	b)	Using the Shift-Or algorithm, demonstrate the matching process for the pattern "101" in the text "1001011".	L3	CO3	5 M
		OR			
11	a)	Analyze the bit-parallelism concept in the Shift-Or algorithm.	L4	CO4	5 M
	b)	Explain the significance of string searching algorithms in computer science and mention at least two real-world applications.	L2	CO1	5 M

PVP 23

Code: 23AM3501

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

INFORMATION RETRIEVAL SYSTEMS
(CSE - AIML)

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)	Write short note on digital libraries.	L2	CO1
1.b)	Differentiate between data retrieval and information retrieval.	L2	CO1
1.c)	Mention one advantage and one limitation of signature files.	L2	CO1
1.d)	What is an inverted file in information retrieval system?	L2	CO1
1.e)	Define stoplist with an example.	L2	CO1
1.f)	Explain position based and context aware index.	L2	CO1
1.g)	How does stemming help in compressing inverted files?	L2	CO1
1.h)	List any two features of a thesaurus.	L2	CO1
1.i)	Define prefix function in KMP algorithm.	L2	CO1
1.j)	What is the principle of the shift-OR algorithm?	L2	CO1

PART – B

			BL	CO	Max. Marks
		UNIT-I			
2	a)	Explain in detail about the four major functional processes in information retrieval systems.	L2	CO1	5 M
	b)	Differentiate Data Base Management System and Information Retrieval System.	L2	CO1	5 M
		OR			
3	a)	Describe an inverted index and its role in storing and retrieving document-term information in IR systems.	L2	CO1	5 M
	b)	Explain IR system evaluation using precision, recall, F1 score and MAP with examples.	L2	CO1	5 M
		UNIT-II			
4	a)	Explain about Dictionary and Posting list in an inverted index.	L2	CO1	5 M
	b)	Illustrate the process of building an inverted file using a sorted array with an example dataset.	L3	CO2	5 M
		OR			
5	a)	Explain how signature file techniques can be modified to reduce false positives. Justify.	L4	CO4	5 M

	b)	Distinguish between vertical and horizontal partitioning in signature files with an example.	L2	CO1	5 M
		UNIT-III			
6	a)	Demonstrate how the N-gram data structure helps in handling spelling errors.	L2	CO1	5 M
	b)	Describe the structure and applications of hypertext data structures in IR.	L2	CO1	5 M
		OR			
7	a)	Explain different types of linkages used in indexing systems.	L2	CO1	5 M
	b)	Compare and Contrast manual indexing and automatic indexing.	L4	CO4	5 M
		UNIT-IV			
8	a)	Analyze the differences between over stemming and under stemming.	L4	CO4	5 M
	b)	Describe the effect of stemming on inverted file structure.	L2	CO1	5 M
		OR			
9	a)	Analyze the challenges of constructing a thesaurus from domain-specific texts.	L4	CO4	5 M
	b)	Compare and Contrast stemming and lemmatization with examples.	L4	CO4	5 M
		UNIT-V			
10	a)	Explain the concept of pattern, text, shift and match with example.	L2	CO1	5 M

Code: 23AM3501

PVP 23

III B.Tech. – I Sem- Regular Examinations
NOVEMBER 2025
CSE (AI & ML)
INFORMATION RETRIEVAL SYSTEMS

PART –A

SHORT SCHEME

Q.NO	PART - A	CO	Blooms Level	Marks
1.a)	Write Short note on digital libraries. Short description on digital libraries – 2M	CO1	L2	2
1.b)	Differentiate between data retrieval and information retrieval Any Two differences – 2M	CO1	L2	2
1.c)	Mention one advantage and one limitation of signature files. One advantage and one limitation – 2M	CO1	L2	2
1.d)	What is an inverted file in information retrieval system? Definition of inverted file – 2M	CO1	L2	2
1.e)	Define Stoplist with an example. Stoplist definition – 2M	CO1	L2	2
1.f)	Explain position based and concept aware index. Explanation – 2M	CO1	L2	2
1.g)	How does stemming help in compressing inverted files? Explanation – 2M	CO1	L2	2
1.h)	List any two features of a thesaurus. Any two features – 2M	CO1	L2	2
1.i)	Define Prefix function of a thesaurus. Explanation – 2M	CO1	L2	2
1.j)	What is the principle of the Shift – OR algorithm? Explanation – 2M	CO1	L2	2

