Task 1.1 – Evaluation (multiple choices)

Select the correct answer for the following questions.

1) **Scenario:** As a patent lawyer, you are tasked with conducting a prior art search to assess the novelty of a new invention before filing a patent application. In this context, would you prioritize precision or recall?

   A. Prioritize precision, as it ensures that only highly relevant prior art is considered, reducing the costs of reviewing irrelevant references.

   B. Prioritize recall, as it ensures that you identify as much relevant prior art as possible, reducing the risk of missing crucial references.

   C. Prioritize both precision and recall equally, as they are equally important in the patent search process.

   D. Prioritize neither precision nor recall, as they are not relevant considerations in a patent search.

2) **Scenario:** You are a movie enthusiast who wants to find great movie suggestions based on your preferences and past watch history. Would you prioritize precision or recall?

   A. Prioritize precision, as it ensures that the recommended movies closely match your specific preferences.

   B. Prioritize recall, as it ensures that you discover a wide variety of movie options, even if some may not perfectly align with your preferences.

   C. Prioritize both precision and recall equally, as they are equally important when seeking movie recommendations.

   D. Prioritize neither precision nor recall, as they are not relevant considerations when searching for movie suggestions.

3) **Scenario:** You are conducting research to find information about the current CEO of GreatBank, a globally leading financial institute In this context, would you prioritize precision or recall?

   A. Prioritize precision, as it ensures that you quickly find the information related to the current CEO of GreatBank.

   B. Prioritize recall, as it ensures that you retrieve a comprehensive list of potential sources mentioning the CEO of GreatBank.

   C. Prioritize both precision and recall equally, as they are equally important when seeking information about a CEO.

   D. Prioritize neither precision nor recall, as they are not relevant considerations when searching for information about a CEO.

Task 1.1 – Evaluation (multiple choices, continued)

Select the correct answer for the following questions.

4) **Scenario:** You are a cybersecurity analyst responsible for detecting and security threats. Your goal is to identify all potential security breaches. In this context, which would you prioritize: precision or recall, and why?

   A. Prioritize precision, as it ensures that the alerts you investigate are highly likely to be actual security breaches, reducing false alarms.

   B. Prioritize recall, as it ensures that you detect as many potential security breaches as possible, even if it means investigating some false alarms.

   C. Prioritize both precision and recall equally, as they are equally important in maintaining network security.

   D. Prioritize neither precision nor recall, as they are not relevant considerations in cybersecurity.

5) **Scenario:** You are developing a retrieval system for a large digital library that provides search results to researchers such that no critical information is missed. Which of the following statement fits best your intentions:

   A. Optimize the retrieval system to produce relevant documents at the top of result list, as precision is most important.

   B. Optimize the retrieval system to return all relevant documents, as researches must cite all relevant literature.

   C. Optimize for both precision and recall to find most relevant documents minimizing the overhead incurred by non-relevant documents.

   D. Optimize for precision, and return more documents to increase the recall (note: if you return all documents, you have maximum recall)

6) **Scenario:** You are developing a search forum for technical questions and answers (Q&A) for users when they encounter technical problems and need quick solutions. In this context, which metric do you optimize and why?

   A. You should prioritize precision, as it ensures that users receive the most accurate and relevant answers to their technical questions.

   B. You should prioritize recall, as it ensures that users do not miss any potentially helpful answers in the forum.

   C. You should prioritize both precision and recall equally to provide a balanced user experience.

   D. You should prioritize neither precision nor recall, as they both are not suitable for this scenario.

Task 1.1 – Evaluation (multiple choices, continued)

Select the correct answer for the following questions.

7) **Scenario:** You are preparing a text retrieval competition. You consider the pros and cons of sparse and dense assessments in retrieval benchmarks. You want to ensure a clear understanding of their impact on specific evaluation metrics. Which statement describes the relationships the best?

