e Spearman

1) Calcolare il coefficiente di correlazione di Kendall Tau tra i due ranking seguenti:

R_1: D2 D1 D4 D3
R_2: D3 D4 D1 D2

R_1	R_2	$\lvert ERR \rvert$	ERR^2
D2	D3	3	9
D1	D4	1	1
D4	D1	1	1
D3	D2	3	9
			20

MAX ERR $\frac{k(k^2-1)}{3} = \frac{4 \times 15}{3} = 20$ (5)

$$R(R_1, R_2) = 1 - \frac{6 \sum_{i=1}^{4} (R_1 - R_2)^2}{4(4^2-1)} = 1 - \frac{6 \times 20}{60} =$$

$$= 1 - \frac{120}{60} = 1 - 2 = -1$$

$R_1 \Rightarrow$ (D2, D1) (D2, D4) (D2, D3)
(D1, D4) (D1, D3)
(D4, D3)

$R_2 \Rightarrow$ (D3, D4) (D3, D1) (D3, D2)
(D4, D1) (D4, D2)
(D1, D2)

$R_1 \to R_2$ D D D
D D
D

$R_2 \to R_1$ D D D
D D
D

$$\tau(R_1, R_2) = \frac{0}{10} - \frac{10}{10} = -1$$

1. Sia *q* una query che ha 6 documenti rilevanti nella collezione. Supponiamo che un algoritmo di ritrovamento riporti il seguente ranking R_q (R indica che il documento è rilevante; N indica che il documento è non rilevante; il risultato più a sinistra è il top della lista):

R_q: RRRNNNNRNR

1. Fornire la descrizione sintetica delle metriche: *Precision*, *Recall*, *Average Precision*

(PUNTI 2)

2. Calcolare *Precision*, *Recall* ed *Average Precision* per la query *q*

(PUNTI 4)

3. Riportare la curva di precision-recall per la query *q*, usando gli 11 livelli standard di recall

(PUNTI 4)

$P = \frac{\#\text{doc rel. restr.}}{\#\text{restituiti}}$ $R = \frac{\#\text{doc ril ritrovati}}{\#\text{rilevanti nella GT.}}$

$P = \frac{5}{10} = 0.5$ $R = \frac{5}{6}$ $R - P = P@6 = \frac{3}{6}$

$AP = \frac{1 + 2/2 + 3/3 + 4/8 + 5/10 + 0}{6}$

	P	R
R	1	1/6 = 0.1666
R	1	2/6 = 0.333
R	1	3/6 = 0.5
N		
N		
N		
R	0.5	4/6 = 0.667
N		
R	0.5	5/6 = 0.833

P	R	
1	0	
1	01	
1		0.1666
1	02	
1	03	
1		0.333
1	04	
1	05	
0.5	06	
0.5		0.667
0.5	07	
0.5	08	
0.5		0.833
0	09	
0	1	

$P(r_j) = \max_{\forall r_i \geq r_j} P(r_i)$

1) Siano dati i seguenti documenti e la query Q rappresentati come vettori di pesi TF-IDF non normalizzati:

	T1	T2	T3	T4	T5	T6
D1	2	2	0	0	0	0
D2	0	0	1	2	3	0
D3	2	1	0	2	0	0
D4	5	1	1	0	2	0
Q	0	0	3	4	0	0

a) Calcolare il ranking dei documenti rispetto alla query Q utilizzando la similarità del coseno.
(PUNTI 3)

b) Assumendo che D3 e D4 siano rilevanti, mentre D1 e D2 non siano rilevanti, riformulare la query utilizzando l'algoritmo di Rocchio (utilizzare i pesi 1, 0,75 e 0,25 per query iniziale, centroide dei documenti rilevanti e centroide dei documenti non rilevanti, rispettivamente).
(PUNTI 5)

a) $COSIM(D,Q) = \dfrac{\sum_{i=1}^{|V|} d_i \cdot q_i}{\|d_i\| \, \|q\|}$

$\|D_1\| = \sqrt{2^2+2^2} = \sqrt{8}$

$\|D_2\| = \sqrt{14}$

$\|D_3\| = \sqrt{9} = 3$

$\|D_4\| = \sqrt{31}$

$\|Q\| = \sqrt{25} = 5$

$COSIM(D_1, Q) = 0$

$COSIM(D_2, Q) = \dfrac{1\times 3 + 2\times 4}{\sqrt{14} \times 5} = \dfrac{11}{5\sqrt{14}} = 0.58$

