

TEMA 5. EVALUACIÓN DE SISTEMAS DE RI

Contenidos

1. Objetivo de la evaluación
2. Eficacia en RI
3. Relevancia
4. Colecciones de test
5. Métricas para la eficacia.
 - 5.1. Evaluar resultados de RI no ordenados.
 - 5.2. Evaluar resultados de RI ordenados.

En este tema se estudian los métodos de evaluación de los sistemas de RI, es decir, se definen algunas medidas que intentan representar la calidad de los documentos recuperados por un sistema. En base a estas medidas objetivas podemos decidir qué sistema es mejor, respecto a los requisitos que nos hayamos planteado a la hora de desarrollar el sistema.

Bibliografía

A Introduction to Information Retrieval:

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.
Cambridge University Press, **2009**.

Capítulo 8

Speech and Language Processing: International Version, 2/E.

Daniel Jurafsky, James H. Martin.

Pearson International Edition, **2009**. ISBN-10: 0135041961.

Capítulo 23



1. OBJETIVO DE LA EVALUACIÓN

Objetivo de la evaluación

Comparar sistemas entre sí.

¿Qué podemos comparar?

- Eficacia (precisión y cobertura).
- Eficiencia (temporal y espacial).
- Satisfacción del usuario (interfaz, modo de presentación de resultados, etc.).



La Eficacia mide la calidad del conjunto de documentos recuperados (esta medida es la que estudiaremos en este tema)

La Eficiencia es una medida objetiva de costes computacionales (temporal y espacial)

La Satisfacción del usuario es una medida subjetiva, que sólo puede obtenerse si se le solicita la opinión al usuario.

2. EFICACIA EN RI

¿Cómo medir la eficacia en RI ?

- Necesitamos una colección de test consistente en tres elementos:
 1. Una colección de documentos.
 2. Un conjunto de consultas.
 3. Un conjunto de juicios de relevancia.*

**(gold standard o ground truth o judgment of relevance)*

Por Juicio de relevancia de una consulta se entiende el conjunto de documentos que debería devolverse al realizar esa consulta.

Evidentemente es difícil disponer de estos Juicios de relevancia. En el caso ideal habría que leerse todos los documentos y decidir los que son adecuados para la consulta (y eso para todas las consultas). Aún así podría haber errores ya que los criterios sobre si un documento es adecuado o no pueden ser subjetivos.

Ajuste del rendimiento del sistema

- Cálculo de pesos (parámetros) para ajustar el rendimiento del sistema.
- **Incorrecto:** evaluar un sistema tomando la colección sobre la que se han ajustado los parámetros para maximizar su rendimiento.
- **Correcto:** utilizar una colección de desarrollo para entrenar y ajustar los parámetros, y otra colección para evaluar el rendimiento imparcial.

Cuando se dispone de un conjunto de test (es decir, consultas, documentos y juicios de relevancia) y queremos ajustar nuestro sistema para que funcione lo mejor posible no podemos usar los datos con los que nos vamos a evaluar para ir mejorando, ya que nos estaríamos ajustando a esos datos de test, de forma que un cambio posterior de conjunto de test podría comportar un empeoramiento del comportamiento del sistema. Además, si se trata de comparar diferentes sistemas, el ajuste de los sistema debe ser ciego a los datos de test.

Hay que apartar un subconjunto de esos datos para realizar los ajustes, y una vez preparado el sistema entonces pasamos el test de verdad (que son preguntas y respuestas que nunca ha visto el sistema).

3. RELEVANCIA

¿Cómo expresar la relevancia?

- Forma estándar: valoración binaria de si es relevante o no relevante para cada par consulta-documento.
- Otras alternativas, considerar una relevancia dentro de una escala, como documentos altamente relevantes o marginalmente relevantes u otras.

Por estándar se entiende que sólo se dice si un documento es relevante o no, pero no se aporta ningún orden respecto a la relevancia.