A. Sparse assessments are ideal for evaluating Precision, while dense assessments are essential for Mean Reciprocal Rank (MRR).

B. Dense assessments are crucial for assessing Mean Average Precision (MAP), whereas sparse assessments help evaluate Normalized Discounted Cumulative Gain (nDCG).

C. Sparse assessments are best suited for assessing Recall, while dense assessments are necessary to calculate Precision.

D. Dense assessments are advantageous for measuring Precision, while sparse assessments are required for computing Recall.

8) **Scenario:** You are a working on a text retrieval project and assessing the performance of a new retrieval algorithm. You want to understand the factors that influence the choice between micro and macro evaluation methods in this context. Which statement describes the selection the best?

A. The computational resources available for processing precision-recall pairs, with micro evaluation being much faster.

B. The number of relevant documents in the entire collection, which makes micro evaluation slow if numbers are very large.

C. The result size and importance of individual queries in your task, which micro evaluation is better able to factor in than macro evaluation.

D. The average precision values achieved by the retrieval method, which requires micro evaluation for low values and macro evaluation otherwise.

9) **Scenario:** You are tasked with evaluating two different search methods, A and B, for a web search engine. Your goal is to determine which method performs better and decide to use Mean Reciprocal Rank (MRR) as your evaluation metric. What is the key factor you should consider?

A. The number of queries with no relevant documents found by either method.

B. The total number of relevant documents found by both methods across all queries.

C. The average rank of the first relevant document for each method across all queries.

D. The speed of retrieval for both methods in milliseconds per query.

Task 1.1 – Evaluation (multiple choices, continued)

Select the correct answer for the following questions.

10) **Scenario:** You are working on a search engine project, and you are tasked with evaluating the performance of two different retrieval methods, Method X and Method Y. Which metric should you use to measure how well these methods maintain high precision as more relevant documents are found?

A. Precision at k (P@k)

B. R-precision

C. Average precision (AP)

D. System efficiency (E)

11) **Scenario:** You are working in an e-commerce company, and you are responsible for evaluating the performance of the search algorithm. Your company has recently implemented a graded relevance assessment system, where relevance values range from 0 to 3, with 0 being "not relevant" and 3 being "highly relevant." You want to measure the effectiveness of the search algorithm and consider both the graded relevance of documents and their ranking in the search results. Which metric is best suited for the task?

A. Mean Reciprocal Rank (MRR)

B. Precision at k (P@k)

C. Cumulative Gain (CG-k)

D. Discounted Cumulative Gain (DCG-k)

12) **Scenario:** You are a medical researcher working on a diagnostic test for a rare disease. The test is designed to identify individuals who may have the disease based on specific biomarkers. In the context of your medical research on the diagnostic test, which of the following best defines the term "false negatives"?

A. Patients who tested positive for the disease using the diagnostic test but do not actually have the disease.

B. Patients who tested negative for the disease using the diagnostic test and are confirmed to be disease-free.

C. Patients who tested positive for the disease using the diagnostic test and are later found to be disease-free upon further examination.

D. Patients who tested negative for the disease using the diagnostic test but actually have the disease.

Task 1.1 – Evaluation (multiple choices, continued)

Select the correct answer for the following questions.

13) **Scenario:** You are a security analyst testing the effectiveness of a biometric lock system on a smartphone. You want to evaluate the system's ability to correctly identify authorized. Which metric would be most relevant to measure the systems performance?

A. True Positive Rate (TPR)

B. True Negative Rate (TNR)

C. False Positive Rate (FPR)

D. False Negative Rate (FNR)

14) **Scenario:** You work as a wildlife researcher studying the behavior of a rare and elusive species of nocturnal owls. To monitor their activities, you've set up a camera trap system that captures images of these owls at night. Your goal is to use a machine learning classifier to automatically identify the owls in the images. However, due to the owls' nocturnal habits and their excellent camouflage, the classifier occasionally misclassifies them. You want to optimize the classifier's performance by adjusting the classification threshold. Which method can help you determine the optimal classification threshold?

