Métricas de Evaluación

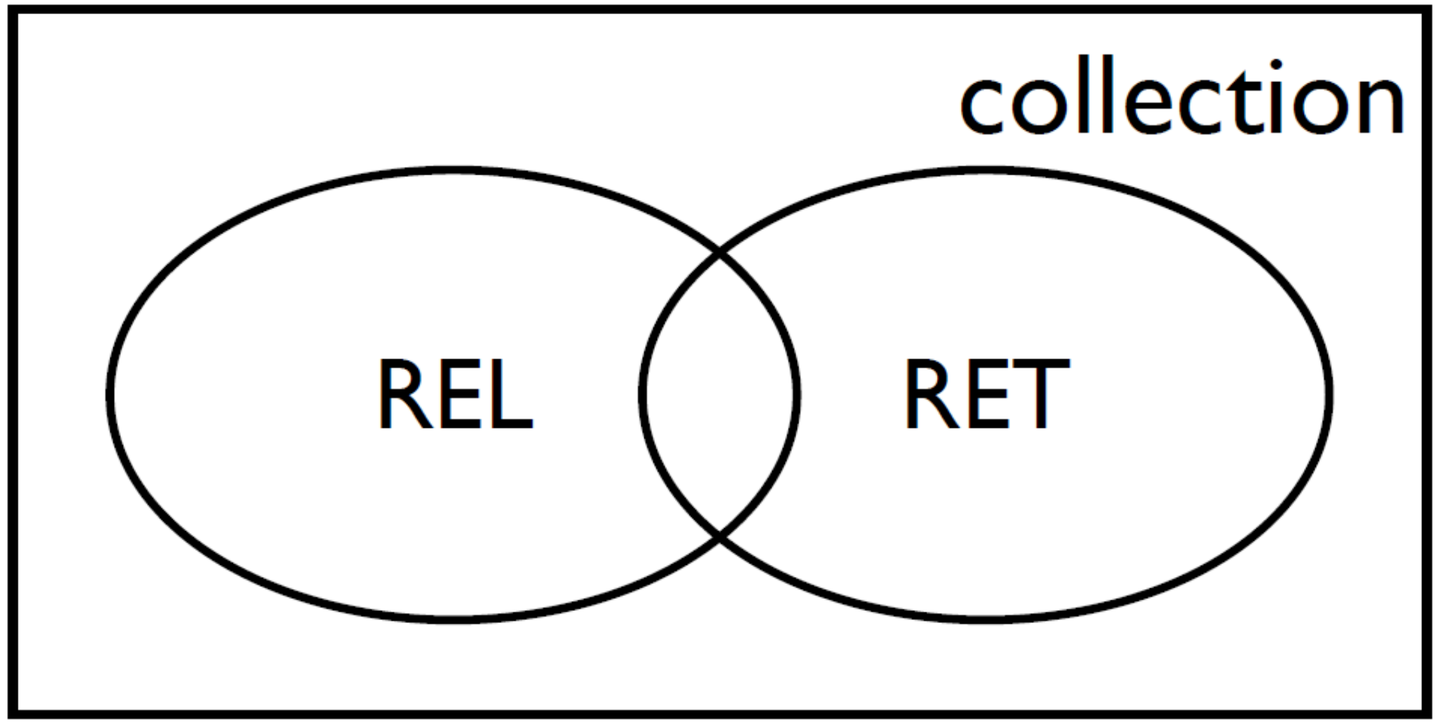
Rubén Francisco Manrique
rf.manrique@uniandes.edu.co

Métricas de evaluación

- Tenemos un conjunto de consultas con resultados clasificados.
- El objetivo de una métrica de evaluación es ***medir la calidad de un ranking*** de documentos relevantes/no relevantes conocidos.
- En otras palabras, necesitamos un ***conjunto de datos de evaluación***.
- ***Conjunto de datos de evaluación***: conjunto de consultas para las cuales se conocen los documentos relevantes.