Ejemplo- Listas de consultas evaluadas (Colección Times)

1.

- KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP SUPPRESSING THE BUDDHISTS .
- Documentos relevantes: 268 288 304 308 323 326 334.

2.

- EFFORTS OF AMBASSADOR HENRY CABOT LODGE TO GET VIET NAM'S PRESIDENT DIEM TO CHANGE HIS POLICIES OF POLITICAL REPRESSION .
- Documentos relevantes: 326 334

donde cada documento relevante de una consulta tiene el mismo grado de relevancia.

4. COLECCIONES DE TEST

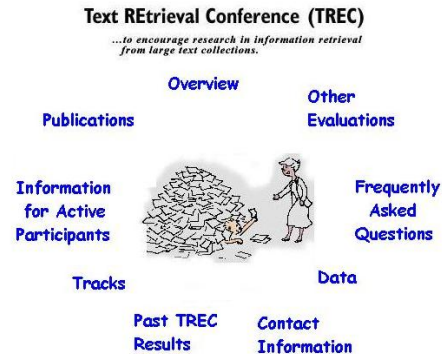
¿Cómo construir Colecciones de test?

- **Tradicional**: se estudia la relevancia de cada documento para cada consulta.
- **Pooling** (en competencias): sólo se estudia la relevancia de los documentos devueltos por los participantes del concurso.
 - *Ventaja*: permite trabajar con colecciones de gran volumen de documentos.
 - *Desventaja*: no se sabe la relevancia de todos los documentos.

Por Tradicional se entiende que tenemos identificada la relevancia de cada documento para cada consulta. Esto es prácticamente imposible en colecciones de millones de documentos.

Colecciones de test más extendidas:

- **TREC Ad Hoc track**
(primeras 8 evaluaciones del TREC entre 1992 y 1999): 1.89 millones de documentos y juicios de relevancia para 450 consultas.
- La mejor subcolección, más consistente, la constituyen TREC's 6-8 con 150 consultas sobre 528.000 artículos. <http://trec.nist.gov>



Para apoyar la investigación dentro de la comunidad de recuperación de información se recopilieron grandes colecciones de documentos, con sus correspondientes consultas y juicios de relevancia, que se pueden usar para estudiar el comportamiento de los sistemas.

National Institute of Standards and Technology (NIST).

Colecciones de test más extendidas:

- **CLEF- Conference and Labs of the Evaluation Forum**
(Cross Language Evaluation Forum)

<http://www.clef-initiative.eu/>

Promover la investigación, la innovación y el desarrollo de sistemas de acceso a la información con énfasis en la información multilingüe y multimodal.

- **Reuters-21578**, una colección de documentos
(inicialmente con 21.578 artículos que amplió a 806.791)
diseñada para tareas de clasificación de texto.

5. MÉTRICAS PARA LA EFICACIA

- 5.1. Evaluar resultados de RI no ordenados.
- 5.2. Evaluar resultados de RI ordenados.

Métricas para la eficacia

La forma de evaluación depende de si el conjunto de documentos resultante de la recuperación está ordenado o no (ranked or unranked).

5.1. Evaluar resultados de RI no ordenados.

Dos métricas muy extendidas en el área de la RI, precisión y cobertura (en inglés, precision y recall):

precision (P) en RI sirve para medir la fracción de documentos recuperados que son relevantes.

$$P = \text{nº de docs relevantes recuperados} / \text{nº de docs recuperados}$$

recall (R) en RI se define como la fracción de documentos relevantes que son recuperados.

$$R = \text{nº de docs relevantes recuperados} / \text{nº de docs relevantes en la colección}$$

Otra manera de presentar estas métricas

	Relevante	No relevante
Recuperado	true positives (tp)	false positives (fp)
No Recuperado	false negatives (fn)	true negatives (tn)

Precision $P = tp / (tp + fp)$

Recall $R = tp / (tp + fn)$

Recuperado-Relevante: Número de documentos que devuelve nuestro sistema que son relevantes para la consulta.

