

Role in text mining – 1M

1.e)Differentiate between focused crawlers and universal crawlers. **2M**

Any two between focused crawlers and universal crawlers – 2M

1.f)Outline crawler ethics. Give one example of an ethical consideration. **2M**

Crawler Ethics – 1M

Example – 1M

1.g)Explain the main objective of opinion mining. **2M**

Objective of Opinion Mining – 2M

1.h)Outline any two techniques used to detect opinion spam. **2M**

Any two techniques to detect opinion spam – 1M + 1M = 2M

1.i)Explain about content-based recommendation. Provide a simple example. **2M**

Content-based recommendation – 1M

Example – 1M

1.j)Outline web usage mining. Briefly explain its importance. **2M**

Web Usage Mining – 1M

Importance – 1M

PART-B

UNIT – I

2 a) Discuss the importance of analytics in social media. How do businesses use it to drive decision-making and engagement? 5M

Importance of analytics in social media – 3M

Used in decision-making and engagement – 2M

2 b) Explain the process of web mining. How does it differ from traditional data mining and what are its main components? 5M

Process of web mining – 2M

Differ from traditional Mining – 2M

Main components – 1M

OR

3 a) Explain the future of social media analytics. Discuss upcoming trends and technologies that are shaping the field. 5M

Future of Social Media Analytics – 3M

Upcoming trends and technologies that are shaping the field – 2M

3 b) Explain the analytics process used in social media analysis. Illustrate with an example. 5M

Analytics process used in social media analysis – 3M

Example – 2M

UNIT – II

4 a) Explain the HITS algorithm. How does it differentiate between hubs and authorities in a network? 5M

HITS algorithm–3M

differentiate between hubs and authorities in a network – 2M

4 b) Describe the process and significance of web page pre-processing. Include methods for handling noise and detecting duplicates. 5M

process and significance of web page pre-processing – 3M

methods for handling noise and detecting duplicates – 2M

OR

5 a) Summarize the relationship of the HITS algorithm with and bibliographic coupling. How are these concepts interconnected? 5M

relationship of the HITS algorithm with and bibliographic coupling – 5M

5 b) Discuss the concept of duplicate detection in web mining. What are some common techniques used for identifying duplicate content? 5M

duplicate detection in web mining – 3M

techniques used for identifying duplicate content – 2M

UNIT –III

6 a) Outline the performance of a web crawler evaluated. Discuss the metrics used for evaluation.5M

performance of a web crawler – 3M

metrics used for evaluation – 2M

6 b) Summarize the ethical considerations in web crawling. What are the potential conflicts with website owners and users? 5M

ethical considerations in web crawling – 3M

potential conflicts with website owners and users – 2M

OR

7 a)Discuss the differences between universal, focused and topical crawlers Provide examples and use cases for each. 5M

differences between universal, focused and topical crawlers – 3M

Examples and usecases for each – 2M

7 b) Discuss recent developments in web crawling technology. How are modern crawlers adapting to dynamic and large-scale web environments? 5M

developments in web crawling technology – 3M

How are modern crawlers adapting to dynamic and large-scale web environments – 2M

UNIT-IV

8 a) Differentiate between regular sentiment classification and comparative opinion mining. Provide examples.
5M

Differentiate between regular sentiment classification and comparative opinion mining – 4M
Examples – 1M

8 b) Explain the problem of opinion mining. Why is it challenging to extract sentiments from text? 5M

problem of opinion mining – 3M
opinion mining challenges to extract sentiments from text – 2M

OR

9 a) Explain how opinion search and retrieval systems work. What are the challenges in retrieving sentiment-rich content? 5 M

how opinion search and retrieval systems work – 3M
challenges in retrieving sentiment-rich content – 2M

9 b) Discuss various approaches to document sentiment classification. What are the limitations of document-level analysis? 5M

approaches to document sentiment classification – 3M
limitations of document-level analysis – 2M

UNIT-V

10 a) Discuss the discovery and analysis of web usage patterns. How can these patterns improve website performance? 5M

discovery and analysis of web usage patterns – 3M
patterns improve website performance – 2M

10 b) Explain collaborative filtering using the K-Nearest Neighbour (KNN) method. How does it generate recommendations? 5M

collaborative filtering using the K-Nearest Neighbour (KNN) method – 4M
generate recommendations – 1M

OR

11 a) Explain the process of data modeling for web usage mining. What types of data are used in this process? 5M

process of data modeling for web usage mining – 3M
types of data are used in this process – 2M

11 b) Explain matrix factorization in collaborative filtering. Why is it considered effective for large-scale recommendation problems? 5M

matrix factorization in collaborative filtering – 3M
effective for large-scale recommendation problems – 2M

Code: 23IT6501

PVP23

IIIB.Tech–I Semester– Honors Examinations -NOVEMBER 2025

SOCIAL MEDIA ANALYTICS

(HONORS in INFORMATION TECHNOLOGY)

Solution Set

Duration: 3 Hours **Max. Marks: 70**

PART-A

1.a)Define the term "social network landscape." Why is it important in analytics? 2M

Social network landscape refers to the overall structure, patterns, and relationships present within a social network.It gives community structures and patterns of interaction, which support targeted marketing, recommendation systems, and behaviour prediction.