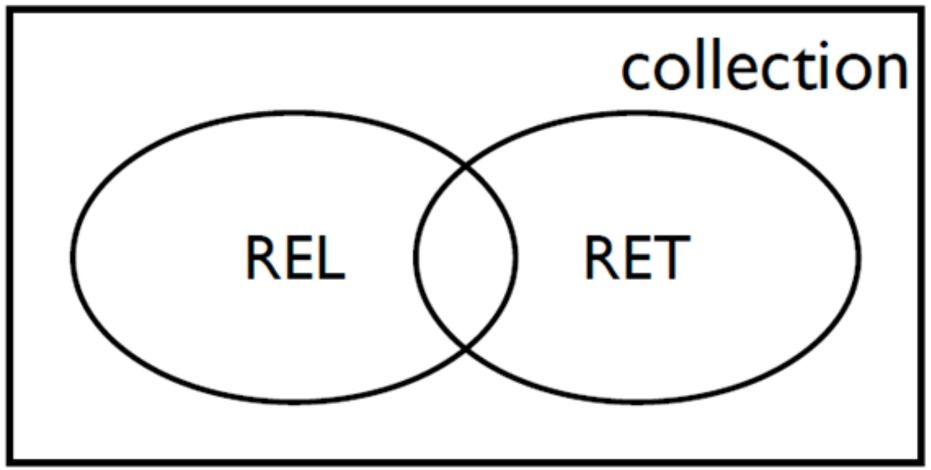
Conjunto Recuperación

- Un conjunto de documentos relevantes (***REL***) y un conjunto de documentos recuperados (***RET***) – Conjunto de datos de evaluación.



Precision y Recall

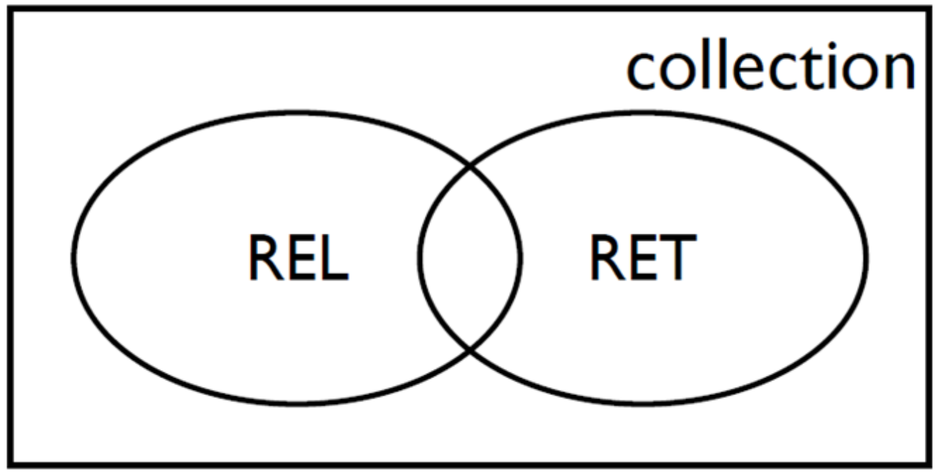
- ***Precision (P):*** la proporción de documentos recuperados que son relevantes.



$$\mathcal{P} = \frac{|RET \cap REL|}{|RET|}$$

Precision y Recall

- ***Recall (P):*** la proporción de documentos relevantes que son recuperados.



$$\mathcal{R} = \frac{|RET \cap REL|}{|REL|}$$

Precision y Recall

- El ***recall*** mide la capacidad del sistema para encontrar todos los documentos relevantes.

- La ***precision*** mide la capacidad del sistema para rechazar cualquier documento no relevante en el conjunto recuperado.

Precision y Recall

- Un sistema puede cometer dos tipos de errores:
  - un error de falso positivo: el sistema recupera un documento que no es relevante (no debería haberse recuperado)
  - un error de falso negativo: el sistema no puede recuperar un documento que es relevante (debería haber sido recuperado)

- ¿Cómo afectan estos tipos de errores a la precisión y el recall?

Combinar Precision y Recall

- A menudo, queremos un sistema que tenga alta precisión y alto recall.
  - Queremos una métrica que mida el equilibrio entre precisión y recuperación.
  - Una posibilidad sería utilizar la media aritmética:

$$\text{arithmetic mean}(\mathcal{P}, \mathcal{R}) = \frac{\mathcal{P} + \mathcal{R}}{2}$$

Combinar Precision y Recall

- ***Malo***: un sistema que obtiene una precisión de 1,0 y un recall cercano a 0,0 obtendría un valor medio de alrededor de 0,50.
  - Un sistema que recupera un único documento relevante obtendría una precisión de 1,0 y una recuperación cercana a 0,0.

- ***Malo***: un sistema que obtiene un recall de 1,0 y una precisión cercana a 0,0 obtendría un valor medio de alrededor de 0,50.
  - Un sistema que recupera toda la colección obtendría una recuperación de 1,0 y una precisión cercana a 0,0.

