COLLEGE OF INNOVATION & TECHNOLOGY

Mid Term Exam

Winter Semester: 2025

Paper Information			
Course Title:	**Advanced Information Storage & Retrieval**	**Course Code:**	**CSC-530**
Exam Duration:		**Number of Pages:** (including cover page)	**8**
Exam Day/Date:	**Tuesday, February 25, 2025**	**Exam Time:**	**10:30 AM**

Student Details			
Student Name:	**TARAKA RAM DONEPUDI**	**Student ID:**	**80678315**
CRN:	**25869**	**Instructor:**	**Dr. MAKRAM SOUI**
Deadline of submission:	**Thursday, February 27, 2025**	**Time:**	**18:00 PM**

Exam Instructions

- Fill the above 'Student Details' section carefully and complete.
- Cheating is strictly prohibited and action will be taken.

Marking Scheme	
Question No.	**Actual Score / Max. Score**
Q1	**/20**
Q2	**/20**
Q3	**/60**
Total Score:	**/ 100**

Learning Outcomes Achievement		
Learning Outcome	**Question #**	**Student's Achievement**

Q1) Color with red ONE correct answer (letter) for each of the following MCQs in the table below.

1	2	3	4	5	6	7	8	9	10
a	a	a	a	a	a	a	a	a	a
b	b	b	b	b	b	b	b	b	b
c	c	c	c	c	c	c	c	c	c
d	d	d	d	d	d	d	d	d	d

1. Which of the following is an example of a conventional IR system?

   a. Lexis-Nexis
   b. IBM's QBIC
   c. Library catalog
   d. AskJeeves

2. What type of IR system searches by keywords?

   a. Conventional
   b. Text-based
   c. Question answering systems
   d. Multimedia

3. Which of the following is an example of a data retrieval system?

   a. Google Search
   b. Banking System
   c. Image Search Engine
   d. Voice Assistant

4. Which of the following is an example of stemming?

   a. Removing punctuation from a sentence
   b. Converting "computational" to "comput"
   c. Recognizing "machine learning" as a single phrase
   d. Removing words like "the" and "it"

5. What is the purpose of an inverted index? Pop
    a. To store entire documents in a compressed format
    b. To map keywords to the list of documents containing them
    c. To delete duplicate words from a document
    d. To encrypt text for secure storage

6. What is the purpose of stripping unwanted characters/markup in text preprocessing?
    a. To add more content to the text
    b. To remove irrelevant parts of the text like HTML tags, punctuation, and numbers
    c. To translate the text into multiple languages
    d. To replace all words with synonyms

7. What does tokenization involve in text preprocessing?
    a. Merging all words into a single string
    b. Breaking text into keywords based on whitespace
    c. Replacing words with their synonyms
    d. Encrypting text for security

8. If a term appears in only one document out of a collection of 10 documents, what will its IDF value be like?
    a. High
    b. Low
    c. Zero
    d. Negative
9. What happens when recall is low?
    a. Many relevant documents are missed
    b. Many irrelevant documents are retrieved
    c. More retrieved documents are relevant
    d. It improves precision automatically
10. Which of the following is NOT true about recall?
    a. It is calculated as the number of relevant documents retrieved divided by the total number of relevant documents
    b. It focuses on retrieving as many relevant documents as possible
    c. High recall always guarantees high precision
    d. A recall value of 1 means all relevant documents were retrieved

[*20 Marks*]

Q2) For each question, color True (T) or False (F) as appropriate.

1	AskJeeves is an example of a question-answering IR system	T	F
2.	IR is the process of finding material of an unstructured nature that satisfies an information need from within large collections.	T	F
3	Information Retrieval (IR) works with unstructured data like text, multimedia, and images	T	F
4	Data Retrieval (DR) designed for structured databases like relational databases.	T	F
5	F Measure measure of performance takes into account both recall and precision	T	F
6	Precision measures how well the system retrieves all relevant documents.	T	F
7.	Efficiency measures how well the system retrieves relevant results (e.g., ranking quality, precision, recall).	T	F
8.	Vector space models predict the probability that a document is relevant to a given query	T	F
9.	Searching is an offline process of organizing documents using keywords extracted from the collection	T	F
10.	During the browsing process, users enter a query into the system, which then searches the database to retrieve the most relevant matching documents.	T	F

[*20 Marks*]

Q3) **A)** Consider the following three documents

S_1: *Machine learning is a subset of artificial intelligence.*
S_2: *Deep learning is a type of machine learning.*
S_3: *Artificial intelligence includes machine learning and deep learning.*
S_4: *Deep learning is a type of artificial intelligence.*

1. Construct the Positional Inverted Index

Term	Postings (Document ID : Word Positions)
a	[1,4], [2,4], [4,4]
and	[3,6]
artificial	[1,7], [3,1], [4,7]
deep	[2,1], [3,7],[4,1]
includes	[3,3]
intelligence	[1,8], [3,2], [4,8]
is	[1,3], [2,3], [4,3]
learning	[1,2], [2,2], [2,8], [3,5], [3,8], [4,2]
machine	[1,1], [2,7], [3,4]
of	[1,6], [2,6], [4,6]
subset	[1,5]
type	[2,5], [4,5]

2. Suppose we have the following Query: "machine learning". determine the matching doFor the query "machine learning", we will look for documents where both "machine" and "learning" appear together.
Solution:
Document S_1:
"machine" at position 1
"learning" at position 2
Match found at positions (1, 2).
Document S_2:
"machine" at position 6
"learning" at position 7
Match found at positions (6, 7).