1.b)State any two basic concepts of Information Retrieval (IR). 2M

Two basic concepts of Information Retrieval (IR) are:
1) Indexing –It involves creating structures like inverted indexes to enable fast retrieval.
2) Relevance –IR systems aim to rank results by their relevance to improve search quality.

1.c)Outline stemming. How does it help in text pre-processing? 2M

Stemming is a text-processing technique that reduces words to their root or base form by removing prefixes and suffixes. For example, running, runner, and ran may all be reduced to the stem "run."
It minimizes vocabulary size, making text analysis, indexing, and retrieval more efficient and improving the performance of Information Retrieval tasks.

1.d)Define stop word removal and explain its role in text mining. 2M

Stop word removal is the process of eliminating commonly used words (such as the, is, of, and, a) that carry little meaningful information for analysis.
It decreases the size of the dataset, improving processing efficiency.

1.e)Differentiate between focused crawlers and universal crawlers. 2M

Focused Crawler	Universal Crawler
Designed to collect web pages related to a specific topic or theme.	Designed to collect all accessible pages on the web, regardless of topic.
Useful for specialized search engines or topic-specific data gathering.	Useful for general-purpose search engines like Google or Bing.

1.f)Outline crawler ethics. Give one example of an ethical consideration. 2M

Crawler ethics refers to the guidelines and best practices that web crawlers must follow to avoid harming websites, violating privacy, or misusing data while collecting information from the web.
Respecting the robots.txt file and only crawling permitted pages.

1.g)Explain the main objective of opinion mining. 2M

The main objective of opinion mining (also called sentiment analysis) is to identify, extract, and analyse people's opinions, emotions, attitudes, or sentiments expressed in text.

1.h)Outline any two techniques used to detect opinion spam. 2M

Two common techniques used to detect opinion spam are:

1. Behavioural Analysis: Suspicious behaviour can indicate fake or manipulated reviews.
2. Text-based Analysis: can detect patterns typical of fake reviews.

1.i)Explain about content-based recommendation. Provide a simple example. 2M

Content-based recommendation is a technique in which a system recommends items to a user based on the features or attributes of items the user has previously liked.

Example: If a user watches many action movies, the system identifies features like "action, adventure, fast-paced" and recommends other movies with similar attributes

1.j)Outline web usage mining. Briefly explain its importance. 2M

Web usage mining is the process of analysing user behaviour data collected from web logs, clickstreams, and browsing patterns to understand how users interact with websites.

Helps improve website structure and navigation by understanding user behaviour.

PART-B

UNIT – I

2 a) Discuss the importance of analytics in social media. How do businesses use it to drive decision-making and engagement? 5M

Social media analytics (SMA) refers to the process of gathering, measuring, analyzing, and interpreting data from social media platforms to support decision-making and business strategies.To understand how users interact with content, what drives engagement, and how social presence affects brand perception.

1. Improving Content Performance:By checking which posts get more likes, comments, or shares, businesses learn what type of content works best. This allows them to create more effective posts and grow their reach.
2. Increasing Engagement: Analytics helps identify the best times to post and the kind of content people enjoy. Businesses use this to interact better with customers, respond faster, and build stronger relationships.

2 b) Explain the process of web mining. How does it differ from traditional data mining and what are its main components? 5M

Process of Web Mining:Web mining is the process of discovering useful patterns, information, and insights from the World Wide Web. It involves collecting data from websites, analysing it, and converting it into meaningful knowledge for decision-making.

Web Mining Differs from Traditional Data Mining:

Web Mining	Traditional Data Mining
Extracts data from the web (web pages, logs, links, user behaviour)	Extracts data from structured databases or warehouses
Deals with unstructured and semi-structured data like text, images, HTML	Deals mostly with structured, clean, organised data

Main Components of Web Mining:

Web mining is divided into three main types:

1. Web Content Mining:Extracting information from the content of web pages such as text, images, videos, anddocuments.

Used for: sentiment analysis, search engines, text mining, etc.