- ***Mejor***: un sistema que obtiene una precisión de 0,50 y un recall cercana a 0,50 obtendría un valor medio de alrededor de 0,50.

Combinar Precision y Recall

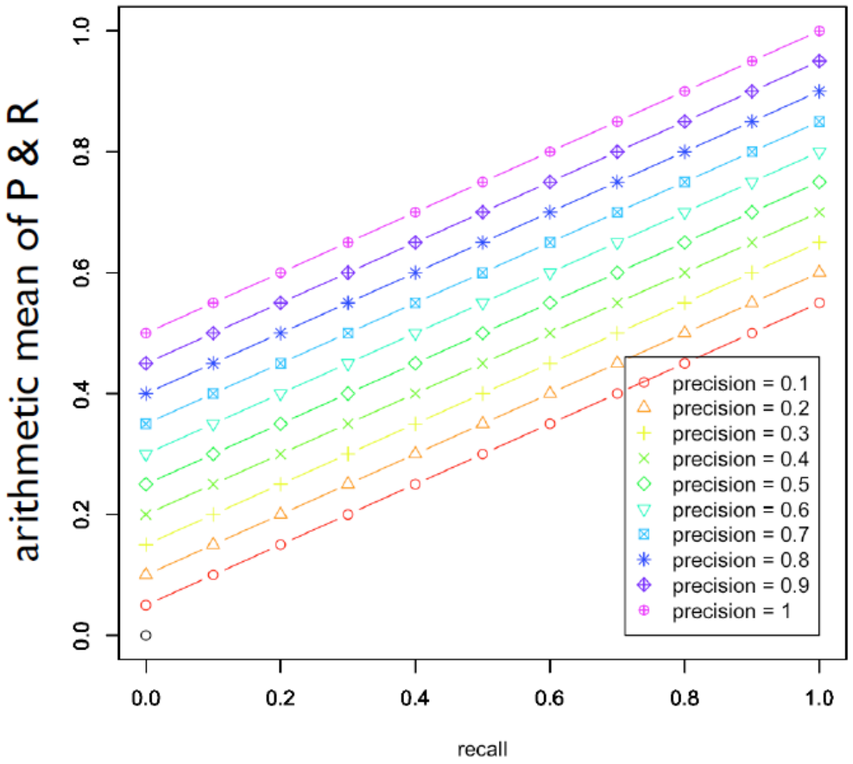
- ***Solución:*** utilice la media armónica en lugar de la media aritmética.

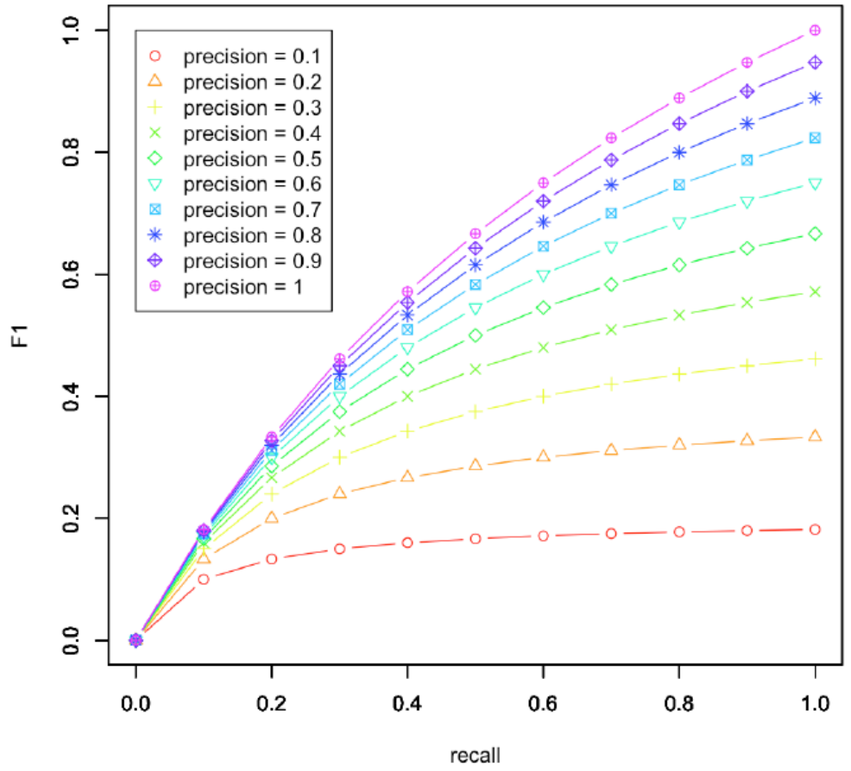
- F-medida:

$$\mathcal{F} = \frac{1}{\frac{1}{2}\left(\frac{1}{\mathcal{P}} + \frac{1}{\mathcal{R}}\right)} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$$

F-medida

- La media armónica castiga los valores pequeños.





