Prova scritta di
METODI PER IL RITROVAMENTO DELL'INFORMAZIONE
C.d.L. in Informatica - A.A. 2020-21
Docenti: P. Lops, P. Basile - 25 Febbraio 2021

1) Siano dati l'insieme delle categorie $C = \{c_1, c_2\}$ e una collezione di documenti definiti sul vocabolario $V = \{T_1, T_2, T_3, T_4, T_5, T_6\}$.
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:
$d_1=\{T_1:2, T_2:3, T_3:4\}$ $d_2=\{T_1:1, T_4:2\}$ $d_3=\{T_2:1, T_4:2\}$ $d_4=\{T_1:1, T_2:2, T_6:4\}$
NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

Determinare la classe di appartenenza del documento $d_x=\{T_4:2, T_5:2\}$

(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):
$$R_q: \text{ RNNRNNR}$$
a) Fornire la descrizione sintetica delle metriche: *F1, R-Precision* ed *Average Precision*

(PUNTI 3)

b) Calcolare *F1, R-Precision* ed *Average Precision* per la query q

(PUNTI 3)

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 *iponimia-iperonimia*.

(PUNTI 4)

5) Descrivere i problemi principali dei recommender systems di tipo collaborativo.

(PUNTI 5)

6) Calcolare il coefficiente di correlazione di Spearman tra i due ranking seguenti:
R_1: D1 D2 D3 D4 D5
R_2: D3 D4 D1 D5 D2

(PUNTI 3)