Recuperado-NoRelevante: Número de documentos que devuelve nuestro sistema que no debía haber devuelto para esa consulta

NoRecuperado-Relevante: Número de documentos que son relevantes y que nuestro sistema ha sido incapaz de recuperar

NoRecuperado-NoRelevante: Número de documentos que no son relevantes y que nuestro sistema ha clasificado correctamente, ya que no los ha recuperado.

La Precision mide la calidad del conjunto de documentos recuperados, es decir, el porcentaje de documentos buenos entre todos los documentos recuperados.

El Recall (cobertura) mide la cantidad de documentos correctamente recuperados en relación al total de documentos relevantes para esa consulta en la colección.

- **F-Medida** (F-measure o media armónica ponderada) es otra métrica que combina la precisión y la cobertura según un parámetro β .

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{Si } \beta = 1 \quad F_1 = \frac{2PR}{P + R}$$

- Si $\beta < 1$ se da mayor importancia a la Precision
- Si $\beta > 1$ se le da mayor importancia al Recall.

Habitualmente se utiliza el valor de 1 para la beta.

Ejercicio #1: Hay un total de 20 documentos relevantes en la colección. El sistema devuelve 8 relevantes y 10 no relevantes.

	Relevante	No relevante
Recuperado	8 tp	10 fp
No Recuperado	12 fn	tn

Calcula la Precisión, el Recall y la F_1 -medida

Ejercicio #1_Sol: Hay un total de 20 documentos relevantes en la colección. El sistema devuelve 8 relevantes y 10 no relevantes.

$$P = \frac{tp}{tp + fp} = \frac{8}{18} = 0.444 \quad R = \frac{tp}{tp + fn} = \frac{8}{20} = 0.4$$

$$F_1 = \frac{2xPxR}{P + R} = \frac{0.3552}{0.844} = 0.4208$$

5.2. Evaluar resultados de RI ordenados.

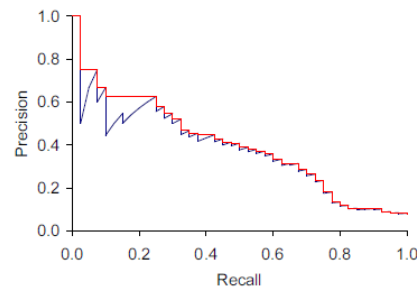
- La Precision, el Recall, y la F-medida son medidas basadas en conjuntos de documentos no ordenados.
- Necesitamos extender estas medidas, o definir otras nuevas si queremos mostrar los k documentos de la parte superior de la lista ordenada de documentos recuperados (ranking).
- La métrica debería preferir un sistema de RI que ponga más arriba los documentos relevantes.

Las medidas de precisión y el recall definidas anteriormente no dependen de ninguna ordenación de los resultados.

Para tener en cuenta la ordenación, necesitamos adaptarlas para captar si un sistema es capaz de colocar documentos relevantes en un lugar más alto en la ordenación. Los métodos estándar en la recuperación de información para lograr esto se basan en trazar curvas de precisión/recall y en promediar medidas de precisión de varias maneras.

Curva precision-recall

- La curva (línea azul) tendrá forma de sierra debido a que si el $(k+1)$ -ésimo documento recuperado no es relevante la cobertura es la misma que la de los k documentos del tope pero la precisión descenderá.
- Una manera de suavizar estos dientes es calculando la **precisión interpolada** (línea roja).



En un contexto de recuperación ordenada, los “mejores” documentos recuperados son proporcionados por los k primeros documentos recuperados. Para cada uno de esos conjuntos, los valores de precisión y recall se pueden trazar para obtener una curva de precisión/recall.