Recuperación ranqueada

- Recordar: El sistema genera una lista clasificada de documentos en lugar de un conjunto desordenado.

- Suposición de comportamiento del usuario: El usuario examina el ranking de salida de arriba a abajo hasta que está satisfecho o se da por vencido.

- La precisión y el recall también se pueden utilizar para evaluar una clasificación.
  - Precision/Recall @ rank K

Recuperación ranqueada: P@K, R@K

- P@K: proporción de documentos top-K recuperados que son relevantes.
- R@K: proporción de documentos relevantes que se recuperan en el top-K.
- Suposición: el usuario solo examinará los resultados top-K.

Recuperación ranqueada: P@K, R@K

Asuma que hay 20 documentos relevantes.

K	P@K	R@K
1	(1/1) = 1.0	(1/20) = 0.05
2		
3		
4		
5		
6		
7		
8		
9		
10		

Recuperación ranqueada: P@K, R@K

- Asuma que hay 20 documentos relevantes.

K	P@K	R@K
1	(1/1) = 1.0	(1/20) = 0.05
2	(1/2) = 0.5	(1/20) = 0.05
3		
4		
5		
6		
7		
8		
9		
10		

Recuperación ranqueada: P@K, R@K

Asuma que hay 20 documentos relevantes.

K	P@K	R@K
1	(1/1) = 1.0	(1/20) = 0.05
2	(1/2) = 0.5	(1/20) = 0.05
3	(2/3) = 0.67	(2/20) = 0.10
4	(3/4) = 0.75	(3/20) = 0.15
5	(4/5) = 0.80	(4/20) = 0.20
6		
7		
8		
9		
10		

Recuperación ranqueada: P@K, R@K

Asuma que hay 20 documentos relevantes.

K	P@K	R@K
1	(1/1) = 1.0	(1/20) = 0.05
2	(1/2) = 0.5	(1/20) = 0.05
3	(2/3) = 0.67	(2/20) = 0.10
4	(3/4) = 0.75	(3/20) = 0.15
5	(4/5) = 0.80	(4/20) = 0.20
6	(5/6) = 0.83	(5/20) = 0.25
7	(6/7) = 0.86	(6/20) = 0.30
8	(6/8) = 0.75	(6/20) = 0.30
9	(7/9) = 0.78	(7/20) = 0.35
10	(7/10) = 0.70	(7/20) = 0.35

¿En que momento R@K será 1?

Recuperación ranqueada: P@K, R@K

- Problema: ¿qué valor de K debemos usar para evaluar?
- Si no sabemos qué valor de K elegir, podemos calcular y reportar varios: P/R@{1,5,10,20}
- Hay métricas de evaluación que no requieren elegir K (como veremos).
- Tenga en cuenta que este cálculo es solo para una consulta, debe informar el promedio entre todas las consultas disponibles.