PART –B
SHORT SCHEME

Q.No		Questions	BL	CO	Max Marks
UNIT I					
2	a)	Explain in detail about the four major functional processes in Information Retrieval Systems. Explanation of any four functional processes in IRS – 5M	L2	CO1	5M
	b)	Differentiate Data Base Management System and Information Retrieval System. Any Three Differences – 5M	L2	CO1	5M
OR					
3	a)	Describe an inverted index and its role in storing and retrieving document-term information in IR systems. Explanation – 5M	L2	CO1	5M
	b)	Explain IR system evaluation using precision, recall, F1 score and MAP with examples. Explanation – 5M	L2	CO1	5M
UNIT II					
4	a)	Explain about Dictionary and Posting list in an inverted index. Explanation – 5M	L2	CO1	5M
	b)	Illustrate the process of building an inverted file using a sorted array with an example dataset Explanation – 4M Any example of their choice – 1M	L3	CO2	5M
OR					
5	a)	Explain how signature file techniques can be modified to reduce false positives. Justify. Explanation – 5M	L4	CO4	5M
	b)	Distinguish between vertical and horizontal partitioning in signature files with an example. Any Three differences – 5M	L2	CO1	5M
UNIT-III					

6	a)	Demonstrate how the N-gram data structure helps in handling spelling errors Explanation – 5M	L2	CO1	5M
	b)	Describe the structure and applications of hypertext data structures in IR. Structure and applications – 5M	L2	CO1	5M
		OR			
7	a)	Explain different types of linkages used in indexing systems. Any three types of linkages explanation – 5M	L2	CO1	5M
	b)	Compare and Contrast manual indexing and automatic indexing. Any Three differences – 5M	L3	CO2	5M
		UNIT-IV			
8	a)	Analyze the differences between over stemming and understemming. Any Three differences – 5M	L4	CO4	5M
	b)	Describe the effect of stemming on inverted file structure Explanation – 5M	L2	CO1	5M
		OR			
9	a)	Analyze the challenges of constructing a thesaurus from domain specific texts. Any two challenges explanation – 5M	L4	CO4	5M
	b)	Compare and Contrast stemming and lemmatization with examples. Any Three differences – 5M	L4	CO4	5M
		UNIT - V			
10	a)	Explain the concept of pattern, text, shift and match with example. Definitions – 5M	L2	CO1	5M
	b)	Using the Shift-Or algorithm demonstrate the matching process for the pattern "101" in the text "1001011". Demonstration of process – 3M Matching the pattern in the text – 2M	L3	CO3	5M
		OR			

11	a)	Analyze the bit-parallelism concept in the Shift-Or algorithm. Explanation – 5M	L4	CO4	5M
	b)	Explain the significance of string searching algorithms in string search algorithms in computer science and mention at least two real world applications. Explanation – 4M Examples – 1M	L2	CO1	5M

Code: 23AM3501

PVP 23

III B.Tech. – I Sem- Regular Examinations
NOVEMBER 2025
CSE (AI & ML)
INFORMATION RETRIEVAL SYSTEMS

PART – A
DETAILED SCHEME

Q.NO	PART - A	CO	Blooms Level	Marks
1.a)	Write Short note on digital libraries. Short description on digital libraries – 2M • Digital libraries like IEEE Xplore use Information Retrieval systems to help users find relevant academic papers	CO1	L2	2
1.b)	Differentiate between data retrieval and information retrieval Any Two differences – 2M • Data Retrieval – Exact Match • Information Retrieval – Relevant match, Natural Language	CO1	L2	2
1.c)	Mention one advantage and one limitation of signature files. One advantage and one limitation – 2M • Signature Files are easy to construct 0's and 1's, fixed length • Limitations : False Positives, mismatches	CO1	L2	2
1.d)	What is an inverted file in information retrieval system? Definition of inverted file – 2M • An inverted file is a data structure used in Information Retrieval Systems to map each term in the document collection to a list of documents	CO1	L2	2
1.e)	Define Stoplist with an example. Stoplist definition – 2M • Stoplists are lists of common words (stopwords) that are excluded from indexing because they are not useful • Ex: a,an,the,and,is,was	CO1	L2	2
1.f)	Explain position based and concept aware index. Explanation – 2M • Position based indexing indicates position of the term	CO1	L2	2