Si representamos en un gráfico los valores de la precisión en función de los valores del recall se obtiene una curva (en azul) como la de la figura. Es decir, en el eje ‘x’ se representan los posibles valores de recall (0, 0,2, 0,4... evidentemente pueden usarse intervalos más pequeños), y el eje ‘y’ representa la precisión que se consigue. Por ejemplo, si para conseguir un recall de 0,2 ha sido necesario que nuestro sistema devuelva 10 documentos (de los cuales 4 son buenos y los restantes no), entonces la precisión es 0,4 (punto (0,2 0,4) del espacio de representación).

5.2.1. Precisión interpolada

Definición (Manning et al., 2009-Cap.8):

P_{interp} a un cierto nivel de recall r se define como la precisión más alta que se encuentra para cualquier nivel de recall $r' \geq r$:

$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

Es decir, nos situamos en el mejor caso. Si obtenemos un mismo recall con distintas tallas de la lista de documentos recuperados, nos quedamos con la opción que proporciona mayor precisión.

5.2.1. Precisión interpolada

Aunque toda la curva interpolada puede ser informativa, generalmente se trabaja con un número bajo de valores.

Precisión interpolada en 11 niveles de recall:

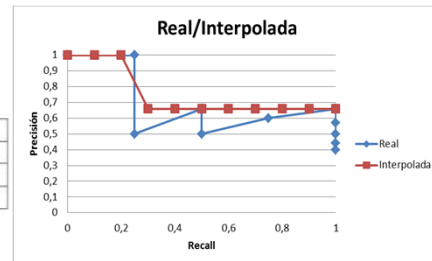
Para cada valor de recall estándar **i** desde 0.0 a 1.0 con incrementos de 0.1, se toma la precisión máxima obtenida en cualquier valor de recall real mayor o igual a **i**.

Ejercicio#2.

Para la consulta Q1 tenemos 4 documentos relevantes.
Calcula Precision&Recall Reales e Interpolados.

- Tabla Precision&Recall Reales:

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00



- Tabla Precision&Recall Interpoladas:

Precision	1	1	1	2/3	2/3	2/3	4/6 = 2/3	4/6	4/6	4/6	4/6
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Para $r' \in [0.5-1] \geq r=0.3$
 $\max P(r') = \frac{2}{3}$

P_{interp} a un cierto nivel de recall r se define como la precisión más alta que se encuentra para cualquier nivel de recall $r' \geq r$

A la vista de estas curvas se puede comparar el comportamiento de los sistemas. Aquellas curvas que muestran una precisión superior para los mismos recall representan un mejor comportamiento. También puede analizarse el comportamiento en determinados segmentos del recall: cuando la curva tiene valores altos en la parte izquierda significa que hay una buena precisión para valores bajos de recall (es decir, los primeros documentos que se recuperan son bastante buenos). Cuando se mantiene un valor alto al final de la curva significa que el sistema no tiene mucha pérdida (no introduce muchos documentos irrelevantes) cuando aumenta el número de documentos recuperados.

Ejercicio#3.

Para la consulta Q2 tenemos 5 documentos relevantes.
Calcula Precision&Recall Reales e Interpolados.

- Tabla Precision&Recall Reales

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision										
Recall										

Ejercicio#3_sol.

Para la consulta Q2 tenemos 5 documentos relevantes.
Calcula Precision&Recall Reales e Interpolados.

- Tabla Precision&Recall Reales:

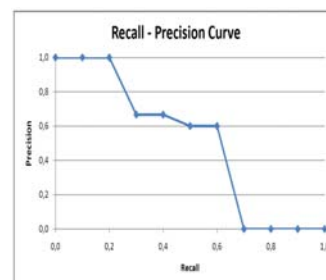
Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	3/9	3/10
Recall	0.2	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6

- Tabla Precision&Recall Interpoladas:

Precision	1	1	1	2/3	2/3	3/5	3/5	0	0	0
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	1.0

Para $r' \in [0.4-0.6] \geq r=0.3$
 $\max P(r') = \frac{2}{3}$

Ya no hay valores $r' \geq r$
 $\max P(r') = 0$



5.2.2 Promedio interpolado

11-puntos de precisión promedia interpolada

Cuando se dispone de más de una consulta:

- Para cada consulta, la precisión interpolada se mide en los 11 niveles de Recall de 0.0, 0.1, 0.2, . . . , 1.0.
- Para cada nivel de recall se calcula la *media aritmética* de la precisión interpolada a ese nivel de recall para cada consulta en la colección de test.