2. Web Structure Mining:Analysing how web pages are linked to each other using hyperlinks. Used for: PageRank, influence detection, website organisation.

3. Web Usage Mining:Studying user behaviour by analysing server logs, clickstreams, browsing history, and user sessions.

Used for: recommendation systems, user profiling, personalisation.

OR

3 a) Explain the future of social media analytics. Discuss upcoming trends and technologies that are shaping the field. 5M

The future of social media analytics lies in AI-driven insights, real-time tracking, emotion detection, and cross-platform analysis, supported by emerging technologies and stronger privacy-focused systems.

1. AI-Driven Insights:Artificial intelligence will automate data analysis and provide deeper, faster understanding of user behaviour.
2. Real-Time Analytics:Businesses will increasingly use real-time dashboards to respond instantly to trends and customer actions.
3. Predictive Analytics:Machine learning models will forecast future customer interests, engagement patterns, and market shifts.
4. Sentiment and Emotion Detection:Advanced tools will analyse not just opinions but user emotions through text, voice, and video content.

3 b) Explain the analytics process used in social media analysis. Illustrate with an example. 5M

Analytics Process Used in Social Media Analysis

1. Data Collection:Social media data such as likes, comments, shares, hashtags, views, and user profiles are gathered from platforms (e.g., Facebook, Instagram, Twitter).

2. Data Cleaning and Preparation:Irrelevant data, spam, bots, and duplicate posts are removed, and the remaining content is organised for analysis.

3. Data Analysis & Pattern Discovery:Techniques like sentiment analysis, trend detection, engagement measurement, and keyword analysis are applied to identify patterns and insights.

4. Interpretation of Results:The discovered insights are evaluated to understand audience behaviour, content performance, and brand perception.

5. Reporting & Decision-Making:Results are presented using dashboards or charts, helping businesses plan better content, improve marketing strategies, and engage their audience.

UNIT – II

4 a) Explain the HITS algorithm. How does it differentiate between hubs and authorities in a network? 5M

HITS Algorithm (Hyperlink-Induced Topic Search)

The HITS algorithm is a link analysis algorithm used to rank web pages by identifying two types of important pages in a network: authorities and hubs. It was introduced by Jon Kleinberg.

HITS assigns each webpage two scores:

1. Authority Score:A page is an authority if many high-quality hub pages link to it.It represents how trustworthy or valuable the content is.

2. Hub Score:A page is a hub if it links to many high-quality authority pages.It represents how good the page is at directing users to useful information.

How HITS Differentiates Hubs and Authorities

- Authorities:Pages with valuable information.They receive links from many hub pages.Example: A medical research article linked by many health blogs.
- Hubs:Pages that act as resource lists.They give links to many authority pages.Example: A blog post listing top 10 trusted medical websites.

4 b) Describe the process and significance of web page pre-processing. Include methods for handling noise and detecting duplicates. **5M**

Web page pre-processing is the step of cleaning, filtering, and structuring raw web data before applying web mining or analysis. Since web pages contain ads, HTML tags, scripts, and repeated content, pre-processing ensures only meaningful and useful information is extracted.

Significance

- Improves data quality by removing irrelevant content.
- Reduces processing time for mining algorithms.

Process of Web Page Pre-processing

1. Data Cleaning
2. Content Extraction
3. Tokenization
4. Stop-word Removal
5. Stemming/Lemmatization
6. Feature Representation

Duplicate Detection Methods

1. Exact Matching
2. URL Normalization

OR

5 a) Summarize the relationship of the HITS algorithm with and bibliographic coupling. How are these concepts interconnected?

5M

HITS Algorithm:HITS (Hyperlink-Induced Topic Search) identify authorities (pages with valuable content) and hubs (pages linking to authorities) in a network.It relies on link structure, where a page's importance is determined by its connections.

Bibliographic Coupling:Bibliographic coupling measures similarity between two documents based on shared references; two papers are "coupled" if they cite the same documents.It also uses link information. but focuses on shared outgoing links rather than incoming links.

Interconnection

- Both methods analyse network/graph structures to discover relationships and importance.
- HITS uses the network to rank nodes (hubs/authorities), while bibliographic coupling uses it to measure similarity between nodes.

5 b) Discuss the concept of duplicate detection in web mining. What are some common techniques used for identifying duplicate content?

5M

Duplicate detection is the process of identifying web pages that contain exactly or nearly the same content.Duplicate pages can bias search engines, increase storage costs, and reduce the quality of mining results.Detecting duplicates ensures that only unique and relevant data is analysed.

