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
I turno

1) Siano dati l'insieme delle categorie $C = \{c_1, c_2\}$ e una collezione di 150 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*:
$$TR = \{<d_1,c_1>, <d_2,c_1>, <d_3,c_2>, <d_4,c_2>\}$$
dove per ogni documento d_j si riporta di seguito l'elenco delle parole in esso presenti, con le relative occorrenze:
$d_1=\{T_1:2, T_2:3, T_3:4\}$ $d_2=\{T_1:1, T_4:2\}$
$d_3=\{T_2:1, T_4:2\}$ $d_4=\{T_1:1, T_3:2\}$
NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

Determinare la classe di appartenenza del documento $d_x=\{T_2:2, T_5:2\}$

(PUNTI 3)

2) 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: RNRRNNNRNN
   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 in maniera sintetica i principi alla base del PageRank, focalizzando l'attenzione sulla formulazione basata sul Flow model

(PUNTI 6)

4) Descrivere il processo di modifica delle query basato sul metodo del *relevance feedback* (algoritmo di Rocchio).

(PUNTI 5)

5) Illustrare in maniera sintetica il problema dell'overspecialization nei recommender systems di tipo content-based

(PUNTI 4)