$\text{Cosim}(D3, Q) = \frac{2 \times 4}{3 \times 5} = \frac{8}{15} = 0.53$

$\text{Cosim}(D4, Q) = \frac{1 \times 3}{\sqrt{31} \times 5} = \frac{3}{5\sqrt{31}} = 0.10$

Ranking
- D2 0
D3 X
D4 X
- D1 0

$\vec{Q}_0 = (0, 0, 3, 4, 0, 0)$

$\vec{Q}_1 = 1 \times \vec{Q}_0 + 0.75\ RIL - 0.25\ NRIL$

$D3 = \left(\frac{2}{3}, \frac{1}{3}, 0, \frac{2}{3}, 0, 0\right)$

$D4 = \left(\frac{5}{\sqrt{31}}, \frac{1}{\sqrt{31}}, \frac{1}{\sqrt{31}}, 0, \frac{2}{\sqrt{31}}, 0\right)$

Centroide RIL

$\vec{RIL} = \left(\frac{\frac{2}{3} + \frac{5}{\sqrt{31}}}{2}, \frac{\frac{1}{3} + \frac{1}{\sqrt{31}}}{2}, \frac{0 + \frac{1}{\sqrt{31}}}{2}, \frac{\frac{2}{3}}{2}, \frac{\frac{2}{\sqrt{31}}}{2}, \right.$

Centr. NRIL

$\vec{NRIL} = \left(\frac{\frac{2}{\sqrt{81}}}{2}, \frac{\frac{2}{\sqrt{81}} + }{2}, \frac{\frac{1}{\sqrt{14}}}{2}, \frac{\frac{2}{\sqrt{14}}}{2}, \frac{\frac{3}{\sqrt{14}}}{2}, 0\right)$

$\vec{Q}_1 = 1 \times \vec{Q}_0 + 0.75\ \vec{RIL} - 0.25\ \vec{NRIL}$

$= 0 + \underbrace{\left(0.75 \times \frac{\frac{2}{3} + \frac{5}{\sqrt{3}}}{2}\right) - 0.25 \left(\frac{\frac{2}{\sqrt{8}}}{2}\right)}_{\text{I Coordinate}}$

1. Sia *q* una query che ha 5 documenti rilevanti nella collezione. Supponiamo che un algoritmo di ritrovamento applicato a *q* riporti il seguente ranking `Rq: D1 D3 D5 D7 D9 D4`

Supponiamo che `D1`, `D7` e `D9` siano documenti rilevanti per *q*

1. Calcolare Precision, Recall ed Average Precision per *q*, fornendo anche una descrizione formale delle metriche

(PUNTI 4)

2. Supponendo di avere dei giudizi di rilevanza non binari, e assumendo che `D1` e `D9` abbiano un grado di rilevanza pari a 3, mentre `D7` abbia un grado di rilevanza pari a 2, calcolare il valore dell'*nDCG* (normalized Discounted Cumulative Gain) per *q*, fornendo anche una breve descrizione della metrica.

(PUNTI 5)

D1	D3	D5	D7	D9	D4
R	0	0	R	R	0

$P = \frac{3}{6}$

$R = \frac{3}{5}$

$$AP = \frac{1 + 2/4 + 3/5 + 0 + 0}{5}$$

D1	D3	D5	D7	D9	D4
3	0	0	2	3	0

$$DCG[i] = \begin{cases} G[1] & i = 1 \\ \frac{G[i]}{\log_2 i} + DCG[i-1] & i > 1 \end{cases}$$

$$DCG = \left(3, 0+3=3, 3, \frac{2}{\log_2 4} + 3 = 4, \frac{3}{\log_2 5} + 4, \frac{3}{\log_2 5} + 4\right)$$

$$= \left(3, 3, 3, 4, \frac{3}{\log_2 5} + 4, \frac{3}{\log_2 5} + 4\right)$$

I DCG

partendo dall'IG = (3, 3, 2, 0, 0, 0)

$$IDCG = \left(3, \frac{3}{\log_2^2} + 3 = 6, \frac{2}{\log_2^3} + 6, \frac{2}{\log_2^3} + 6, \frac{2}{\log_2^3} + 6, \frac{2}{\log_2^3} + 6\right)$$

$$nDCG = \left(\frac{3}{3}, \frac{3}{6}, \frac{3}{\frac{2}{\log_2^3} + 6}, \frac{4}{x}, \frac{\frac{3}{\log_2^5} + 4}{x}, \longrightarrow \text{UGLAGU}\right)$$