Document S_3:
"machine" at position 4
"learning" at position 5
Match found at positions (4, 5).
Document S_4:
"machine" is not present in S_4, so no match.

3. What is the advantage of using a Positional Inverted Index compared to a standard Inverted Index for phrase searching?

Solution:

The main advantage of a Positional Inverted Index over a standard Inverted Index for phrase searching is that it stores the positions of terms within each document. This allows for efficient searching of phrases where the terms must appear in a specific order and close together.
In a standard Inverted Index, we can only determine which documents contain the query terms, but we cannot tell if the terms are close enough to form a phrase.
A Positional Inverted Index allows us to check the relative positions of the terms, making it easier to find phrases (such as "machine learning") that appear in a specific order within documents.
Thus, the Positional Inverted Index enables us to efficiently search for exact phrases, improving both precision and search accuracy.

[*20 Marks*]

B) A document retrieval system was used to search for articles related to Artificial Intelligence (AI) in a research database. The results are as follows:

The database contains 150 documents related to AI.
The search retrieved 90 documents.
Out of the 90 retrieved documents, 50 were relevant to AI.

Tasks:

- Calculate the Precision Score for the search.
- Calculate the Recall Score for the search.
- Explain the significance of Precision and Recall in evaluating search effectiveness.
- If the search retrieved 10 additional relevant documents without adding irrelevant ones, how would the Precision and Recall scores change?

Solution:
Precision measures how many of the retrieved documents are actually relevant. It is calculated as:
Precision= Number of Relevant Records Retrieved/(Total Number of records retrieved)
Given:
Relevant documents retrieved = 50
Total retrieved documents = 90
Precision= 50/90=0.556= 55.56%
Recall measures how many of the total relevant documents in the database were retrieved. It is calculated as:
Recall = (Number of relevant records retrieved) / (Total number of relevant records in the database)
Relevant documents retrieved = 50
Total relevant documents in the database = 150
Recall= 50/150= 0.3333=33.33%
Precision focuses on the quality of retrieved results. A high precision score means that most retrieved documents are relevant, reducing the effort needed to filter out irrelevant ones.
Recall focuses on completeness, ensuring that as many relevant documents as possible are retrieved. A high recall score means fewer relevant documents are missed.
A high precision, low recall system retrieves only the most relevant documents but may miss many others.
A high recall, low precision system retrieves many documents (including irrelevant ones), increasing the burden on the user to filter through results.
The ideal system balances both precision and recall, depending on the application's needs.
If the search retrieves **10 more relevant documents** without adding irrelevant ones:
New relevant documents retrieved = 50 + 10 = 60
New total retrieved documents = 90 + 10 = 100
Total relevant documents in the database = 150

Updated Precision:

Precision= 60/100= 0.6 =60%

Updated Recall:

Recall= 60/150= 0.4= 40%

Conclusion:
Precision increased from 55.56% to 60% because all newly retrieved documents were relevant.
Recall increased from 33.33% to 40%, as more relevant documents were retrieved.
Since no irrelevant documents were added, the search effectiveness improved in both precision and recall.

[*20 Marks*]

~~~~~~
~~~~~~

C)

A spam email classifier is evaluated based on its performance. The classifier produces the following results when tested on an email dataset:

- **True Positives (TP) = 50** (Spam emails correctly identified as spam)
- **False Positives (FP) = 10** (Non-spam emails incorrectly identified as spam)
- **False Negatives (FN) = 20** (Spam emails incorrectly classified as non-spam)

a. Calculate Precision
b. Calculate Recall
c. Compute the F1-score
d. Interpret the result

✓ If the F1-score is high, what does it mean for the classifier's performance?
✓ If the F1-score is low, what should be improved?

Let's calculate the performance metrics for the spam email classifier:
a. Precision:
Precision = TP / (TP + FP)
Precision = 50 / (50 + 10) = 50 / 60 ≈ 0.8333 = 83.33%

b. Recall:
Recall = TP / (TP + FN)
Recall = 50 / (50 + 20) = 50 / 70 ≈ 0.7143=71.43%

c. F1-score:
F1-score = 2 * (Precision * Recall) / (Precision + Recall)
F1-score = 2 * (0.8333 * 0.7143) / (0.8333 + 0.7143) ≈ 0.7692 or 76.94%

d. Interpretation of the Result

F1-score = 76.94%, which indicates a good balance between precision and recall.
The classifier correctly identifies most spam emails (high recall of 71.43%), while also avoiding too many false positives (high precision of 83.33%).

If the F1-score is high, what does it mean for the classifier's performance?
A high F1-score means that the classifier is performing well in both precision (low false positives) and recall (low false negatives).
This ensures that spam emails are effectively caught while minimizing incorrect labeling of non-spam emails.
If the F1-score is low, what should be improved?
If the F1-score is low, either precision or recall (or both) are low.
If precision is low, it means too many false positives—the classifier is incorrectly marking normal emails as spam.
Solution: Improve spam detection rules to avoid classifying non-spam as spam.
If recall is low, it means too many false negatives—actual spam emails are missed.

Solution: Enhance spam detection to ensure more spam emails are correctly caught. Thus, improving the classifier depends on whether false positives or false negatives are a bigger concern in the application scenario.

[*20 Marks*]

~~~~~~

End of Examination… Good Luck!
~~~~~~