# ACTO1 – SAR

**(28/03/2022 – 2 puntos)**

**Apellidos y Nombre:** ……

**(IMPORTANTE: todos los cálculos se mostrarán redondeados a dos decimales; se deben justificar las respuestas)**

**1.** Sea una colección de documentos con 30 documentos, identificados con los números de 1 al 30. Sabemos que los documentos relevantes para una determinada consulta son los numerados del 1 al 10.

Dos sistemas de recuperación de información devuelven el siguiente resultado para la consulta:

S1= [2,13,1,15,11,3,14,4,22,19,27,5]
S2= [21,24,3,15,2,16,22,19,1,17,8,18]

Para cada uno de los sistemas se pide:

**a)** Calcular la eficacia (Precisión, Recall y la F-medida con β=1) para la consulta. **(0,2 puntos)**

| Sistema | Precisión | Recall | F-1 |
|---|---|---|---|
| S1 | | | |
| S2 | | | |

**b)** Completar las Tablas de Precision y Recall (expresando la operación de división realizada y el resultado redondeado en dos decimales, p.e. 2/3 = 0,67). **(0,3 puntos)**

**Tabla Precision&Recall Reales**

| S1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevante | | | | | | | | | | | | |
| Precisión | | | | | | | | | | | | |
| Recall | | | | | | | | | | | | |

| S2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevante | | | | | | | | | | | | |
| Precisión | | | | | | | | | | | | |
| Recall | | | | | | | | | | | | |

**c)** Calcular el MAP de ambos sistemas. **(0,2 puntos)**

**d)** Teniendo en cuenta los resultados de los anteriores apartados, ¿cuál de los dos sistemas es mejor? (justifíquese la respuesta) **(0,1 puntos)**

**2.** Sea una colección de documentos compuesta únicamente por los documentos Doc1 y Doc2 y sea la siguiente consulta:

**Doc1**: Conservación Parque Doñana
**Doc2**: Parque Nacional Garajonay
**Consulta**: Conservación Parque Nacional Garajonay

Todas las palabras de los documentos y la consulta son consideradas como términos para su representación.

**Se pide:**

**a)** Completar la tabla para un esquema de pesado lnc.ltc. **(0,4 puntos)**

| Term | | | Consulta | | | | Doc1 | | | | Doc2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $df_t$ | $idf_t$ | $f_{t,q}$ | $tf_{t,q}$ | $w_{t,q}$ | L-Norm | $f_{t,d}$ | $tf_{t,d}$ | $w_{t,d}$ | L-Norm | $f_{t,d}$ | $tf_{t,d}$ | $w_{t,d}$ | L-Norm |
| Conservación | | | | | | | | | | | | | | |
| Parque | | | | | | | | | | | | | | |
| Doñana | | | | | | | | | | | | | | |
| Nacional | | | | | | | | | | | | | | |
| Garajonay | | | | | | | | | | | | | | |

**b)** Indicar qué documento es más relevante para la consulta en base a la similitud coseno con esquema de pesado lnc.ltc. **(0,2 puntos)**

**3.** Sea un sistema de recuperación de información que dada una consulta devuelve el conjunto de documentos que contienen todos los términos de la consulta (AND entre los términos de la consulta). Se han desarrollado las 3 configuraciones siguientes y formulamos la misma consulta: **(0,3 puntos)**

**Configuración 1**: No se realiza ni eliminación de stopwords ni stemming. Sea A1 el conjunto de documentos devueltos.

**Configuración 2:** No se realiza eliminación de stopwords, pero se realiza stemming. Sea A2 el conjunto de documentos devueltos.

**Configuración 3:** Se realiza eliminación de stopwords, pero no se realiza stemming. Sea A3 el conjunto de documentos devueltos.

a) ¿Qué relación hay entre los conjuntos A1 y A2?
b) ¿Qué relación hay entre los conjuntos A1 y A3?

**4.** Se pide calcular la distancia de Levenshtein entre las palabras **arena** y **atenta,** considerando que el coste de la operación Borrado es 1, Inserción es 1, y Sustitución es 1. Utiliza la cuadrícula para representar los costes acumulados. La cuadrícula tiene un tamaño fijo, que no tiene por qué ajustarse exactamente al espacio que se requiere. **(0.3 puntos)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

## Soluciones:

1.
S1= [2,13,1,15,11,3,14,4,22,19,27,5]
S2= [21,24,3,15,2,16,22,19,1,17,8,18]

a) Calcular la eficacia (Precisión, Recall y la F-medida con β=1) para la consulta. **(0,2 puntos)**

Precisión = nº de docs relevantes recuperados/ nº de docs recuperados
Recall = nº de docs relevantes recuperados/ nº de docs relevantes en la colección

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

| Sistema | Precisión | Recall | F-1 |
|---|---|---|---|
| S1 | 5/12=0.42 | 5/10= 0.5 | 2(0.42x0.5)/(0.42+0.5)=0.46 |
| S2 | 4/12=0.33 | 4/10=0.4 | 2(0.33x0.4)/(0.33+0.4)=0.36 |

c) Completar las Tablas de Precision y Recall (expresando la operación de división realizada y el resultado redondeado en dos decimales, p.e. 2/3 = 0,67). **(0,3 puntos)**

