# ESERCIZIO

$q_1$ restituisce 70 doc di cui 38 rilevanti

$q_2$ " 50 " 25 rilevanti.

Quale query conviene fornire?

Calcolo la PRECISIONE

$$P_{q1} = \frac{38}{70} = 0.54 \qquad P_{q2} = \frac{25}{50} = 0.50$$

CONVIENE $q_1$

Se SAPPIAMO anche che per entrambe le query i doc rilevanti nella collezione sono 90, la risposta cambia?

Posso calcolare il recall

$$R_{q1} = \frac{38}{90} = 0.42 \qquad R_{q2} = \frac{25}{90} = 0.27$$

1) Siano dati i seguenti documenti estratti da una collezione di 100 documenti:

```
D1 = "T1 T2 T1 T3"
D2 = "T3 T4"
D3 = "T1 T5 T4 T4"
```

a) Fornire la rappresentazione dei documenti sotto forma di bag-of-words
(PUNTI 2)

b) Costruire l'indice invertito della collezione
(PUNTI 2)

c) Calcolare la rappresentazione TF-IDF per i 3 documenti (*usare il numero di occorrenze non normalizzato per il TF*)
(PUNTI 3)

d) Utilizzando la similarità del coseno, definire il ranking dei documenti in risposta alla query Q = "T1 T2 T6"
(PUNTI 3)

a) BOW

$D1 = \langle T1: 2, T2: 1, T3: 1 \rangle$

$D2 = \langle T3: 1, T4: 1 \rangle$

$D3 = \langle T1: 1, T4: 2, T5: 1 \rangle$

b) Indice invertito

$T1 \rightarrow D1 \quad D3$

$T2 \rightarrow D1$

$T3 \rightarrow D1 \quad D2$

$T4 \rightarrow D2 \quad D3$

$T5 \rightarrow D3$

c) Rappr. TFIDF - Per calcolare IDF ho bisogno di sapere la cardinalità della collezione, che è 100 e i doc frequency dei termini -

Non ho tutte le info per calcolare IDF, allora faccio ipotesi realistica:

$df_{T1} = 50 \qquad IDF_{T1} = \log \frac{100}{50} = \log 2$

$df_{T2} = 50 \qquad IDF_{T2} = \log 2$

$df_{T3} = 90 \qquad IDF_{t3} = \log \frac{100}{90}$

$df_{T4} = 80 \qquad IDF_{T4} = \log \frac{100}{80}$

$df_{T5} = 10 \qquad IDF_{T5} = \log \frac{100}{10} = \log 10$

$df_{T6} = 30 \qquad IDF_{T6} = \log \frac{100}{30}$

| | T1 | T2 | T3 | T4 | T5 | T6 |
|---|---|---|---|---|---|---|
| D1 | $2 \times \log 2$ | $\log 2$ | $\log 10/9$ | 0 | 0 | 0 |
| D2 | 0 | 0 | $\log 10/9$ | $\log 10/8$ | 0 | 0 |

D3 $\log 2$ 0 0 $2\log 10/8$ $\log 10$ 0

d) $Q = \langle T1, T2, T6 \rangle$

Q $\log 2$ $\log 2$ 0 0 0 $\log 10/3$

$$GSIM(D1, Q) = \frac{2 \times \log 2 \times \log 2 + \log 2 \log 2}{|D1|\ |Q|} =$$

$$|D1| = \sqrt{(2\log 2)^2 + (\log 2)^2 + (\log 10/3)^2} \qquad |Q| = \sqrt{(\log 2)^2 + (\log 2)^2 + (\log 10/3)^2}$$

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 $R_q$: D1 D5 D3 D7 D9 D4. Supponendo di avere dei giudizi di rilevanza non binari espressi in una scala a 5 valori (1-5), e assumendo che D1 e D9 abbiano rilevanza pari a 5, mentre D5 abbia rilevanza pari a 3, calcolare il valore dell'$nDCG$ per $q$.

$R_q$ = D1 D5 D3 D7 D9 D4

$G = (5, 3, 0, 0, 5, 0)$

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

$$DCG = \left(5, \underbrace{\frac{3}{\log_2 2} + 5}_{8}, \underbrace{\frac{0}{\log_2 3} + 8}_{8}, \underbrace{\frac{0}{\log_2 4} + 8}_{8}, \frac{5}{\log_2 5} + 8, 0 + \frac{5}{\log_2 5} + 8\right)$$

$DCG = (5, 8, 8, 8, 10.2, 10.2)$

$IG = (5, 5, 3, 0, 0, 0)$

$$IDCG = \left(5, \underbrace{\frac{5}{\log_2 2} + 5}_{10}, \frac{3}{\log_2 3} + 10, \frac{3}{\log_2 3} + 10, \frac{3}{\log_2 3} + 10, \frac{3}{\log_2 3} + 10\right) =$$