A. Calculate the Mean Reciprocal Rank (MRR) and choose the threshold that maximizes it.

B. Use the Mean Absolute Error (MAE) to assess the classifier's performance at different thresholds.

C. Plot and analyze a Receiver Operating Characteristic (ROC) curve.

D. Calculate the F1-score for various threshold values and select the threshold with the highest F1-score.

15) **Scenario:** You are a data scientist working for a medical imaging company, developing an image classifier to detect early signs of a specific medical condition in X-ray images. Your goal is to evaluate the performance of your classifier and determine how well it can distinguish between positive and negative cases. What does the area under the ROC curve (AUC-ROC) represent when evaluating the image classifier's performance?

A. The accuracy of the classifier in correctly identifying all positive cases.

B. The overall classification accuracy of the classifier.

C. The balance between the classifier's true positive rate and false positive rate across different thresholds.

D. The precision of the classifier in correctly identifying all negative cases.

Task 1.2: Precision, Recall, and Reciprocal Rank (theoretical)

Two search engines, A and B, search the same collection and return the top 30 documents for a single query, ranked by relevance. The table below shows the rankings, using a '+' to indicate relevance and an empty cell for non-relevance. There are a total of 12 relevant documents in the collection for this query.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
A	+	+		+				+				+								+										+
B		+	+				+			+	+		+				+			+			+		+					

a) Draw the precision-recall graph for both engines.

b) Which engine performs better and why?

c) Let us assume we run 10 queries, each with 5 relevant documents and we obtain the following top-10 results for A and B. Again, '+' marks relevant documents. Which engine performs better (precision, recall, MRR)?

Q1		1	2	3	4	5	6	7	8	9	10
	A			+		+			+		
	B		+				+				+

Q2		1	2	3	4	5	6	7	8	9	10
	A				+						+
	B		+				+				

Q3		1	2	3	4	5	6	7	8	9	10
	A						+	+	+	+	
	B				+						+

Q4		1	2	3	4	5	6	7	8	9	10
	A		+	+				+			
	B	+	+	+							+

Q5		1	2	3	4	5	6	7	8	9	10
	A			+		+			+		
	B						+		+	+	

Q6		1	2	3	4	5	6	7	8	9	10
	A		+			+					+
	B	+		+					+		

Q7		1	2	3	4	5	6	7	8	9	10
	A					+			+	+	
	B			+				+			

Q8		1	2	3	4	5	6	7	8	9	10
	A			+						+	
	B				+	+	+		+		

Q9		1	2	3	4	5	6	7	8	9	10
	A		+	+		+		+			
	B				+	+	+				

Q10		1	2	3	4	5	6	7	8	9	10
	A		+		+				+		
	B	+	+	+				+			+

Task 1.3: Graded Relevance (theoretical)

We are assessing the performance of two systems, A and B, using graded relevance and cumulative gain metrics.

a) Suppose you have a set of five queries, and for each query, both System A and System B retrieve a list of documents with graded relevance assessments. Here's the data for the five queries:

• Query 1:	A = [3, 1, 2, 0 ,0]	B = [2, 3, 1, 0 ,0]
• Query 2:	A = [3, 0, 2, 2, 1]	B = [2, 0, 3, 1 ,1]
• Query 3:	A = [1, 0, 2, 2, 3]	B = [3, 0, 2, 1, 2]
• Query 4:	A = [1, 0, 1, 1, 1]	B = [3, 3, 0, 1, 1]
• Query 5:	A = [2, 3, 1, 3, 2]	B = [0, 0, 0, 3, 3]

Assuming each query contains 2 relevant documents with grade 3, and 3 relevant documents with grade 2, calculate the nDCG for each query and system. Then, average the nDCG scores for Systems A and B to determine which one performs better.