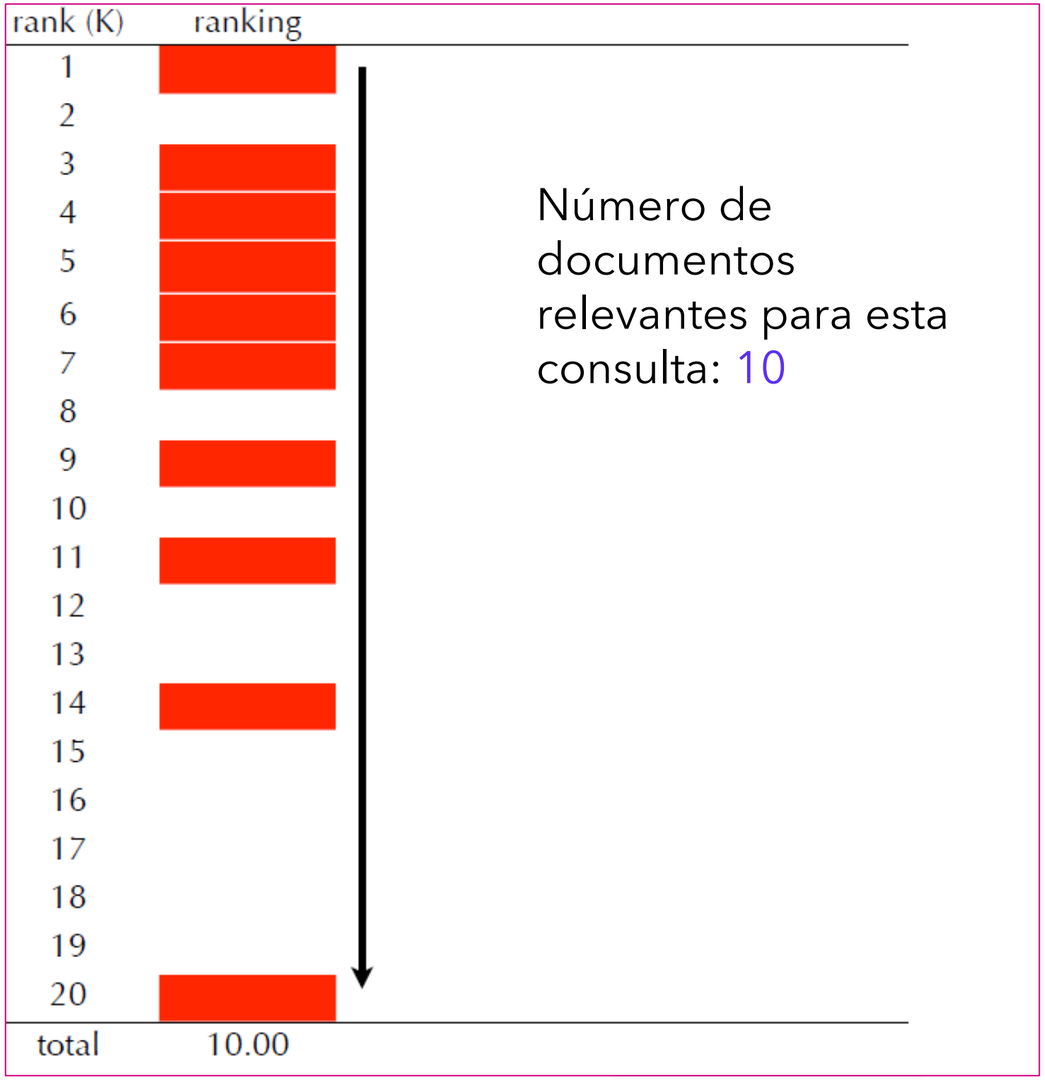
Recuperación ranqueada: Average Precision

- Idealmente, queremos que el sistema logre una alta precisión para valores variables de K.