Common Techniques for Identifying Duplicates

1. Exact Matching: Compares the full text of two web pages. Works for perfect duplicates, but fails for near-duplicates.
2. URL Normalization:Removes tracking parameters and variations in URLs.Helps detect same content accessed via multiple URLs.

UNIT –III

6 a) Outline the performance of a web crawler evaluated. Discuss the metrics used for evaluation.5M

The performance of a web crawler is evaluated to determine how efficiently and effectively it collects relevant web pages while minimizing redundant or irrelevant downloads. Evaluation ensures the crawler meets its goals, such as coverage, speed, and accuracy.

Common Metrics for Evaluating Web Crawlers

1. Coverage:Measures the percentage of relevant pages discovered by the crawler compared to all available relevant pages.High coverage indicates the crawler effectively explores the web.
2. Precision:Ratio of relevant pages retrieved to total pages retrieved.High precision means the crawler retrieves mostly useful content, reducing noise.
3. Recall:Ratio of relevant pages retrieved to total relevant pages available on the web.High recall indicates that the crawler finds most of the relevant content.
4. Efficiency / Speed:Assesses pages downloaded per unit time and resource usage. High efficiency indicates the crawler operates quickly with minimal bandwidth or memory usage.
5. Redundancy:Measures the number of duplicate pages downloaded.Low redundancy indicates effective handling of duplicate content.

6 b) Summarize the ethical considerations in web crawling. What are the potential conflicts with website owners and users? 5M

Ethical Considerations in Web Crawling:Crawlers should obey the rules set by website owners in the robots.txt file to avoid accessing restricted areas.Crawlers must avoid overloading servers by limiting request rates and scheduling accesses responsibly.Crawlers should not collect sensitive personal data without consent, respecting user privacy and legal regulations.Extracting content for redistribution or commercial purposes may violate copyright laws, so proper usage and attribution are necessary.Crawlers should be identified clearly and provide contact information, allowing website owners to manage or block them if needed.

Potential Conflicts

1. With Website Owners:Excessive crawling can slow down servers or consume bandwidth.Copying content may violate intellectual property rights.
2. With Users:Crawlers may collect personal or private information, raising privacy concerns.Automated data collection can affect user experience on interactive websites.

OR

7 a) Discuss the differences between universal, focused and topical crawlers Provide examples and use cases for each. 5M

Differences between universal, focused and topical crawlers

Feature	Universal Crawler	Focused Crawler	Topical Crawler
Scope	Entire web	Specific topic	Specific topic with prioritized links
Efficiency	Low	Medium	High
Technique	BFS / DFS crawl	Topic filters	Scoring, ML-based link selection

Feature	Universal Crawler	Focused Crawler	Topical Crawler
Example	Googlebot	Academic paper crawler	Climate change news crawler
Use Case	Search engines, web archives	Domain-specific search, datasets	Topical aggregators, trend analysis

7 b) Discuss recent developments in web crawling technology. How are modern crawlers adapting to dynamic and large-scale web environments? 5M

Recent Developments in Web Crawling

1. Dynamic Content Handling:Modern crawlers use headless browsersto crawl JavaScript based websites.
2. Focused Crawlers:ML-based crawlers predict link relevance to avoid irrelevant pages and focus on high-value content.
3. Topical Crawlers:This improves efficiency by avoiding irrelevant pages and focusing on high-value content.

UNIT-IV

8 a) Differentiate between regular sentiment classification and comparative opinion mining. Provide examples. 5M

1. Regular Sentiment Classification:Identifies the overall sentiment (positive, negative, or neutral) expressed in a text.Measures general opinion about a single entity, product, or topic.Customer feedback analysis, brand monitoring, social media sentiment tracking.

Example:Review: "The camera quality of this phone is excellent."Sentiment: Positive

2. Comparative Opinion Mining:Detects comparisons between two or more entities and identifies which is preferred or better. Measures relative opinion rather than absolute sentiment.Product comparison, competitive analysis, recommendation systems.

Example:Review: "The battery of Phone A lasts longer than Phone B."Comparative Sentiment: Phone A > Phone B

8 b) Explain the problem of opinion mining. Why is it challenging to extract sentiments from text? 5M

Opinion mining (or sentiment analysis) is the process of automatically identifying and extracting subjective information from text, such as attitudes, emotions, or opinions about products, services, events, or entities. The main problem is to determine whether a text expresses a positive, negative, or neutral sentiment and to identify the target entity of the opinion.

Extracting sentiments from text is challenging because natural language is often ambiguous, and words can have different meanings depending on context. Sarcasm and irony make it difficult to interpret literal words correctly, while sentiment can be context-dependent, varying across domains or situations. Opinions are sometimes implicit, expressed without clear sentiment words, and a single text may contain mixed sentiments, both positive and negative.