Ejercicio #4.

Para las consultas Q1 y Q2 calcula Precision&Recall Promedio Interpolado y la curva de Recall-Precision.

- Tabla Precision&Recall Interpoladas Q1

Precision	1	1	1	2/3	2/3	2/3	4/6 = 2/3	4/6	4/6	4/6	4/6
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

- Tabla Precision&Recall Interpoladas Q2

Precision	1	1	1	2/3	2/3	3/5	3/5	0	0	0	0
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

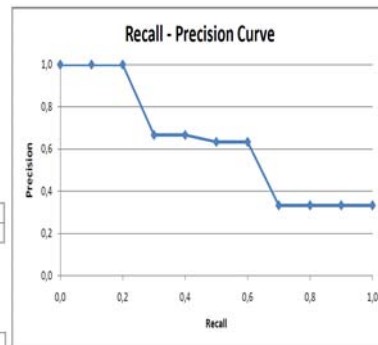
- Tabla Precision&Recall Interpoladas Q1+Q2

Precision	1	1	1	2/3	2/3	19/30	19/30	1/3	1/3	1/3	1/3
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

$$\frac{2/3 + 2/3}{2}$$

$$\frac{2/3 + 3/5}{2} = 19/30$$

$$\frac{4/6 + 0}{2} = 1/3$$



5.2.3 Medir la eficacia mediante un único valor

- Mean Average Precision (MAP)
- R-Precision
- Precision-at- k

Mean Average Precision (MAP)

- Proporciona una medida única de calidad en todos los niveles de recall.
- Para una consulta **simple** la **Precisión media** es el promedio del valor de precisión obtenido después de que cada documento relevante sea recuperado en la lista ordenada de documentos recuperados.
- Si algún documento relevante no se recupera entonces el valor de precisión que sumaremos para promediar será 0.
- Tenemos que conocer a priori n° de docs. relevantes.

Otra forma de evaluar un sistema que devuelve documentos ordenados es mediante la Mean Average Precision. Consiste en calcular la media del valor de las precisiones cada vez que aparece un documento relevante en la lista que devuelve el Sistema. Para el resto de documentos relevantes que no son encontrados se suma el valor 0 como precisión para el cálculo de la media.

Ejemplo. Consulta con 5 documentos relevantes en las posiciones 1, 3, 6, 10 y 15

Obtendríamos precisiones de 1, 0.66, 0.5, 0.4, 0.33 entonces la **precisión media de la consulta** sería:

$$P_{media} = \frac{1 + 0.66 + 0.5 + 0.4 + 0.33}{5} = 0.57$$

- Esta medida favorece a los sistemas que devuelven los documentos relevantes en las primeras posiciones de la lista ordenada.
- El valor de MAP para una colección de test es la media aritmética de los valores de precisión promedio para las consultas individuales.
- No se eligen niveles fijos de recall, y no hay ninguna interpolación.

- Sea $\{d_1, \dots, d_{m_j}\}$ el conjunto de documentos relevantes para una consulta $q_j \in Q$.
- Sea R_{jk} es el conjunto resultante de la recuperación ordenada desde el primer resultado hasta llegar al documento relevante d_k , entonces **MAP(Q)** será:

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

Para un conjunto de consultas Q , se promedian estas medias.

MAP tiene la ventaja de proporcionar una única métrica que se puede utilizar para comparar sistemas.

MAP tenderá a favorecer los sistemas que proporcionan documentos relevantes en rangos altos.