	• Context aware indexing indicate the context of terms.			
1.g)	How does stemming help in compressing inverted files? Explanation – 2M • Stemming is a technique used to reduce words to their root forms, which helps in compressing these inverted files	**CO1**	**L2**	**2**
1.h)	List any two features of a thesaurus. Any two features from the below – 2M • Hierarchical • Associative • Synonymy • Antonomy • Broader and narrower	**CO1**	**L2**	**2**
1.i)	Define Prefix function in KMP algorithm. Explanation – 2M • KMP uses prefix function which builds an LPS array for pattern matching	**CO1**	**L2**	**2**
1.j)	What is the principle of the Shift – OR algorithm? Explanation – 2M • Shift – OR principle is bit parallelism / bit wise operations 0's and 1's	**CO1**	**L2**	**2**

PART – B
DETAILED SCHEME

2. a) Explain in detail about the four major functional processes in Information Retrieval Systems. L2 – CO1 – 5M

Ans: Explanation of any four functional processes in IRS – 5M

- Query Submission and processing : Tokenization, Normalization, Stemming and Lemmatization
- Search with inverted index
- Ranking based on relevance
- Retrieving and displaying results

(OR)

- Item Normalization
- Selective Dissemination of Information
- Document based search
- Index based search

2.b) Differentiate Data Base Management System and Information Retrieval System. L2-CO1-5M

Ans: Any Three Differences – 5M

Information Retrieval (IR)	Database Management System (DBMS)
The input data is Unstructured or semi-structured text (documents, web pages)	Input data is Structured data (tables, rows, columns)
Query can be Keyword-based, natural language queries	SQL-based structured queries (SELECT, WHERE, JOIN, etc.)
The results are approximate answers, ranking-based results	Precise queries expecting exact matches
Uses full-text search, inverted index, similarity models (e.g., TF-IDF)	Uses indexes (B-trees, hash), relational algebra for searching
Ranked list of relevant documents or snippets	Exact data records or tuples that match the query
Evaluation metrics are Precision, Recall, F1-score, MAP, NDCG	Correctness, completeness, response time
No strict schema; may use flat files, XML, JSON, document stores	Schema-based storage in relational tables
Best for keyword-based access to large text corpora	Best for structured querying and transaction management
Examples are Search engines (Google, Bing), Lucene, ElasticSearch	MySQL, PostgreSQL, Oracle, MongoDB
Usecases are Web search, digital libraries, document filtering	Banking systems, inventory control, ERP, customer databases

OR

3. a) Describe an inverted index and its role in storing and retrieving document-term information in IR systems. L2-CO1-5M

Explanation – 5M

Ans: An Inverted Index is a core data structure used in Information Retrieval Systems to map terms (words) to the documents in which they appear. It allows for fast full-text searches and is the foundation of search engines.

It is called "inverted" because it inverts the typical relationship: instead of mapping documents to the terms they contain, it maps terms to the documents they occur in.

documents->terms,
terms → documents.

3.b) Explain IR system evaluation using precision, recall, F1 score and MAP with examples L2-CO1-5M

Explanation – 5M

Ans:
Precision: Precision is the proportion of retrieved documents that are relevant to the total no.of retrieved documents.

Recall: Recall measures the proportion of retrieved relevant items to the total no.of possible relevant items for the given query

F1 Score: F1 Score is the harmonic mean of precision and recall, giving a balanced measure. It's useful when both false positives and false negatives matter.

Mean Average Precision: MAP calculates the average precision for each query and then takes the mean across all queries.

UNIT – II

4. a) Explain about Dictionary and Posting list in an inverted index. L2-CO1-5M

Explanation – 5M

Ans: An inverted index consists of two main components:

1. **Dictionary**:

- A list of all unique terms appearing in the document collection.
- Often sorted alphabetically or hashed for quick access.

2. **Posting Lists**:

- For each term, a list of document IDs where the term appears.
- May also include term frequencies, positions, or weights.

4.b) Illustrate the process of building an inverted file using a sorted array with an example dataset

Explanation – 4M L3-CO2-5M

Any example of their choice – 1M

Ans: Step 1: Input the Document Collection

Step 2: Preprocess the Text

Step 3: Extract All (Term, Document ID) Pairs

Step 4: Sort the List by Term (Alphabetically)

Step 5: Build the Dictionary and Posting Lists

Any example of choice

OR

5.a) Explain how signature file techniques can be modified to reduce false positives. Justify. L4-CO4-5M

Explanation – 5M

Ans: A Signature File is a compact, fixed-length binary representation of a document that serves as a quick filtering mechanism in information retrieval. Instead of storing the actual text or a full index for every search, each document is represented by a bit string, known as its signature.

