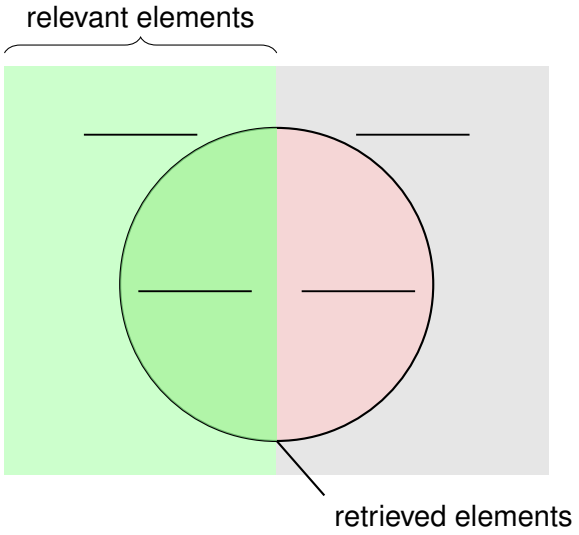
Lab Class IR:II

Exercise 1 : Effectiveness measures in Set Retrieval

The diagram below illustrates the relationship between “relevant” and “retrieved” elements and True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) in an information retrieval task.

(a) Label the four regions in the diagram with TP, FP, FN, and TN.

(b) Complete the formulas for precision and recall in the fields provided.



Precision = _____

Recall = _____

Exercise 2 : Measure Intuition (Set & Ranked Retrieval)

Describe the following measures, each in 1–3 sentences:

(a) Recall

(b) Query throughput

(c) Indexing space overhead

(d) Normalized Discounted Cumulative Gain (nDCG)

Exercise 3 : Measure Suitability (Set Retrieval)

Name two search tasks where the Precision measure (not Precision@k!) is good or bad at modelling the user behavior, respectively. Give a short explanation (1 sentence).:

(a) For which search task is it a good fit?

(b) For which search task is it a bad fit?

Exercise 4 : Effectiveness vs. Efficiency (Set & Ranked Retrieval)

Which of the following measures describe the effectiveness, and which describe the efficiency of a retrieval system? Give a short reason (up to 10 words)

(a) Query latency

(b) nDCG

(c) Recall

(d) Index size

Exercise 5 : Practical Limitations of Measuring Recall (Set Retrieval)

Recall measures how many of the relevant documents were successfully retrieved.

(a) Explain in your own words what happens when recall is very high.

(b) Is there a trivial way to maximize recall? What would be the consequence for precision?

(c) In practice, information retrieval systems often work with very large document collections. What difficulties arise when trying to measure recall in such settings?

Exercise 6 : Precision, Recall, and F-Measure (Set Retrieval)

Two information retrieval systems A and B, have been tested on the same topics. The number of relevant documents retrieved by each system is shown below:

System	Relevant Retrieved	Total Retrieved
A	8	10
B	5	6

Assume the total number of relevant documents in the collection is 12.

(a) Compute the precision p and recall r for each system.

(b) Determine which system is better in terms of precision p and which is better in terms of recall r.

(c) Compute the F-measure (i.e., harmonic mean $\frac{2 \cdot p \cdot r}{p+r}$ of precision p and recall r) for each system.

(d) Based on the F-score, determine which system performs better overall.

(e) Explain why the F-measure is a better single measure in this scenario.

Exercise 7 : Harmonic vs. Arithmetic and Geometric Means (Set Retrieval)

The F-measure induces a total order and is computed as the harmonic mean of precision p and recall r. Compare the harmonic, geometric, and arithmetic means of precision and recall. Explain why the harmonic mean (Figure 1) is often preferred over the geometric mean (Figure 2) or the arithmetic mean (Figure 3) in evaluating performance metrics. Discuss how each mean behaves when precision and recall differ significantly, and justify why the harmonic mean provides a more balanced evaluation in such cases.

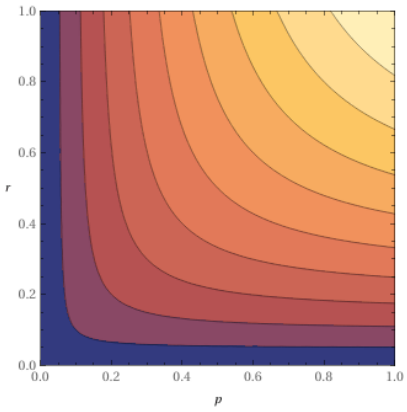


Figure 1: Harmonic mean $\frac{2 \cdot p \cdot r}{p+r}$ [plot]