Pero dado que la medida esencialmente ignora el recall, puede favorecer aquellos sistemas que están ajustados para devolver pequeños conjuntos de documentos en los que tienen mucha confianza, a expensas de los sistemas que intentan ser más exhaustivos al tratar de alcanzar niveles más altos de recall.

Ejercicio#5. Para cada consulta Q1 y Q2 calcula Precisión media y MAP del conjunto.

- Tabla Precision&Recall Reales Q1 (4 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00

- Tabla Precision&Recall Reales Q2 (5 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	3/9	3/10
Recall	0.2	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6

- Precisión media Q1 =
- Precisión media Q2 =
- MAP =

Ejercicio#5. Para cada consulta Q1 y Q2 calcula Precisión media y MAP del conjunto.

- Tabla Precision&Recall Reales Q1 (4 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00

- Tabla Precision&Recall Reales Q2 (5 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	3/9	3/10
Recall	0.2	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6

- **Precisión media Q1** = $(1 + 2/3 + 3/5 + 4/6) / 4 = 0.73$
- **Precisión media Q2** = $(1 + 2/3 + 3/5 + 0 + 0) / 5 = 0.45$
- **MAP** = $(0.73 + 0.45) / 2 = 0.59$

Nótese que en la Precisión media de Q2, donde sólo se encuentran 3 de los 5 documentos relevantes, se suma 0 como valor de precisión de esos dos documentos para el cálculo de la media.

R-Precision

- Sea **R** el número total de documentos relevantes para una consulta,
- **R-Precision** será el número total de documentos relevantes encontrados entre los **R** primeros documentos devueltos, dividido por **R** .
- También se puede calcular la **media de R-Precision** entre un conjunto de preguntas.

Ejemplo.

Si hay 20 documentos relevantes para una pregunta, y entre los primeros 20 documentos devueltos se encuentran 10 documentos relevantes, este valor sería:

$$R - \text{Precision} = \frac{10}{20} = 0.5$$

Ejercicio#6. Para cada consulta Q1 y Q2 calcula R-Precision y R-Precision media.

- Tabla Precision&Recall Reales Q1 (4 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00

- Tabla Precision&Recall Reales Q2 (5 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	3/9	3/10
Recall	0.2	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6

Q1: Total de documentos relevantes=4.

R-Precision para Q2 =

Q2: Total de documentos relevantes=5.

R-Precision para Q1=

Q1 + Q2: **Media de R-Precision** para Q1 y Q2 =

Ejercicio#6. Para cada consulta Q1 y Q2 calcula R-Precision y R-Precision media.

- Tabla Precision&Recall Reales Q1 (4 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00

- Tabla Precision&Recall Reales Q2 (5 docs. relevantes)

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	No	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	3/9	3/10
Recall	0.2	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6

Q1: Total de documentos relevantes=4.

2 documentos relevantes recuperados del tope de 4 documentos. → **R-Precision** = $2/4 = 1/2$

Q2: Total de documentos relevantes=5.

3 documentos relevantes recuperados del tope de 5 documentos. → **R-Precision** = $3/5$

Q1 + Q2: **Media de R-Precision** para Q1 y Q2 = $(1/2+3/5)/2=11/20=0.55$

Precision-at- k

Precision-at- k : Precisión de los k primeros documentos recuperados (nºdocs relevantes encontrados entre los k docs vistos)

- Precisión a un nivel de recuperación fijo.
- Apropiado para búsquedas en web: la mayoría de usuarios espera encontrar lo que busca en la primera o segunda página.
- Ventaja: no es necesario saber de antemano el total de docs relevantes para cada consulta.
- Desventaja: es la métrica menos estable, no promedia bien, y arbitrariedad al tomar el parámetro de k .

Ejemplo,

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	Yes	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0.25	0.25	0.50	0.50	0.75	1.00	1.00	1.00	1.00	1.00

Precision at 5 = 3/5

Precision at 10 = 4/10