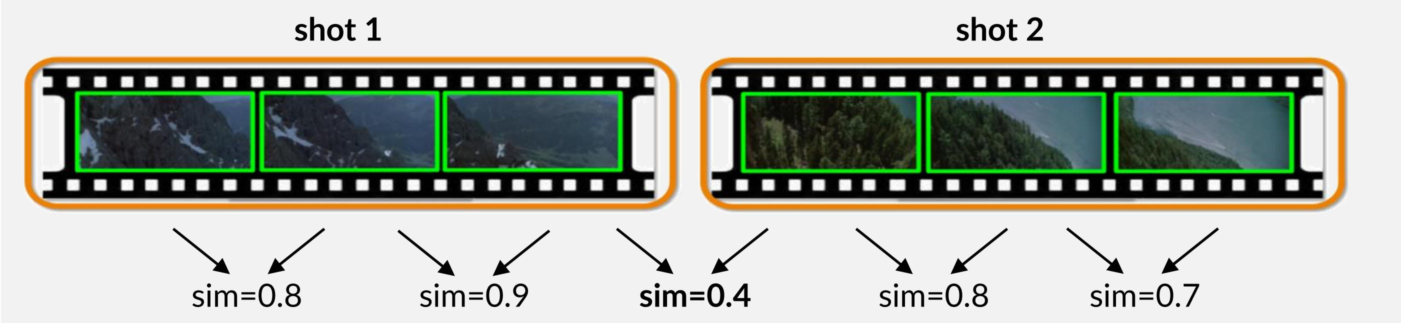
b) Consider the 10 queries on the previous page for Task 1.2.c. Both systems evaluated 10 queries and returned the top-10 answers. In this case, a '+' denotes a graded relevance of 1, and a blank cell denotes a non-relevant document (grade = 0). Compare both systems using the average nDCG.

c) We have introduced two DCG calculation variants. Apply both variants to the scenario in (b), varying the grades assigned to relevant documents (marked with '+') from 1 to 3. What observations can you make?

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \qquad \text{variant:} \quad DCG_k^{`} = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Task 1.4: Performance of Video Shot Detection (practical)

We developed software to segment videos into shots, which are sequences of frames that belong together. The software calculates frame-to-frame similarity and detects shot boundaries when the similarity falls below a set threshold, while frames with higher similarity are considered part of the same shot.



In the example above, we have two shots, each comprising 3 frames, with similarity values between consecutive frames. Notably, the similarity between frames at the shot boundaries is significantly lower than within the shots. By applying a threshold, such as 0.5, we can reliably identify the shot divisions.

Let's consider a practical example. You can download two text files on the course homepage, each representing the output of two different versions of our shot detection system. These files contain similarity values between consecutive frames, along with labels "shot" or "noshot" indicating the presence or absence of a shot boundary between the frames (these labels are used for system training). Preferably, use a Jupyter notebook (Python) and use markdown segments to document your answers and to explain your choices (minimal wording).

a) Begin by defining "positive" and "negative" for the actual condition observed in the text file as "shot" and "noshot." Similarly, establish what "True" and "False" mean for the predicted condition expressed as a predicate on similarity values between consecutive frames (and a chosen threshold).

b) Create a function to compute the confusion matrix's true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values for a given threshold. Next, calculate the sensitivity (True Positive Rate or TPR), specificity (True Negative Rate or TNR), and accuracy (ACC) from these values.

c) Compute the ROC curve and create a plot using various thresholds. Determine the range of thresholds that is required. Explore whether there is a more efficient method for calculating the ROC curve rather than iteratively invoking the function created in step b) for the entire threshold value range.

d) Determine the ideal threshold for shot detection. What criterion are you using to define "optimal"? Evaluate the performance of both methods using their respective optimal thresholds.

e) Define a function that calculates the area under the ROC curve and compute it for both methods. How does the area relate to the values calculated in the previous sub tasks? Can you implement an efficient way to compute the area?