Techniques to reduce false positives:

1. **Use Multiple Hash Functions**

   This reduces the probability that unrelated terms produce the same bit pattern.

2. **Increase Signature Length**

   A longer bit string decreases collision chances.

3. **Compression & Error-Tolerant Techniques**

   Block signatures, superimposed coding, or compression techniques can reduce redundancy while balancing false positives.

5.b) Distinguish between vertical and horizontal partitioning in signature files with an example.

Any Three differences – 5M L2-CO1-5M

Ans:

Vertical Partitioning	Horizontal Partitioning
Divides the signature file **column-wise**, i.e., splits the bit strings by bit positions into multiple smaller files.	Divides the signature file **row-wise**, i.e., splits the file into separate chunks where each chunk contains complete signatures of a subset of documents.
Bits (columns) in the signature matrix.	**Signatures (rows)** in the signature matrix.
Reduces the number of bits to be checked at one time, thus lowering memory usage and possibly improving query matching speed.	Allows processing of only a subset of documents at a time, useful for large collections.
The bit string of each document is split into equal-length segments (columns), and each segment is stored in a separate subfile.	The signature file is divided into multiple smaller files, each containing complete signatures for a subset of documents.
If each signature has 12 bits, split into 3 partitions of 4 bits each → Store separately.	If you have 9 documents, split them into 3 partitions, each containing signatures for 3 documents.
- Less storage per subfile → faster access. - Can process in parallel per bit segment.	- Only a subset of signatures needs to be loaded into memory at a time. - Suitable for distributed processing.
Slightly more complex matching logic.	Still processes entire signature per document in a subset.

6.a) Demonstrate how the N-gram data structure helps in handling spelling errors L2-CO1-5M

Explanation – 5M

Ans: Handling Noisy Input Using gram Indexing

Problem: User types `"retreval"` instead of `"retrieval"` (typo).

Step 1: Generate trigrams for the input:

```
ret, etr, tre, rev, eva, val
```

Step 2: Compare input trigrams to document trigrams:

Original trigrams for `"retrieval"`:

```
ret, etr, tri, rie, iev, eva, val
```

Step 3: Matching:

Common trigrams between `"retrieval"` and `"retreval"`:

```
ret, etr, eva, val
```

Step 4: Ranking documents:

Documents with maximum **trigram overlap** are ranked higher.

This allows the search system to **retrieve the correct document** even with spelling mistakes

6.b) Describe the structure and applications of hypertext data structures in IR. L2-CO1-5M

Structure and applications – 5M

Ans:

Hypertext Data Structure: Hypertext is a non-linear structure where each item (node) can reference other items through embedded links. Node: Each separate item of information (text, image, audio, video). Link: A pointer from one node to another.Supports multi-format references, e.g., text referencing images or videos.

Applications: HTML Webpages

OR

7.a) Explain different types of linkages used in indexing systems. L2-CO1-5M

Any three types of linkages explanation – 5M

Ans: 1. Equivalence Linkage

Connects synonyms, variants, or equivalent terms.

2. **Hierarchical Linkages**

Shows Broader → Narrower relationship between the terms.

3. **Associative Linkages**

   Related but not identical terms.

4. **Pre-coordination in Indexing**

   Pre-coordination is when indexing terms are combined at the time of indexing, before storage.

5. **Post-coordination in Indexing (for comparison)**

   Terms are assigned individually and combined at search time.

6. Sequential Linkage

7.b) Compare and Contrast manual indexing and automatic indexing. L4-CO4-5M

Any Three differences – 5M

Ans:

Manual Indexing	Automatic Indexing
Done by human indexers (librarians, subject experts)	Computer algorithms (statistical, linguistic, or ML-based methods)
Humans read and analyze documents → assign subject headings, keywords, or descriptors	System extracts index terms automatically from text (e.g., tokenization, stop-word removal, stemming, TF-IDF weighting, embeddings).
Often uses controlled vocabularies or thesauri to ensure consistency.	May use free-text terms, statistical features, or ontology-based term mapping.
High accuracy humans understand context	Quality depends on algorithm, training data, and preprocessing
Can vary between different human indexers	Highly consistent once algorithm is fixed
Time-consuming, expensive, and not scalable for large document collections.	Fast, cheap, and scalable to millions of documents.
Librarian assigning "Polycystic Ovary Syndrome—Diagnosis" as a subject heading. - MeSH terms assigned by human curators in PubMed.	Search engines creating inverted indexes automatically. - Using TF-IDF or BM25 for keyword indexing. - Deep learning methods extracting embeddings for retrieval.
Best suited for small, specialized collections where precision is critical (e.g., medical, legal, archival libraries).	Large, dynamic collections (e.g., web search engines, digital libraries).