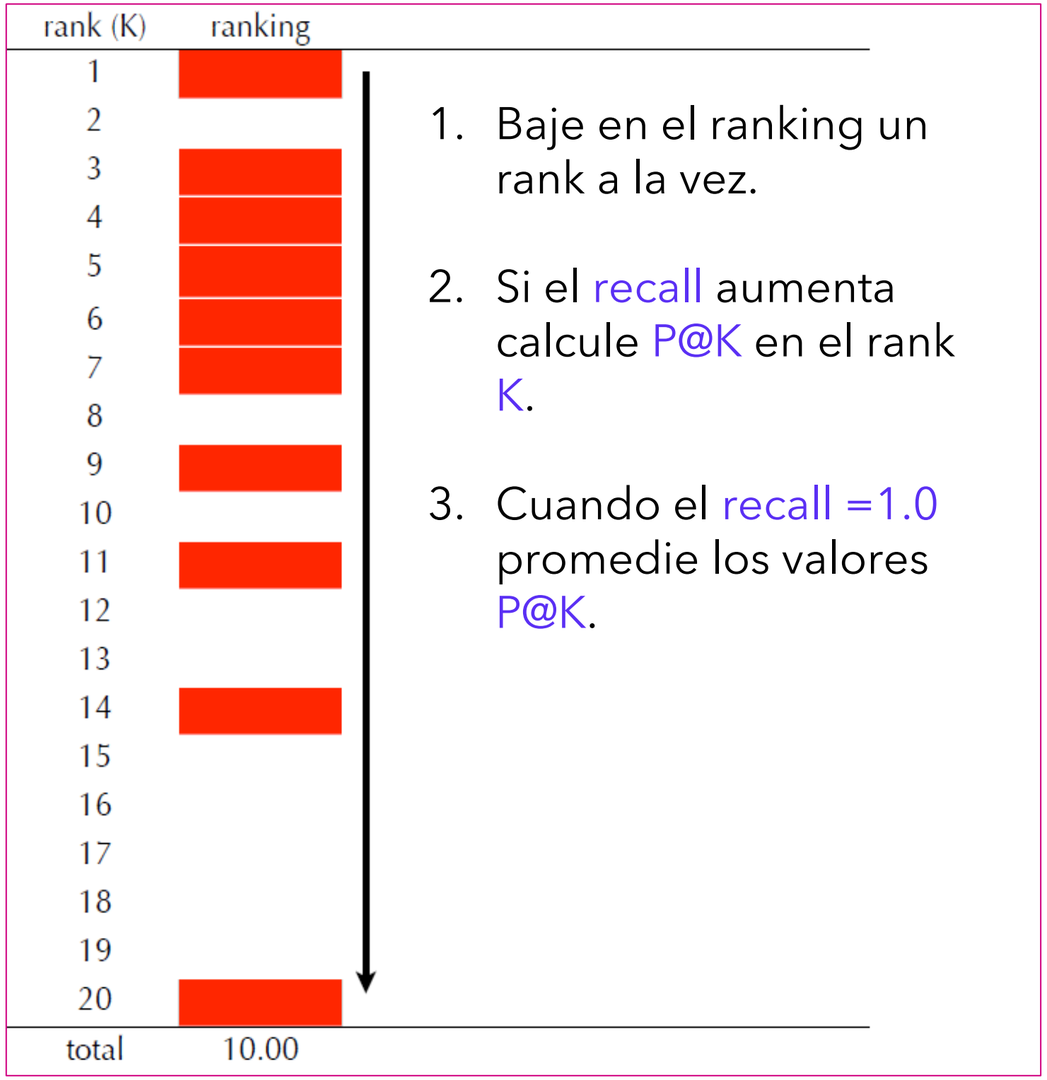
- La precisión promedio (average-precision) métrica da cuenta de la precisión y recall sin tener que establecer K.

- Pasos de calculo:
  - 1. Baja en el ranking de un rango a la vez.
  - 2. Si el documento en el rango K es relevante, mida P@K.
  - 3. Cuando recall = 1.0, tome el promedio de todos los valores P@K.
    - El número de valores P@K será igual al número de documentos relevantes.

Recuperación ranqueada: Average Precision



Número de documentos relevantes para esta consulta: 10



1. Baje en el ranking un rank a la vez.
2. Si el recall aumenta calcule P@K en el rank K.
3. Cuando el recall =1.0 promedie los valores P@K.

Recuperación ranqueada: Average Precision

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

Intercambiar los ranks 2 y 3.

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.79

Recuperación ranqueada: Average Precision

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

Intercambiar los ranks 8 y 9.

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.70	0.88
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.77

Recuperación ranqueada: Average Precision

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.30	1.00
4		0.40	1.00
5		0.50	1.00
6		0.60	1.00
7		0.70	1.00
8		0.80	1.00
9		0.90	1.00
10		1.00	1.00
11		1.00	0.91
12		1.00	0.83
13		1.00	0.77
14		1.00	0.71
15		1.00	0.67
16		1.00	0.63
17		1.00	0.59
18		1.00	0.56
19		1.00	0.53
20		1.00	0.50
total	10.00	average-precision	1.00

rank (K)	ranking	R@K	P@K
1		0.00	0.00
2		0.00	0.00
3		0.00	0.00
4		0.00	0.00
5		0.00	0.00
6		0.00	0.00
7		0.00	0.00
8		0.00	0.00
9		0.00	0.00
10		0.00	0.00
11		0.10	0.09
12		0.20	0.17
13		0.30	0.23
14		0.40	0.29
15		0.50	0.33
16		0.60	0.38
17		0.70	0.41
18		0.80	0.44
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.33

Recuperación ranqueada: MAP

- La precisión promedio (average precision) se calcula para una sola consulta.
- Mean Average Precision (MAP): precisión media promediada en un conjunto de consultas.
- ***Una de las métricas más comunes en la evaluación de IR***

Recuperación ranqueada: Niveles de Relevancia

- Qué sucede cuando hay más de dos niveles de relevancia (por ejemplo, perfecto, excelente, bueno, regular, malo; 5-4-3-2-1).

