Task 1: Repetition — Precision & Recall

Consider an information need for which there are 5 relevant documents in the collection. Given is the following list of relevant (R) and non-relevant (N) documents returned for a query (the leftmost item is the top ranked search result):

R N R N N N N N R R

Task 1.1: Compute the precision, recall and F_1 score of the ranking.

$Precision = \frac{4}{10} = 0.4$
$Recall = \frac{4}{5} = 0.8$

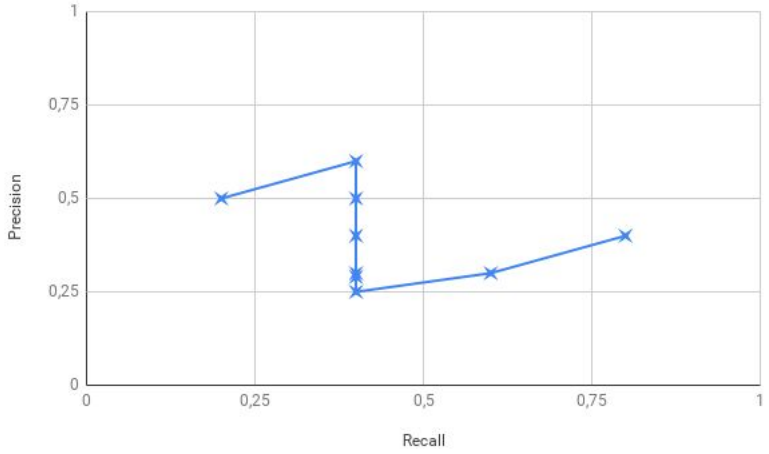
$F_1 = \frac{2PR}{P+R} = \frac{2 \cdot 0.4 \cdot 0.8}{0.4+0.8} = \frac{8}{15} = 0.5\overline{3}$

Task 1.2: Compute the precision@1, recall@1, precision@5 and recall@5.

$P@1 = \frac{1}{1} = 1\,,\ R@1 = \frac{1}{5} = 0.2$
$P@5 = \frac{2}{5} = 0.4\,,\ R@5 = \frac{2}{5} = 0.4$

Task 1.3: Create the precision/recall graph.

Rank	Recall	Precision
1	0.2	1.0
2	0.2	0.5
3	0.4	$0.\overline{6}$
4	0.4	0.5
5	0.4	0.4
6	0.4	$0.\overline{3}$
7	0.4	0.29
8	0.4	0.25
9	0.6	$0.\overline{3}$
10	0.8	0.4



Task 2: Repetition — Average Precision

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows:

System 1	R	N	R	N	N	N	N	N	R	R
System 2	N	R	N	N	R	R	R	N	N	N

Task 2.1: Which of the two system returns a better ranking according to Precision@8?

$$P@8(System\ 1) = \frac{2}{8} = 0.25$$
$$P@8(System\ 2) = \frac{4}{8} = 0.5$$

According to this measure, System 1 retrieves the better ranking.

Task 2.2: Which of the two system returns a better ranking according to their average precision?

relevant document	Precision (System 1)	Precision (System 2)
1	1	0.5
2	$\frac{2}{3}$	0.4
3	$\frac{1}{3}$	0.5
4	0.4	$\frac{4}{7}$

$$AP(System\ 1) = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot 0.4 = \frac{3}{5} = 0.6$$
$$AP(System\ 2) = \frac{1}{4} \cdot 0.5 + \frac{1}{4} \cdot 0.4 + \frac{1}{4} \cdot 0.5 + \frac{1}{4} \cdot \frac{4}{7} = \frac{3}{5} \approx 0.49$$

According to Precision@8, System 2 outpfermorm System 1.

Task 3: Cohen's Kappa

Two raters A and B agree in a classification task as follows:

		B	
		Yes	No
A	Yes	20	5
	No	10	15

Task 3.1: Compute Cohen's Kappa for this agreement.

$$p_a = \frac{20+15}{20+5+10+15} = 0.7$$
$$p_e = \frac{20+5}{20+5+10+15} \cdot \frac{20+10}{20+5+10+15} + \frac{10+15}{20+5+10+15} \cdot \frac{5+15}{20+5+10+15} = 0.5$$
$$\kappa = \frac{p_a - p_e}{1-p_e} = \frac{0.7-0.5}{1-0.5} = 0.4$$

Task 3.2: What do you expect if you swap 20 and 5? Compute Cohen's Kappa for this.