OR

9 a) Explain how opinion search and retrieval systems work. What are the challenges in retrieving sentiment-rich content? 5 M

Opinion search and retrieval systems aim to find sentiment-rich content (positive, negative, or neutral opinions) from large collections such as websites, reviews, or social media. They work by:

1. Crawling and collecting text from online sources.

2. Identifying opinion-bearing sentences.
3. Classifying the sentiment (positive/negative/neutral) and extracting aspects (e.g., "battery", "camera").
4. Ranking and presenting results based on relevance and sentiment strength so that users can easily see summarized opinions.

Challenges in Retrieving Sentiment-Rich Content

Same words can express different sentiments depending on context (e.g., "This phone is sick!").Systems struggle to detect sarcastic statements where literal meaning is opposite of intended meaning.Sentiment words differ across domains and dialects; models trained on one domain often fail in another.Some opinions are hidden ("The earlier model was better") or comparative, making them hard to detect and classify.

9 b) Discuss various approaches to document sentiment classification. What are the limitations of document-level analysis? **5M**

The overall sentiment of a document is calculated by summing the polarity scores of all sentiment words.Uses supervised algorithms for the model is trained on labelled data and predicts sentiment for new documents.Improves accuracy by integrating prior sentiment knowledge with learned features.

Limitations of Document-Level Sentiment Analysis

A single document may contain mixed sentiments about different aspects, making one overall label inaccurate.Positive and negative sentences may cancel each other out, leading to wrong classification.Document-level analysis cannot identify which features are liked or disliked (e.g., "camera good but battery bad").

UNIT-V

10 a) Discuss the discovery and analysis of web usage patterns. How can these patterns improve website performance?
5M

Web usage patterns refer to the behaviours and navigation paths followed by users when they browse a website.

Their discovery involves Web Usage Mining, which includes:

1. Data Collection:Usage data is collected from server logs, browser logs, cookies, and user sessions.This includes pages visited, time spent, clicks, IP address, session IDs, etc.
2. Data Preprocessing:Cleaning and filtering raw log data along with identifying users, sessions, and paths.Removing redundant or irrelevant records.
3. Pattern Discovery:frequently visited page combinations, grouping similar users or sessions, predicting user behaviour, common navigation sequences
4. Pattern Analysis:Validating, interpreting, and visualizing discovered patterns, visualization dashboards to understand user behaviour.

Web Usage Patterns Improve Website Performance by understanding common navigation paths. Patterns help recommend relevant pages, products, or content to users based on past behaviour.Frequently visited links can be moved to prominent positions. Broken paths or dead links can be identified and corrected.

10 b) Explain collaborative filtering using the K-Nearest Neighbour (KNN) method. How does it generate recommendations? **5M**

Collaborative filtering is a recommendation technique that predicts a user's interests by analysing similar users or similar items.Collaborative filtering using the K-Nearest Neighbour (KNN) method recommends items by finding users or items that are most similar to the target user. Items with the highest predicted ratings or items preferred by similar users are recommended.

KNN Generates Recommendations

- If user-based, the system recommends items liked by similar users.
  Example:If User A's neighbours all liked Movie X, then Movie X is recommended to User A.
- If item-based, the system finds items similar to what the user already likes.
  Example:If a user liked Item A and Item B is similar (based on neighbour items), recommend Item B.

OR

11 a) Explain the process of data modeling for web usage mining. What types of data are used in this process? **5M**

Data modelling for web usage mining involves organizing and structuring raw web log data into meaningful patterns that describe user behaviour. The process starts with collecting data from web server logs, browser logs, cookies, and user sessions. This data is then pre-processed through cleaning, user identification, session identification, and path completion to convert it into consistent user–session profiles.After preprocessing, the data is transformed into models such as transaction files, user profiles, or session matrices, which serve as input for mining techniques like clustering, association rules, and sequential pattern mining.

The types of data used in this processinclude usage data (clickstreams, pages visited, time spent), content data (information on the web pages themselves), and structure data (site topology and link structure), all of which help discover meaningful navigation and behavioural patterns.

11 b) Explain matrix factorization in collaborative filtering. Why is it considered effective for large-scale recommendation problems? **5M**

Matrix factorization in collaborative filtering decomposes a large user–item rating matrix into two smaller matrices:

- Latent features of users and
- Latent features of items.

By multiplying these two matrices, the system predicts missing ratings based on how strongly a user's hidden preferences match an item's hidden characteristics. This approach is highly effective for large-scale recommendation problems because it handles very sparse data efficiently.