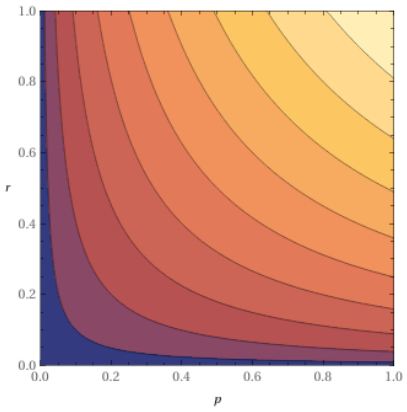


Figure 2: Geometric mean $\sqrt{p \cdot r}$ [plot]

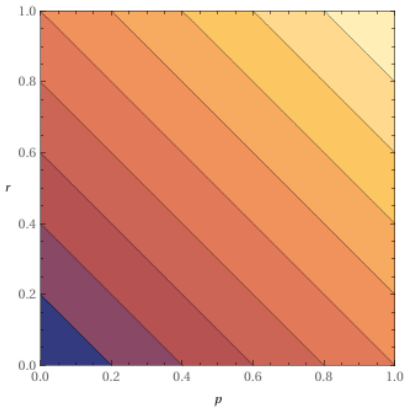


Figure 3: Arithmetic mean $\frac{p+r}{2}$ [plot]

Exercise 8 : Precision and Recall (Set Retrieval)

Consider a retrieval system that achieves a recall of r_1 and a precision of p_1 when retrieving k_1 documents. Suppose we retrieve more documents, $k_2 > k_1$, and compute the corresponding recall r_2 and precision p_2.

Determine whether the following statements are true or false:

(a) $r_2 \geq r_1$

(b) $p_2 \geq p_1$

Exercise 9 : Interactive Retrieval

Name three main differences between ad hoc retrieval and interactive retrieval.

Exercise 10 : Cranfield Paradigm

Which items are required for a laboratory experiment for ad hoc retrieval? This setup is also referred to as an experiment under the Cranfield Paradigm.

Exercise 11 : Experimental setup: Relevance judgements

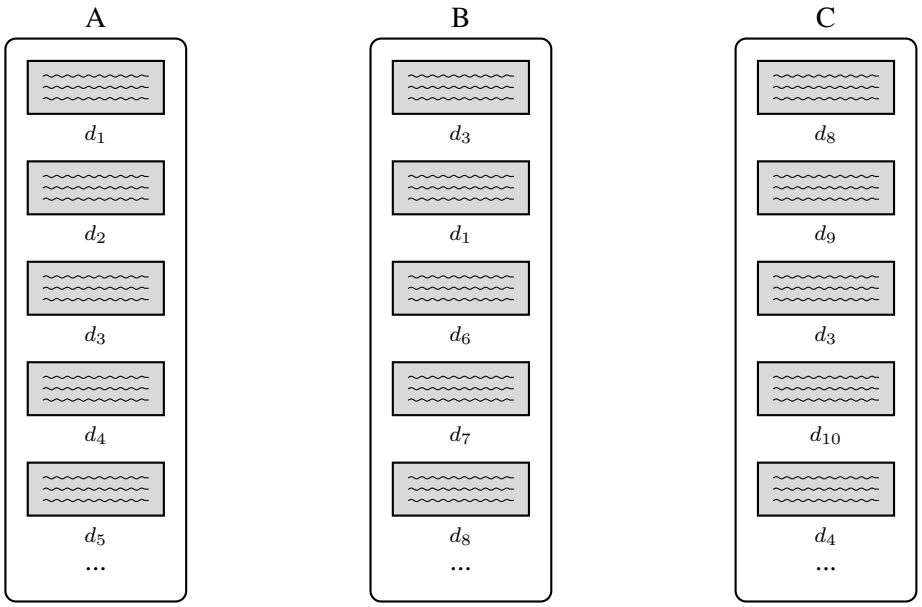
Which dimensions should be considered for carrying out manual assessment for relevance judgement.

Exercise 12 : Pooling

Given a corpus with documents denoted as d_i indexed by three retrieval systems A, B, and C, each producing a ranked list of documents for a query q (where the top is considered most relevant), answer the following:

(a) For $k = 3$ pooling depth, which documents are included in the resulting document "pool"?

(b) Explain the limitations of this pooling approach in the context of this example.



Exercise 13 : Kappa statistics

Given the judgements of two annotators A, B on a given topic, Scott's κ measures their agreement as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o denotes the proportion of agreement observed, and p_e the expected proportion of agreement by chance.

(a) Compute the Scott's κ measure for the binary relevance judgements given in Table 1.

(b) How would you interpret the results (using Table 2)?

Table 1: Annotators A and B have made the following $n = 600$ binary relevance judgements.

		B		$\sum$
		yes	no	
A	yes	456	44	500
	no	13	87	100
$\sum$		469	131	600

Table 2: Interpretations to measured kappa values (disputed).

κ	Agreement
< 0	poor
0.01 - 0.20	slight
0.21 - 0.40	fair
0.41 - 0.60	moderate
0.61 - 0.80	substantial
0.81 - 1.00	almost perfect