lunedì 26 ottobre 2020 11:37

1) Siano dati in input la seguente query `Q` e il documento `D`:

```
Q = "information retrieval"
D = "information retrieval and text retrieval"
```

Calcolare la similarità del coseno tra la query `Q` ed il documento `D`, assumendo che:

- il termine `and` sia una stopword
- il document frequency dei termini `information`, `retrieval` and `text` siano rispettivamente 10, 50 e 100
- il numero di documenti nella collezione sia N = 1000
- sia utilizzato il `tf-idf` come schema di pesatura dei termini nel documento e nella query (non normalizzare il term frequency).

Rappresento Q e D sotto forma di BAG OF WORDS

$Q = \langle information: 1, retrieval: 1 \rangle$

$D = \langle information: 1, retrieval: 2, text: 1 \rangle$

CALCOLIAMO GLI IDF

$IDF_{information}: \log_{10} \frac{1000}{10} = \log_{10} 100 = 2$

$IDF_{retrieval}: \log_{10} \frac{1000}{50} = \log 20 \simeq 1.3$

$IDF_{text}: \log_{10} \frac{1000}{100} = \log 10 = 1$

Definiamo la MATRICE TERMINI – DOCUMENTI che riporta il TF-IDF

	Q	D
information	$1 \times 2 = 2$	$1 \times 2 = 2$
retrieval	$1 \times 1.3 = 1.3$	$2 \times 1.3 = 2.6$
Text	0	$1 \times 1 = 1$

$\vec{Q} = (2, 1.3, 0)$

$\vec{D} = (2, 2.6, 1)$

$$\text{COSIM}(\vec{D}, \vec{Q}) = \frac{\vec{D} \cdot \vec{Q}}{\|\vec{D}\| \|\vec{Q}\|} = \frac{2 \times 2 + 1.3 \times 2.6 + 0 \times 1}{\sqrt{2^2 + (2.6)^2 + 1^2} \sqrt{2^2 + (1.3)^2}} =$$

$$\simeq 0.9$$