https://miro.com/app/board/o9J_lKiwpKk=/

Evaluation in Information Retrieval (Chapter 8)

Exercise 8/1

The following ordered list of 20 letters R and N represents **relevant (R)** and **non-relevant (N)** retrieved documents as an answer for a query on a collection of 10 000 documents.
The leftmost document is expected to be the most relevant. The list contains 6 relevant documents. Assume that the collection contains 8 documents relevant to the query.

IR(Q) -> RRNNNNNNRNRNNNRNNNNR

1 2 5 11 15 20

a) Precision?
   Recall?
b) F-measure?
e) MAP?
f) Best possible MAP?
g) Worst possible MAP?

Definition 1 (Recall)
Recall describes how many of the relevant documents are retrieved.

$$recall = R = \frac{\#relevant\ retrieved}{\#relevant}$$

Definition 2 (Precision)
Precision describes how many of the retrieved documents are relevant.

$$precision = P = \frac{\#relevant\ retrieved}{\#retrieved}$$

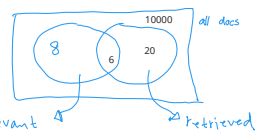
Definition 3 (F-measure)
A balanced F-measure (F_1-measure) defines a recall-precision relationship represented by their weighted harmonic mean:

$$F = \frac{2 \cdot R \cdot P}{R + P}$$

Definition 4 (Mean Average Precision)
MAP expresses the precision in each point a new relevant document is included in the result. It is counted as

$$MAP(Q) = \frac{1}{|Q|} \cdot \left(\sum_{q \in Q} \frac{1}{rel_q} \cdot \left(\sum_{i=1}^{rel_q} prec_i \right) \right)$$

where rel_q is the number of relevant documents for query q and $prec_i$ is the precision at the i-th relevant document.



Exercise 8/4

Below is a table showing how two judges judged the relevance (**0 = non-relevant, 1 = relevant**) of the set of 12 documents with respect to a query. Assume that you developed an IR system, that for this query returns the documents {4, 5, 6, 7, 8}.

IR(Q) -> {4, 5, 6, 7, 8}

a) Calculate the κ statistic.
b) Precision, if judges agree
c) Recall, if at least one judge thinks so

Definition 5 (κ statistic)
Let N be the total number of documents, J is a set of judges and $P(A)$ the number of documents on which the judges agree. Let also define R_j and NR_j be the number of relevant and non-relevant documents, respectively, according to the judge $j \in J$ and

$$P(R) = \frac{\sum_{j \in J} R_j}{|J| \cdot N} \quad \text{and} \quad P(NR) = \frac{\sum_{j \in J} NR_j}{|J| \cdot N}$$

as the number of relevant and non-relevant documents, respectively. Let finally define

$$P(E) = P(R)^2 + P(NR)^2$$

as the approximate number of disagreements between the judges. Then the κ statistic is defined as the measure of agreement between the judges

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}.$$

Doc ID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

Table 1: Judges judging the relevance of documents.