Esercizio sul relevance feedback

Sia data la seguente matrice termini documenti contenente ***pesi TF-IDF non normalizzati***:

	t1	t2	t3	t4	t5	t6
d1	**0,8**	**1,2**	**0**	**0,7**	**0**	**0**
d2	**0,1**	**1**	**1,4**	**0**	**0**	**0**
d3	**0**	**0,2**	**0**	**0**	**3,2**	**0,9**
d4	**0,1**	**0,1**	**0,1**	**0**	**2,3**	**1,7**
d5	**0**	**0**	**2**	**2**	**1**	**0**

e la query q=(t1:1, t2:2)

1) Calcolare il ranking dei documenti rispetto alla query *q* utilizzando la similarità del coseno.
2) Assumendo che il *terzo* ed il *quarto* documento del ranking siano rilevanti e che il *primo* documento del ranking non sia invece rilevante, riformulare la query utilizzando l'algoritmo di Rocchio e ricalcolare il ranking dei documenti.

Come primo passo calcolo le lunghezze dei documenti e della query per normalizzare i vettori sulla propria lunghezza.

$$|d1| = \sqrt{0{,}8^2 + 1{,}2^2 + 0{,}7^2} = 1{,}603$$

$$|d2| = \sqrt{0{,}1^2 + 1^2 + 1{,}4^2} = 1{,}723$$

$$|d3| = \sqrt{0{,}2^2 + 3{,}2^2 + 0{,}9^2} = 3{,}330$$

$$|d4| = \sqrt{0{,}1^2 + 0{,}1^2 + 0{,}1^2 + 2{,}3^2 + 1{,}7^2} = 2{,}865$$

$$|d5| = \sqrt{2^2 + 2^2 + 1^2} = 3$$

$$|q| = \sqrt{1^2 + 2^2} = 2{,}236$$

Dunque la matrice termini documenti con pesi normalizzati è la seguente (rappresento anche la query)

	t1	t2	t3	t4	t5	t6
d1	0,499	0,749	0	0,437	0	0
d2	0,058	0,580	0,812	0	0	0
d3	0	0,060	0	0	0,961	0,270
d4	0,035	0,035	0,035	0	0,803	0,593
d5	0	0	0,667	0,667	0,333	0
q	0,447	0,894	0	0	0	0

Calcolo la similarità del coseno tra la query q e tutti i documenti. Riporto il calcolo esplicito solo per un singolo documento (calcolo solo il prodotto interno perché sto operando su vettori di lunghezza unitaria):

$$cosim(d1, q) = 0{,}499 * 0{,}447 + 0{,}749 * 0{,}894 = 0{,}893$$

$$cosim(d2, q) = 0{,}545$$

$$cosim(d3, q) = 0{,}054$$

$$cosim(d4, q) = 0{,}047$$

$$cosim(d5, q) = 0 \ (nessun\ termine\ in\ comune\ con\ la\ query)$$

Il ranking dei documenti rispetto alla query *q* è il seguente: d1, d2, d3, d4, d5

Assumiamo che l'utente fornisca del feedback ed indichi d3 e d4 rilevanti, mentre d1 non rilevante.

Calcolo il centroide dei documenti rilevanti, facendo la media delle singole coordinate:

	t1	**t2**	**t3**	**t4**	**t5**	**t6**
d3	0	0,060	0	0	0,961	0,270
d4	0,035	0,035	0,035	0	0,803	0,593
Centroide documenti rilevanti	0,017	0,047	0,017	0	0,882	0,432

Il centroide dei documenti non rilevanti corrisponde al vettore del documento d1, unico documento non rilevante.

Applico il metodo di Rocchio per ottenere la query modificata, sommando al vettore della query iniziale *q*, il centroide dei documenti rilevanti (peso di tale vettore pari a 0,75) e sottraendo il centroide dei documenti non rilevanti (peso pari a 0,25). Ottengo la seguente query q1:

	t1	**t2**	**t3**	**t4**	**t5**	**t6**
q	0,447	0,894	0	0	0	0
Centroide documenti rilevanti (RIL)	0,017	0,047	0,017	0	0,882	0,432
Centroide documenti non rilevanti (NRIL)	0,499	0,749	0	0,437	0	0
q1=q0+0,75*RIL-0,25*NRIL	0,336	0,743	0,013	-0,109 (*)	0,661	0,324

(*) i pesi negativi vengono annullati e dunque questa è la rappresentazione del vettore query q1:

q1	0,336	0,743	0,013	0	0,661	0,324

Calcolo nuovamente la similarità del coseno tra la query *q1* e tutti i documenti:

$$cosim(d1, q1) = 0{,}675$$

$$cosim(d2, q1) = 0{,}461$$

$$cosim(d3, q1) = 0{,}767$$

$$cosim(d4, q1) = 0{,}761$$

$$cosim(d5, q1) = 0{,}156$$

Il nuovo ranking dei documenti rispetto alla query *q1* è il seguente: d3, d4, d1, d2, d5.

E' possibile notare che i documenti per cui era stato espresso feedback positivo (d3, d4) sono ora al top del ranking, mentre quello per cui era stato espresso feedback negativo (d1) è ora sceso dalla prima alla terza posizione del ranking.