

**Prova scritta di**  
**METODI PER IL RITROVAMENTO DELL'INFORMAZIONE**  
**C.d.L. in Informatica - A.A. 2020-21**  
**Docenti: P. Lops, P. Basile - 21 Gennaio 2021**  
**II turno**

- 1) Siano dati l'insieme delle categorie  $C = \{c_1, c_2\}$  e una collezione di 200 documenti definiti sul vocabolario  $V = \{T_1, T_2, T_3, T_4, T_5\}$ .

Costruire un classificatore bayesiano per  $C$ , addestrandolo sul seguente training set  $TR$ :

$$TR = \{ \langle d_1, c_1 \rangle, \langle d_2, c_2 \rangle, \langle d_3, c_1 \rangle, \langle d_4, c_2 \rangle \}$$

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

$$\begin{aligned} d_1 &= \{T_2:1, T_5:4\} & d_2 &= \{T_3:1, T_5:2\} \\ d_3 &= \{T_1:3, T_3:2\} & d_4 &= \{T_1:1, T_2:2, T_3:1\} \end{aligned}$$

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

Determinare la classe di appartenenza del documento  $d_x = \{T_2:2, T_4:2\}$

(PUNTI 3)

- 2) Sia  $q$  una query che ha 5 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):

$$R_q: \text{NNNRNRRRNN}$$

- a) Fornire la descrizione sintetica delle metriche: *Precision*, *Recall*, *R-Precision* ed *Average Precision*

(PUNTI 4)

- b) Calcolare *Precision*, *Recall*, *R-Precision* ed *Average Precision* per la query  $q$

(PUNTI 4)

- 3) Descrivere il problema dello *spider trap* e del *dead end* nell'algoritmo PageRank e illustrare una possibile soluzione.

(PUNTI 5)

- 4) Fornire la definizione di *synset* in WordNet e descrivere in maniera sintetica la relazione di *ipponimia-iperonimia*.

(PUNTI 4)

- 5) Descrivere, commentando opportunamente, la funzione per il calcolo delle predizioni dei rating in un algoritmo di filtraggio collaborativo di tipo *user to user*.

(PUNTI 6)