UNIT – IV

8.a) Analyze the differences between over stemming and understemming. L4-CO4-5M

Any Three differences – 5M

Ans:

Over-stemming

1. Different words are reduced to the same stem incorrectly.
2. Increases recall but decreases precision.
3. May retrieve irrelevant results (e.g., *universe* → *university*).

Under-stemming

1. Variants of the same word are not reduced to one stem.
2. Decreases recall but increases precision.
3. May miss relevant results (e.g., *connect* ≠ *connecting*).

8.b) Describe the effect of stemming on inverted file structure L2-CO1-5M

Explanation – 5M

Ans: Stemming means reducing words to their root form (stem).
Example:

connects, *connected*, *connecting* → connect

In Information Retrieval (IR) systems, we build an inverted index — a list that maps terms → documents.

Without stemming → each word form is treated separately.

With stemming → all word forms share one stem

OR

9.a) Analyze the challenges of constructing a thesaurus from domain specific texts. L4-CO4-5M

Any two challenges explanation – 5M

Ans:

Limited Words/controlled vocabulary – Not all important terms appear in the text.

Confusing Meanings – Some words have different meanings in different places.

Different Terms Used – People use many forms of the same idea.

New Terms Coming – Domain vocabulary changes quickly.

Rare Words – Important terms may appear very few times.

Hard to Link Words – Finding correct relationships between terms is difficult.

Needs Experts – Experts are required to check and confirm the thesaurus.

9.b) Compare and Contrast stemming and lemmatization with examples. L4-CO4-5M

Any Three differences – 5M

Ans:

Stemming	Lemmatization
Stemming is reducing the term into the most rooted form rule based	Dictionary-based reduction
May not be valid words (e.g., *comput* for *computing*)	Always valid dictionary words (e.g., *compute*)
No POS tagging	Parts of the speech(POS) tagging will be there
Accuracy is moderate/low	High Accuracy
Fast Processing of word/retreival	Slower processing
Minimal resources required	Requires POS tagger and lexical database
May reduce precision (due to over-stemming)	Increases precision (context-sensitive)
Increases recall (more aggressive matching)	Moderate recall improvement
Best suited for Large-scale IR systems where speed matters	Semantic applications like NLP and QA systems

UNIT – V

10.a) Explain the concept of pattern, text, shift and match with example. L2-CO1-5M

Definitions – 5M

Ans: Text: The main string in which the search is performed.

Pattern: The string/substring to be found in the text.

Shift: The movement of the pattern along the text when a mismatch occurs.

Match: A position in the text where the pattern exactly matches a substring.

10.b) Using the Shift-Or algorithm demonstrate the matching process for the pattern "101" in the text "1001011". L3-CO3-5M

Demonstration of process – 3M

Matching the pattern in the text – 2M

Ans: **Step by Step working process**

- Bitmask Table Creation
- Register Initialization(R)
- Text Processing by applying Rupdate rule in each step until end of the string/text

  Rupdate = ((R << 1) | 1) & B[text[i]]

- If the right most bit becomes 0 then match condition otherwise mismatch condition

Finding 101 in 1001011

Pattern matched from position [4] to position [6]

OR

11.a) Analyze the bit-parallelism concept in the Shift-Or algorithm. L4-CO4-5M

Explanation – 5M

Ans:

- The Shift-Or algorithm (also called Bit-Parallel String Matching) is an efficient pattern matching algorithm that uses bitwise operations to find occurrences of a pattern in a text.
- It is particularly fast for small patterns, as it leverages parallelism using bits.
- Useful in keyword searching, spell checking, and real-time text scanning.
- Uses bitwise OR and SHIFT operations to simulate parallel comparison of pattern characters.

11.b) Explain the significance of string searching algorithms in string search algorithms in computer science and mention at least two real world applications. L2-CO2-5M

Explanation – 5M

Ans:

String search algorithms: String searching (or pattern matching) algorithms are fundamental techniques in Information Retrieval (IR) and computer science for finding occurrences of a pattern string within a text string. Locate the starting indices of the pattern in the text efficiently.

- **Applications in IR:**
  - Keyword searching in documents or databases
  - Detecting phrases or entities in text
  - DNA sequence analysis and computational biology
  - Text editors