- Opción 1: transformar a una función de puntuación de relevancia binaria y aplicar P@K, R@K.

$$rel_b(d_i) = \begin{cases} 1, & rel(d_i) \geq 3 \\ 0, & rel(d_i) < 3 \end{cases}$$

- Opción 2: Ganancia acumulada descontada (DCG).

Ganancia acumulada descontada (DCG)

- Sea REL_i la relevancia asociada con el documento en el rank $i\,(1 \leq i \leq K)$.

- Supongamos una escala no binaria (cada valor posible $\in \mathbb{N}$):

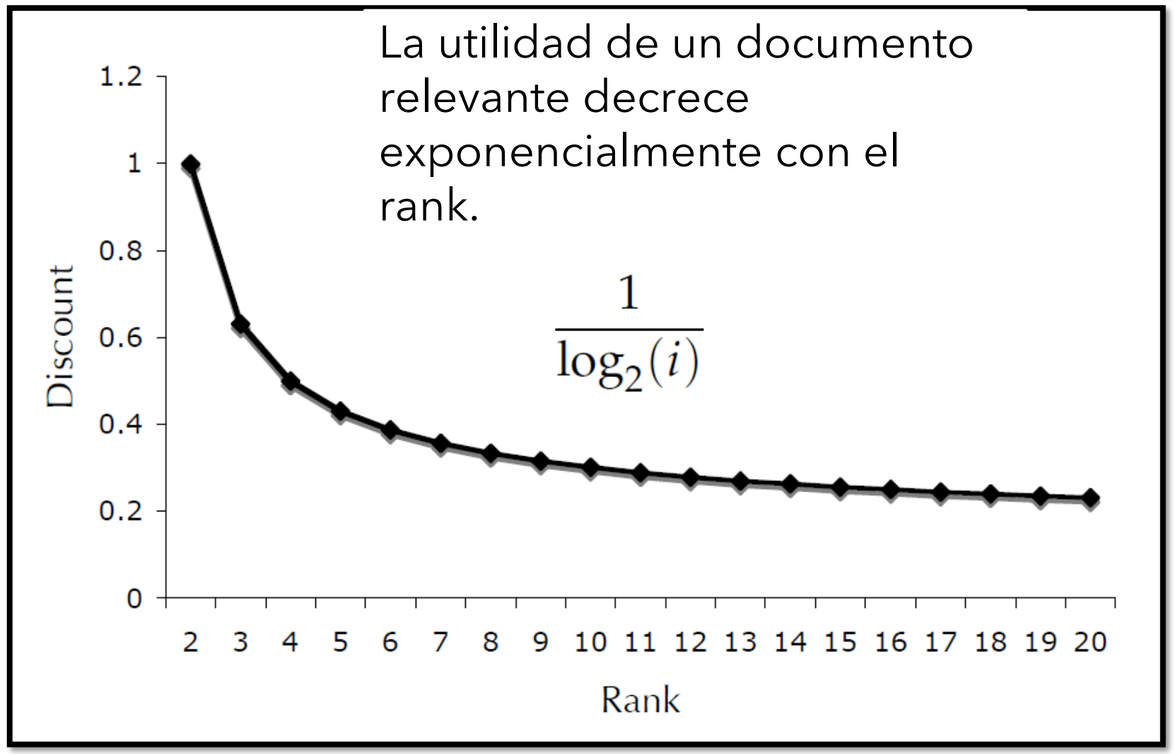
  - perfect ➔ 4
  - excellent ➔ 3
  - good ➔ 2
  - fair ➔ 1
  - bad ➔ 0

Ganancia acumulada descontada

- DCG se define como.

$$DCG@K = \sum_{i=1}^{K} \frac{REL_i}{log_2(\max(i, 2))}$$

La utilidad de un documento relevante para un usuario disminuye rápidamente con el rango (más rápidamente que linealmente)



Ganancia acumulada descontada

$$DCG@K = \sum_{i=1}^{K} \frac{REL_i}{log_2(\max(i, 2))}$$

- perfect ➔ 4
- excellent ➔ 3
- good ➔ 2
- fair ➔ 1
- bad ➔ 0

rank (i)	REL_i
1	4
2	3
3	4
4	2
5	0
6	0
7	0
8	1
9	1
10	0

Ganancia acumulada descontada

$$DCG@K = \sum_{i=1}^{K} \frac{REL_i}{log_2(\max(i, 2))}$$

Cada rank se asocial con un factor de descuento:

$$\frac{1}{log_2(\max(i, 2))}$$

rank (i)	REL_i	discount factor
1	4	1.00
2	3	1.00
3	4	0.63
4	2	0.50
5	0	0.43
6	0	0.39
7	0	0.36
8	1	0.33
9	1	0.32
10	0	0.30