As the agreement decreases, a decrease of Cohen's Kappa value is expected.

New computation with swapped values:

$$p_a = \frac{5+15}{20+5+10+15} = 0.4$$
$$p_e = \frac{5+20}{20+5+10+15} \cdot \frac{5+10}{20+5+10+15} + \frac{10+15}{20+5+10+15} \cdot \frac{20+15}{20+5+10+15} = 0.5$$
$$\kappa = \frac{p_a - p_e}{1-p_e} = \frac{0.4-0.5}{1-0.5} = -0.2$$

Task 4: Inter-rater agreement: Fleiss' Kappa

Ten raters ($n = 10$) assign five subjects ($N = 5$) to a total of three categories ($k = 3$). The categories are presented in the columns, while the subjects are presented in the rows. Each cell lists the number of raters who assigned the indicated (row) subject to the indicated (column) category.

Category j **Subject id**	1	2	3	P_i
1	0	0	10	1
2	1	7	2	$\frac{22}{45} = 0.4\overline{8}$
3	6	2	2	$\frac{17}{45} = 0.3\overline{7}$
4	3	4	3	$\frac{4}{15} = 0.2\overline{6}$
5	8	2	0	$\frac{29}{45} = 0.6\overline{4}$
p_j	$\frac{9}{25} = 0.36$	$\frac{3}{10} = 0.3$	$\frac{17}{50} = 0.34$	

Task 4.1: Compute the p_j values and $\overline{P}_e$.

$$p_1 = \frac{1}{5 \cdot 10} \sum_{i=1}^{5} n_{ij} = \frac{1}{50}(0 + 1 + 6 + 3 + 8) = \frac{9}{25} = 0.36$$

$$\overline{P}_e = \sum_{j=1}^{3} p_j^2 = 0.36^2 + 0.3^2 + 0.34^2 = 0.3352$$

Task 4.2: Compute the P_i values and $\overline{P}$.

$$P_1 = \frac{1}{10 \cdot (10-1)}((\sum_{j=1}^{3} n_{1j}^2) - 3) = \frac{1}{90}((1^2 + 2^2 + 7^2)^2 - 10) = \frac{22}{45} = 0.4\overline{8}$$

$$\overline{P} = \frac{1}{5} \sum_{i=1}^{5} P_i = \frac{1}{5}(1 + 0.4\overline{8} + 0.3\overline{7} + 0.2\overline{6} + 0.6\overline{4}) = \frac{5}{9} = 0.\overline{5}$$

Task 4.3: Compute Fleiss' Kappa value.

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} = \frac{0.\overline{5} - 0.3352}{1 - 0.3352} \approx 0.3315$$

Task 4.4: What does this Fleiss' Kappa value tell us about the rater agreement?
According to the interpretation table by Landis and Koch, this Kappa value indicated fair agreement between the raters. However, this table should not be accepted universally.

Task 5: Graphical Model for Truth Inference

Five workers assigned one out of five categories to ten different items. Apply the graphical model for truth inference to determine the label of each item.

Worker 1

cat.[1] item	1	2	3	4	5
1	1				
2		1			
3	1				
4			1		
5		1			
6				1	
7				1	
8		1			
9			1		
10	1				

Worker 2

cat. item	1	2	3	4	5
1		1			
2				1	
3		1			
4		1			
5		1			
6				1	
7				1	
8		1			
9	1				
10				1	

Worker 3

cat. item	1	2	3	4	5
1	1				
2	1				
3	1				
4	1				
5	1				
6	1				
7	1				
8	1				
9	1				
10	1				

Worker 4

cat. item	1	2	3	4	5
1		1			
2					1
3		1			
4			1		
5		1			
6	1				
7	1				
8		1			
9	1				
10	1				

Worker 5

cat. item	1	2	3	4	5
1	1				
2		1			
3	1				
4			1		
5					1
6				1	
7				1	
8			1		
9			1		
10					1

[1] category

Scores at R1

Worker 1	1
Worker 2	1
Worker 3	1
Worker 4	1
Worker 5	1

$1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 = \mathbf{3}$

Votes at R1

cat. / item	1	2	3	4	5
1	**3**	2	0	0	0
2	1	2	0	1	1
3	3	2	0	0	0
4	1	1	3	0	0
5	1	3	0	0	1
6	2	0	0	3	0
7	2	0	0	3	0
8	1	3	1	0	0
9	3	0	2	0	0
10	3	0	0	1	1

Labels at R1

cat. / item	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	1	0	0	0	0
4	0	0	1	0	0
5	0	1	0	0	0
6	0	0	0	1	0
7	0	0	0	1	0
8	0	1	0	0	0
9	1	0	0	0	0
10	1	0	0	0	0

Scores at R2

Worker 1	0.9
Worker 2	0.5
Worker 3	0.4
Worker 4	0.5
Worker 5	0.6

In R2, Worker 1 agrees with the label of R1 in 9 out of 10 cases.

