Practical 4 : Evaluation Metrics for IR Systems

4a) Calculate precision, recall, and F-measure for a given set of retrieval results.

# Function to calculate precision, recall, and F-measure

def evaluate\_ir\_system(retrieved, relevant):

"""

Calculate Precision, Recall, and F-measure for IR System

retrieved: Set of retrieved documents (as indices or identifiers)

relevant: Set of relevant documents (as indices or identifiers)

"""

# Calculate number of relevant retrieved documents

relevant\_retrieved = len(retrieved & relevant)

# Precision: Relevant Retrieved / Total Retrieved

precision = relevant\_retrieved / len(retrieved) if len(retrieved) > 0 else 0

# Recall: Relevant Retrieved / Total Relevant

recall = relevant\_retrieved / len(relevant) if len(relevant) > 0 else 0

# F-measure: Harmonic mean of Precision and Recall

f\_measure = (2 \* precision \* recall) / (precision + recall) if (precision + recall) > 0 else 0

return precision, recall, f\_measure

# Example Usage

if \_\_name\_\_ == "\_\_main\_\_":

# Example retrieved documents (can be document indices or identifiers)

retrieved\_docs = {1, 2, 3, 4, 5}

# Example relevant documents

relevant\_docs = {3, 4, 5, 6, 7}

# Calculate Precision, Recall, and F-measure

precision, recall, f\_measure = evaluate\_ir\_system(retrieved\_docs, relevant\_docs)

# Output the results

print(f"Precision: {precision:.4f}")

print(f"Recall: {recall:.4f}")

print(f"F-measure: {f\_measure:.4f}")

**Output :**

Precision: 0.6000

Recall: 0.6000

F-measure: 0.6000

**4b) Use an evaluation toolkit to measure average precision and other evaluation metrics.**

from sklearn.metrics import precision\_score, recall\_score, f1\_score, average\_precision\_score, roc\_auc\_score

# Example ground truth (true labels) and predictions

true\_labels = [1, 0, 1, 1, 0, 1, 0, 0, 1, 0]

predicted\_probs = [0.9, 0.1, 0.8, 0.7, 0.4, 0.9, 0.2, 0.3, 0.85, 0.05]

predicted\_labels = [1 if prob > 0.5 else 0 for prob in predicted\_probs]

# Calculate metrics

precision = precision\_score(true\_labels, predicted\_labels)

recall = recall\_score(true\_labels, predicted\_labels)

f1 = f1\_score(true\_labels, predicted\_labels)

average\_precision = average\_precision\_score(true\_labels, predicted\_probs)

roc\_auc = roc\_auc\_score(true\_labels, predicted\_probs)

# Display the results

print("Evaluation Metrics:")

print(f"Precision: {precision:.2f}")

print(f"Recall: {recall:.2f}")

print(f"F1-Score: {f1:.2f}")

print(f"Average Precision (AP): {average\_precision:.2f}")

print(f"ROC-AUC: {roc\_auc:.2f}")

Output :

Evaluation Metrics:

Precision: 1.00

Recall: 1.00

F1-Score: 1.00

Average Precision (AP): 1.00

ROC-AUC: 1.00