$IDCG = (5, 10, 11.9, 11.9, 11.9, 11.9)$

$$nDCG = \frac{DCG}{IDCG} = \left(\frac{5}{5}, \frac{8}{10}, \frac{8}{11.9}, \frac{8}{11.9}, \frac{10.2}{11.9}, \frac{10.2}{11.9}\right)$$

1) Siano dati l'insieme delle categorie C = {c1, c2} e una collezione di 1000 documenti definiti sul vocabolario V = {T1,T2,T3,T4,T5}.

a) Costruire un classificatore *bayesiano* per C, addestrandolo sul seguente training set TR:

TR = {<D1,c1>, <D2,c1>, <D3,c2>, <D4,c2>}

dove per ogni documento si riporta di seguito l'elenco delle parole in esso presenti, con le relative occorrenze:

D1={T1:1, T2:2, T3:3} D2={T4:1}

D3={T1:2, T2:5} D4={T3:4, T4:2}

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 6)

b) Determinare la classe di appartenenza del seguente documento d={T2:2,T5:2}

(PUNTI 2)

Secondo Bayes $P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)}$

Per classif. un doc. devo Trovare $\arg\max_{c_i} P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)}$

Poiché mi interessa trovare il massimo, posso eliminare $P(d)$

Calcolo $P(c_i)$ e $P(d|c_i)$ stimando le dal training set

$P(c_i) = \frac{\#\text{ doc di classe } c_i \text{ nel training set}}{\#\text{ tot doc nel training set}}$

$P(c_1) = \frac{2}{4}$ $P(c_2) = \frac{2}{4}$

$P(d|c_i) = P(t_1 \wedge \dots \wedge t_k | c_i) \overset{\text{ASSUNZIONE INDIPENDENZA}}{=} P(t_1|c_i)P(t_2|c_i)\dots P(t_k|c_i)$

$= \prod_{j=1}^{k} P(t_j|c_i)$

$P(t_k|c_i) = \frac{\#\text{ volte in cui } t_k \text{ compare nei doc classe } c_i + 1}{\#\text{ Tot termini nei doc classe } c_i + |V|}$

$P(T_1|c_1) = (1+1)/(7+5) = 2/12$

$P(T_2|c_1) = 3/12$

$P(T_3|c_1) = 4/12$

$P(T_4|c_1) = 2/12$

$P(T_1|c_2) = (2+1)/(13+5) = 3/18$

$P(T_2|c_2) = 6/18$

$P(T_3|c_2) = 5/18$

$P(T_4|c_2) = 3/18$

$P(T_4|C_1) = 2/12$ $\quad P(T_4|C_2) = 3/18$

$P(T_5|C_1) = 1/12$ $\quad P(T_5|C_2) = 1/18$

b) Classificare $d = \{T_2:2,\ T_5:2\}$

$$P(C_1|d) = P(C_1)\,P(d|C_1) = \frac{2}{4}\left(\frac{3}{12}\right)^2\left(\frac{1}{12}\right)^2 = \alpha$$

$$P(C_2|d) = P(C_2)\,P(d|C_2) = \frac{2}{4}\left(\frac{6}{18}\right)^2\left(\frac{1}{18}\right)^2 = \beta$$

Se $\alpha > \beta \Rightarrow d \in C_1$

Se $\alpha < \beta \Rightarrow d \in C_2$

1) Sia $q$ una query che ha 6 documenti rilevanti nella collezione. Supponiamo che un algoritmo di ritrovamento applicato a $q$ riporti il seguente ranking $R_q$: D1 D2 D3 D4 D5 D6
Supponiamo che D2, D4 e D6 siano documenti rilevanti per $q$

a) Calcolare l'*Average Precision*, e il *Recall* per la query $q$, fornendo anche una descrizione delle metriche

(PUNTI 4)

b) Riportare la curva di precision-recall per la query $q$, usando gli 11 livelli standard di recall

(PUNTI 3)

R = D1 D2 D3 D4 D5 D6

0 X 0 X 0 X

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

$Recall = \frac{\#\text{doc ril. rit}_q}{\#\text{doc ril.}} = \frac{3}{6}$

R =

0

X P=0.5 R= 1/6 = 0.167

0

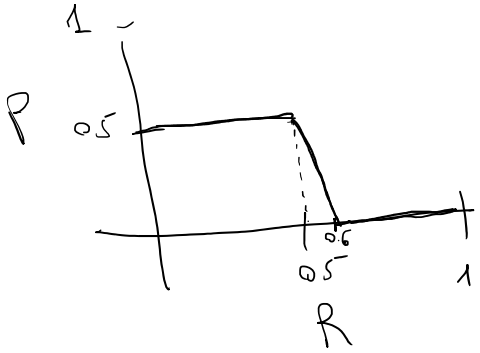
X P = 0.5 R=2/6 = 0.333

0

X P=0.5 R= 3/6 = 0.5

$P(\pi_j) = P(\pi)$

$\forall \pi_i \neq \pi_j$




P R

0.5 0

0.5 01 0.167

0.5 02

0.5 03 0.333

0.5 04

0.5 05

0 06

0 07

0 08

0 09

0 1