$0.9 \cdot 1 + 0.5 \cdot 0 + 0.4 \cdot 1 + 0.5 \cdot 0 + 0.6 \cdot 1 = \mathbf{1.9}$

Votes at R2

cat. / item	1	2	3	4	5
1	**1.9**	1	0	0	0
2	0.4	1.5	0	0.5	0.5
3	1.9	1	0	0	0
4	0.4	0.5	2	0	0
5	0.4	1.9	0	0	0.6
6	0.9	0	0	2	0
7	0.9	0	0	2	0
8	0.4	1.9	0.6	0	0
9	1.4	0	1.5	0	0
10	1.8	0	0	0.5	0.6

Labels at R2

cat. / item	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	1	0	0	0	0
4	0	0	1	0	0
5	0	1	0	0	0
6	0	0	0	1	0
7	0	0	0	1	0
8	0	1	0	0	0
9	0	0	**1**	0	0
10	1	0	0	0	0

Scores at R3

Worker 1	1
Worker 2	0.4
Worker 3	0.3
Worker 4	0.4
Worker 5	0.7

$1 \cdot 1 + 0.4 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 0 + 0.7 \cdot 1 = \mathbf{2}$

Votes at R3

cat. / item	1	2	3	4	5
1	**2**	0.8	0	0	0
2	0.3	1.7	0	0.4	0.4
3	2	0.8	0	0	0
4	0.3	0.4	2.1	0	0
5	0.3	1.8	0	0	0.7
6	0.7	0	0	2.1	0
7	0.7	0	0	2.1	0
8	0.3	1.8	0.7	0	0
9	1.1	0	1.7	0	0
10	1.7	0	0	0.4	0.7

Labels at R3 (**converged**)

cat. / item	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	1	0	0	0	0
4	0	0	1	0	0
5	0	1	0	0	0
6	0	0	0	1	0
7	0	0	0	1	0
8	0	1	0	0	0
9	0	0	1	0	0
10	1	0	0	0	0

Appendix

Precision, Recall & F1

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)}$$

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)}$$

$$F_1 = \frac{2PR}{P+R}$$

Average Precision

If the set of relevant documents for an information need is $\{d_1, ..., d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then the average precision is computed as $AP = \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$.

Cohen's Kappa

		B	
		Yes	No
A	Yes	a	b
	No	c	d

$$p_a = \frac{a+d}{a+b+c+d}$$

$$p_e = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

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

Fleiss' Kappa

n — number of raters, N — number of subjects, k — number of categories

$\kappa = \frac{\overline{P} - \overline{P}_e}{1-\overline{P}_e}$ — Fleiss' Kappa

$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$ — the proportion of all assignments which were to the *j*-th category

$P_i = \frac{1}{n(n-1)} \left(\left(\sum_{j=1}^{k} n_{ij}^2 \right) - n \right)$ — the extent to which raters agree for the *i*-th subject

$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$ $\qquad$ $\overline{P}_e = \sum_{j=1}^{k} p_j^2$

Truth Inference

$score(user) = \Sigma_{c \in categories, i \in items}(vote(user, c, i) \cdot estimatedlabel(c, i))$

$\forall_{c \in categories, i \in items} estimatedvote(c, i) = \Sigma_{user \in users} score(user) \cdot vote(user, c, i)$

$\forall_{c \in categories, i \in items} estimatedlabel(c, i) =$

$1,\ if\ \forall_{cx \in categories} MAX(estimatedscore(cx, i)) = estimatedscore(c, i),\ 0\ otherwise$

Notes & Sources

- Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to Information Retrieval. Vol. 39. Cambridge University Press, 2008.
  - https://nlp.stanford.edu/IR-book/information-retrieval-book.html
- https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html
- https://en.wikipedia.org/wiki/Fleiss'_kappa
- https://en.wikipedia.org/wiki/Cohen's_kappa