Ganancia acumulada descontada

$$DCG@K = \sum_{i=1}^{K} \frac{REL_i}{log_2(\max(i, 2))}$$

rank (i)	REL_i	discount factor	gain
1	4	1.00	4.00
2	3	1.00	3.00
3	4	0.63	2.52
4	2	0.50	1.00
5	0	0.43	0.00
6	0	0.39	0.00
7	0	0.36	0.00
8	1	0.33	0.33
9	1	0.32	0.32
10	0	0.30	0.00

Ganancia acumulada descontada

$$DCG@K = \sum_{i=1}^{K} \frac{REL_i}{log_2(\max(i, 2))}$$

rank (i)	REL_i	discount factor	gain	DCG_i
1	4	1.00	4.00	4.00
2	3	1.00	3.00	7.00
3	4	0.63	2.52	9.52
4	2	0.50	1.00	10.52
5	0	0.43	0.00	10.52
6	0	0.39	0.00	10.52
7	0	0.36	0.00	10.52
8	1	0.33	0.33	10.86
9	1	0.32	0.32	11.17
10	0	0.30	0.00	11.17

$DCG_{10} = 11.17$

Ganancia acumulada descontada

- Problema: DCG no está "limitado", lo que hace que sea problemático promediar entre consultas.
  - Los valores no son comparables en varias consultas.

- NDCG: ganancia acumulada descontada normalizada [0,1]
  - Para una consulta dada, mide DCG_i
  - Luego, divida este valor DCG_i entre el mejor posible DCG_i para esa consulta

El mejor posible DCG_i

- Suponga una consulta con:
  - 2 documentos con una relevancia de 4.
  - 3 documentos con una relevancia de 3.
  - 2 documentos con una relevancia de 2.
  - El resto de documentos son 0s.

- Cual es el mejor posible ranking para i=1:
  - 4,3,3,3,4,2,2,0,0….
  - 4,0,0,0,0,4,0,0,0….

- Cual es el mejor posible ranking para i=4:
  - 4,4,3,3….

El mejor posible DCG_i: Ejemplo

rank (i)	REL_i	discount factor	gain	DCG_i
1	4	1.00	4.00	4.00
2	3	1.00	3.00	7.00
3	4	0.63	2.52	9.52
4	2	0.50	1.00	10.52
5	0	0.43	0.00	10.52
6	0	0.39	0.00	10.52
7	0	0.36	0.00	10.52
8	1	0.33	0.33	10.86
9	1	0.32	0.32	11.17
10	0	0.30	0.00	11.17

rank(i)	*REL_i*	*discount factor*	*gain*	*DCG_i*
1	4	1	4	4
2	4	1	4	8
3	3	0.63	1.89	9.89
4	2	0.5	1	10.89
5	1	0.43	0.43	11.32
6	1	0.39	0.39	11.71
7	0	0.36	0	11.71
8	0	0.33	0	11.71
9	0	0.32	0	11.71
10	0	0.3	0	11.71

$$NDCG_2 = \frac{7}{8} = 0.875$$

$$NDCG_{10} = \frac{11.17}{11.71} = 0.9538$$

Resumen

- ***P@K***: precisión bajo el supuesto de que los resultados top-K son el 'conjunto' recuperado.
- ***R@K:*** recall bajo el supuesto de que los resultados top-K son el 'conjunto' recuperado.
- ***Average Precision***: considera la precisión y recall y se enfoca principalmente en los mejores resultados. ***MAP*** es el promedio del average precision.
- ***DCG:*** ignora el recuerdo, considera múltiples niveles de relevancia y se enfoca muy necesariamente en los rangos superiores
- ***NDCG***: truco para hacer que DCG oscile entre 0 y 1

Referencias

- Jaime Arguello INLS 509: Information Retrieval
- Introduction to information retrieval https://nlp.stanford.edu/IR-book/
- Jurafsky D. and Martin J. (2021) Speech and Language Processing (3rd ed. draft). Online: https://web.stanford.edu/~jurafsky/slp3/
- Yoav Goldberg (2017). Neural Network Methods in Natural Language Processing.
- In Deng, L., & In Liu, Y. (2018). Deep learning in natural language processing.