**Tabla Precision&Recall Reales**
Precisión = nº de docs relevantes recuperados/ nº de docs recuperados
Recall = nº de docs relevantes recuperados/ nº de docs relevantes en la colección

| S1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevante | Y | N | Y | N | N | Y | N | Y | N | N | N | Y |
| Precisión | 1 | 0,50 | 0,67 | 0,50 | 0,40 | 0,50 | 0,43 | 0,50 | 0,44 | 0,40 | 0,36 | 0,42 |
| Recall | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 |

| S2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevante | N | N | Y | N | Y | N | N | N | Y | N | Y | N |
| Precisión | 0 | 0 | 0,33 | 0,25 | 0,40 | 0,33 | 0,29 | 0,25 | 0,33 | 0,30 | 0,36 | 0,33 |
| Recall | 0 | 0 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.4 | 0.4 |

c) Calcular el MAP de ambos sistemas. **(0,2 puntos)**

Para una consulta simple, como es el caso de este ejercicio, la Precisión media (MAP) es el promedio del valor de precisión obtenido después de que cada documento relevante sea recuperado en la lista ordenada de documentos recuperados
MAP(S1) = (1+0,67+0,5+0,5+0,42+0+0+0+0+0) / 10= 0,31
MAP(S2) = (0,33+0,40+0,33+0,36+0+0+0+0+0+0) / 10= 0,14

d) Teniendo en cuenta los resultados de los anteriores apartados, ¿cuál de los dos sistemas es mejor? (justifíquese la respuesta) **(0,1 puntos)**

El S1, presenta unos resultados mejores tanto en Precisión como en Recall, y por tanto en F1. Observamos también un mejor comportamiento en el MAP.

**2.**

**2.a)** El esquema de pesado es lnc.ltc, por lo que en el peso $w_{td}$ de los documentos no se aplica el pesado idf. Se han aplicado las siguientes fórmulas.

$$tf_{t,d} = \begin{cases} 1 + \log_{10} f_{t,d}, & \text{si } f_{t,d} > 0 \\ 0, & \text{otro caso} \end{cases}$$

$$\text{idf}_t = \log_{10}(N/\text{df}_t)$$

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|k|} q_i d_i}{\sqrt{\sum_{i=1}^{|k|} q_i^2}\sqrt{\sum_{i=1}^{|k|} d_i^2}}$$

| Term | | | Consulta | | | | Doc1 | | | | Doc2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $df_t$ | $idf_t$ | $f_{t,q}$ | $tf_{t,q}$ | $w_{t,q}$ | L-Norm | $f_{t,d}$ | $tf_{t,d}$ | $w_{t,d}$ | L-Norm | $f_{t,d}$ | $tf_{t,d}$ | $w_{t,d}$ | L-Norm |
| Conservación | 1 | 0,3 | 1 | 1 | 0,3 | 0,58 | 1 | 1 | 1 | 0,58 | 0 | 0 | 0 | 0 |
| Parque | 2 | 0 | 1 | 1 | 0 | 0,00 | 1 | 1 | 1 | 0,58 | 1 | 1 | 1 | 0,58 |
| Doñana | 1 | 0,3 | 0 | 0 | 0 | 0,00 | 1 | 1 | 1 | 0,58 | 0 | 0 | 0 | 0 |
| Nacional | 1 | 0,3 | 1 | 1 | 0,3 | 0,58 | 0 | 0 | 0 | 0,00 | 1 | 1 | 1 | 0,58 |
| Garajonay | 1 | 0,3 | 1 | 1 | 0,3 | 0,58 | 0 | 0 | 0 | 0,00 | 1 | 1 | 1 | 0,58 |

**2.b)**

cos(q,Doc1)=(0,58x0,58)=0,34

cos(q,Doc2)=(0,58x0,58)+(0,58x0,58)=0,67

Por lo que para la consulta es más relevante Doc2

**3.**

**3.a)** $A1 \subseteq A2$

Todos los documentos devueltos con la Configuración 1, el conjunto A1, serán también devueltos por la configuración 2. Por otra parte, el proceso de stemming de la configuración 2 puede agrupar varios términos en uno, por lo que el número de documentos que satisfacen la consulta puede ser mayor.

**3.b)** $A1 \subseteq A3$

Todos los documentos devueltos con la Configuración 1, el conjunto A1, serán también devueltos por la configuración 3. Por otra parte, El proceso de eliminación de stopwords de la configuración 3 comporta que el número de términos de la consulta a localizar en los documentos es menor, y como se realiza un AND entre los términos puede que el número de documentos que satisfacen la consulta sea mayor.

4.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| a | 5 | 4 | 4 | 3 | 2 | 2 | 2 | | |
| n | 4 | 3 | 3 | 2 | 1 | 2 | 3 | | |
| e | 3 | 2 | 2 | 1 | 2 | 3 | 4 | | |
| r | 2 | 1 | 1 | 2 | 3 | 4 | 5 | | |
| a | 1 | 0 | 1 | 2 | 3 | 4 | 5 | | |
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | | |
| | # | a | t | e | n | t | a | | |

La distancia de Levenshtein es 2, valor de la esquina superior derecha de la tabla.